

Semantics @ Roche Forum 2015

September 8, 2015 | Basel, Switzerland

Terminologies and ontologies - do we need standards for semantic artefacts?

Stefan Schulz

Institute for Medical Informatics,
Statistics und Documentation

Medical University of Graz

stefan.schulz@medunigraz.at

✓ TECH & SCIENCE

There Are 3 Trillion Trees on Earth, 8 Times What We Previously Thought

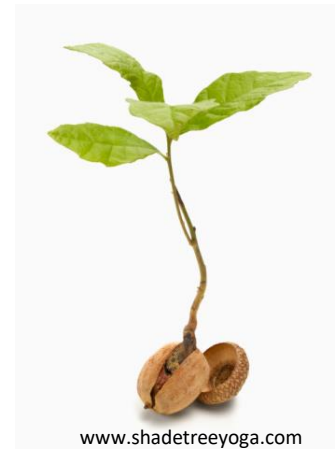
BY DOUGLAS MAIN 9/3/15 AT 12:38 PM



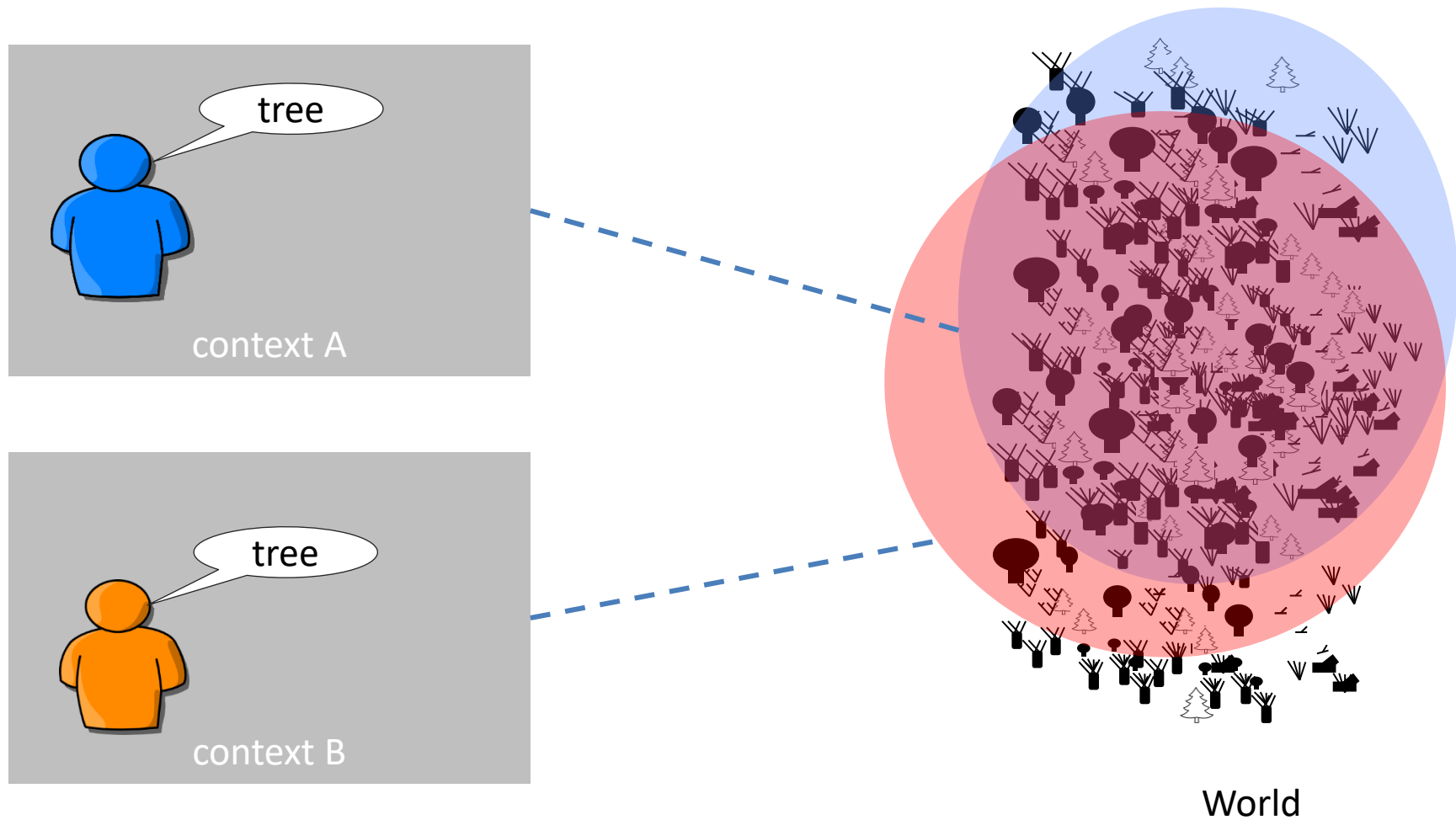
Deforestation in northwestern Brazil. Humans have cut down about half of the Earth's original tree cover. LUNAE
PARRACHO / REUTERS

Newsweek, 3 Sept 2015

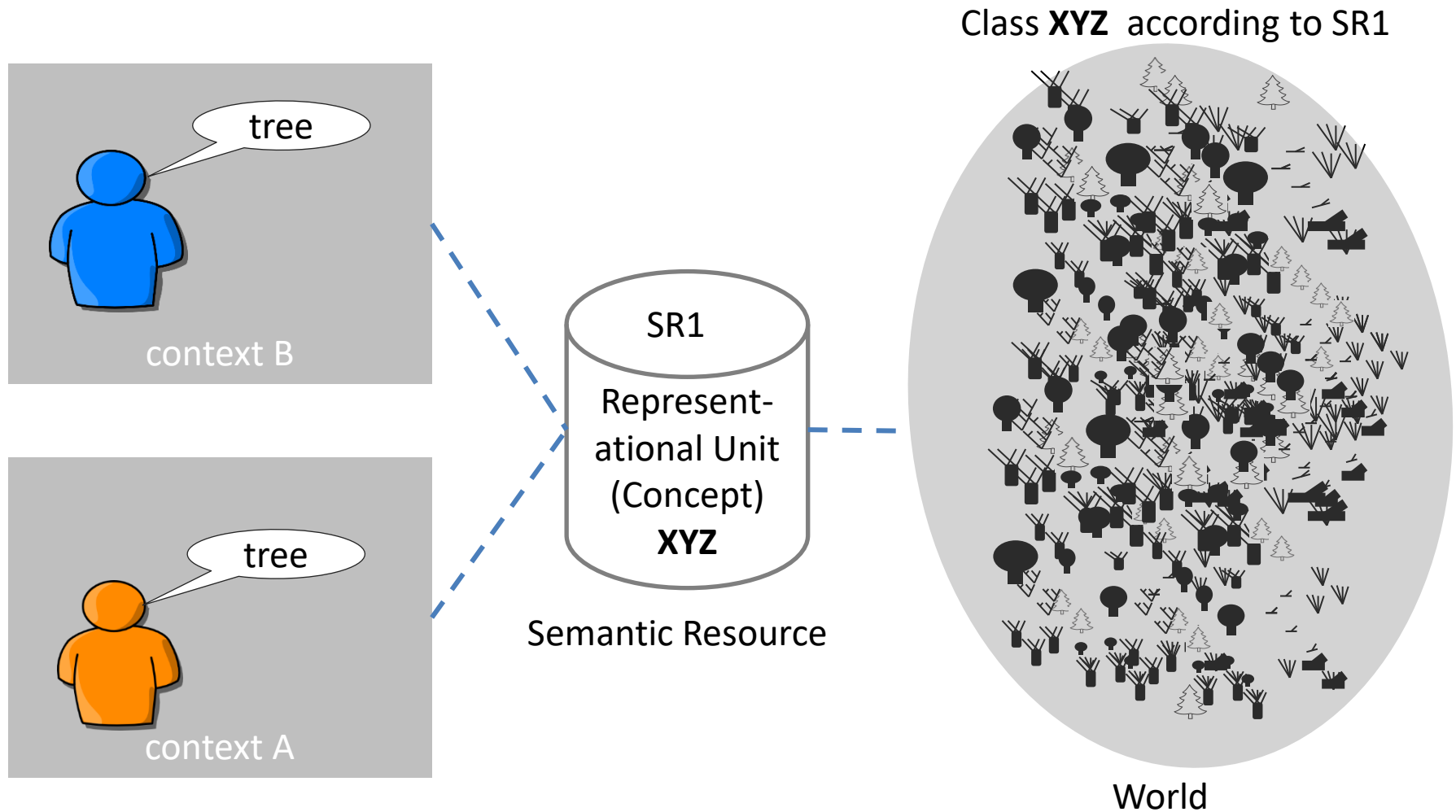
What is a tree ?



Different views



Semantic normalization



Terminology vs. Ontology

- Terminological aspects

- Preferred label
- Synonyms, translations
- Hypernyms / Hyponyms



bla bla bla

Tree

Preferred term (English): tree (plant):

Other terms

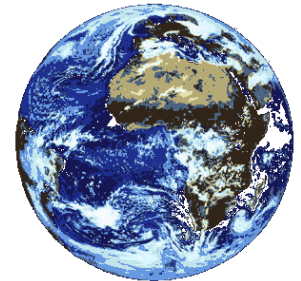
English: tree

German: Baum (m., pl. Bäume)

French: arbre (f.)

- Ontological aspects

- textual definition
- formal definition



"a perennial plant with an elongated stem, or trunk, supporting branches and leaves"

PerennialPlant and

hasPart some Stem and

atSomeTime some (hasPart some Leaf) and
atSomeTime some (hasPart some Branch)

Existing semantic resources for life sciences

- **Bioportal** hosts 461 ontologies and other terminology systems
- **The Unified Medical Language System (UMLS)** hosts and links 179 biomedical terminology systems
- Large content overlap

The screenshot shows the BioPortal homepage. At the top is a navigation bar with links: BioPortal, Browse, Search, Mappings, Recommender, Annotator, Resource Index, and Projects. Below the bar, a welcome message states: "Welcome to BioPortal, the world's most comprehensive repository of biomedical ontologies." A link for help is provided. The main content area has two search sections. The first, "Search all ontologies", has a text input field with the placeholder "Enter concept, e.g. Melanoma", a "Search" button, and a link to "Advanced Search". The second, "Find an ontology", has a text input field with the placeholder "Enter ontology name, e.g. NCI Thesauri", an "Explore" button, and a link to "Browse Ontologies". Below these are two lists. "Ontology Visits (August 2015)" is a table with two columns: ontology name and visit count. "Latest Notes" is a list of recent updates with timestamps.

Ontology	Visits
Current Procedural Terminology (CPT)	29937
RxNORM (RXNORM)	11243
Systematized Nomenclature of Medicine – Clinical Terms (SNOMEDCT)	11204
Medical Dictionary for Regulatory Activities (MEDDRA)	9569
National Drug Data File (NDDF)	3868
More	

Latest Notes

- Change Property Value Proposal: Synonym proposed for use for "viral hemagglutination inhibition assay" OBI:0000873 (Ontology for Biomedical Investigations) 4 months ago by sbaur
- Change Property Value Proposal: Not reflective of what the scientific community uses (Ontology for Biomedical Investigations) 4 months ago by sbaur
- The RDF format of MeSH is inaccessible (Medical Subject Headings) 6 months ago by A.Mina
- The RDF format of MeSH is inaccessible, I can't download it for the french and the english langu...

The screenshot shows the UMLS Terminology Services (UTS) homepage. At the top is a header: "A service of the U.S. National Library of Medicine | National Institutes of Health". Below this is a logo for the "Unified Medical Language System" and the text "UMLS Terminology Services". A navigation bar contains links: UTS Home, Applications, SNOMED CT, Resources, Downloads, Documentation, and UMLS Home. The main content area has a "Welcome »" section with links: Request a License, News/Announcements, Training, and Supported Browsers. To the right is a "Welcome to the UTS" section with the heading "The UMLS Terminology Services (UTS) allows you to:" followed by a bulleted list of services.

Welcome to the UTS

The UMLS Terminology Services (UTS) allows you to:

- Request a UMLS Metathesaurus License and create a UTS account
- Search and display content from UTS Applications including:
 - Metathesaurus Browser
 - Semantic Network Browser
 - SNOMED CT Browser
- Download data files including:
 - UMLS Knowledge Sources
 - RxNorm weekly and monthly updates
 - SNOMED CT
 - CORE Problem List and Route of Administration Subsets

Problems

- Resources are tailored to specific use cases
 - E.g.: in ICD 10 "Thrombosis" does not include "Thrombosis in pregnancy" (use for health statistics)
- Resources address implicit contexts
 - E.g.: the Foundational Model of Anatomy describes *canonical* anatomy
- Resources are no longer maintained
 - 50 source vocabularies in UMLS not "active"
- Resources are semantically shallow
 - Relations like "broader than", "associated with"
- Resources are just bad quality
 - e.g. use OWL ignoring OWL semantics (NCI Thesaurus)

Problems (cont.)

- Resources are incomplete
 - missing definitions, e.g. in most of ICD 10
 - fuzzy text definitions (MeSH: trees are *usually* tall (...) having *usually* a main stem)
 - undefined primitives
(unclear if pericardium is part of heart)
 - ambiguous preferred terms
"eye": same label for human and drosophila eyes
 - missing synonyms / entry terms
for most of GO terms no match with any text passage in literature, e.g. "*tetrahydromethanopterin-dependent serine hydroxymethyltransferase activity*"

Three Strategies for tailored semantic resources

1. Re-use existing resources, tolerate heterogeneity
2. Create and maintain application-specific resources
3. Join terminology / ontology standardisation / activities

1. Reuse existing resources

- Tolerate semantic heterogeneity and underspecification including errors, unknown contexts
 - Hendler: "A Little Semantics Goes A Long Way" (?)
- Accept lack of precision when doing terminology / ontology mapping at term level
- Appropriate where results do not need to be precise:
 - High recall document or fact retrieval



Three Strategies for tailored semantic resources

1. Re-use existing resources, tolerate heterogeneity
2. Create and maintain application-specific resources
3. Join terminology / ontology standardisation / activities

2. Create application-specific resources from scratch

- Use case driven terminology / ontology engineering
- Tailored content, no unnecessary ballast
- Pragmatic / idiosyncratic solutions prevent re-use / interoperability
- Engineering / maintenance costs
- Yet another species in the ontology zoo

"Deciding whether a particular concept is a class in an ontology or an individual instance depends on what the potential applications of the ontology are."

Natasha Noy & Deborah McGuinness:
Ontology Development 101

http://protege.stanford.edu/publications/ontology_development/ontology101.pdf

Three Strategies for tailored semantic resources

1. Re-use existing resources, tolerate heterogeneity
2. Create and maintain application-specific resources
3. Join terminology / ontology standardisation / activities

3. Contribute to develop existing (content) standards / specifications

- Join communities that use common terminology / ontology specifications
- Contribute to development / maintenance
- Ontologies
 - objective descriptions of a domain and not as application-specific knowledge bases (scientific realism*)
 - Only express what is universally true
- Examples
 - SNOMED CT
 - OBO Foundry
 - Upper-level ontologies (BFO, DOLCE, BioTop)



SNOMED CT



SNOMED CT

- Terminology / Ontology that represents entities relevant for clinical documentation
- Approx. 300, 000 representational units ("concepts")
- Formal definitions in OWL-EL
- Terms in several languages
 - Fully specified names: non-ambiguous labels
 - Synonyms: close-to user terms
- Maintained by IHTSDO

IHTSDO: International Health Standards Development Organisation

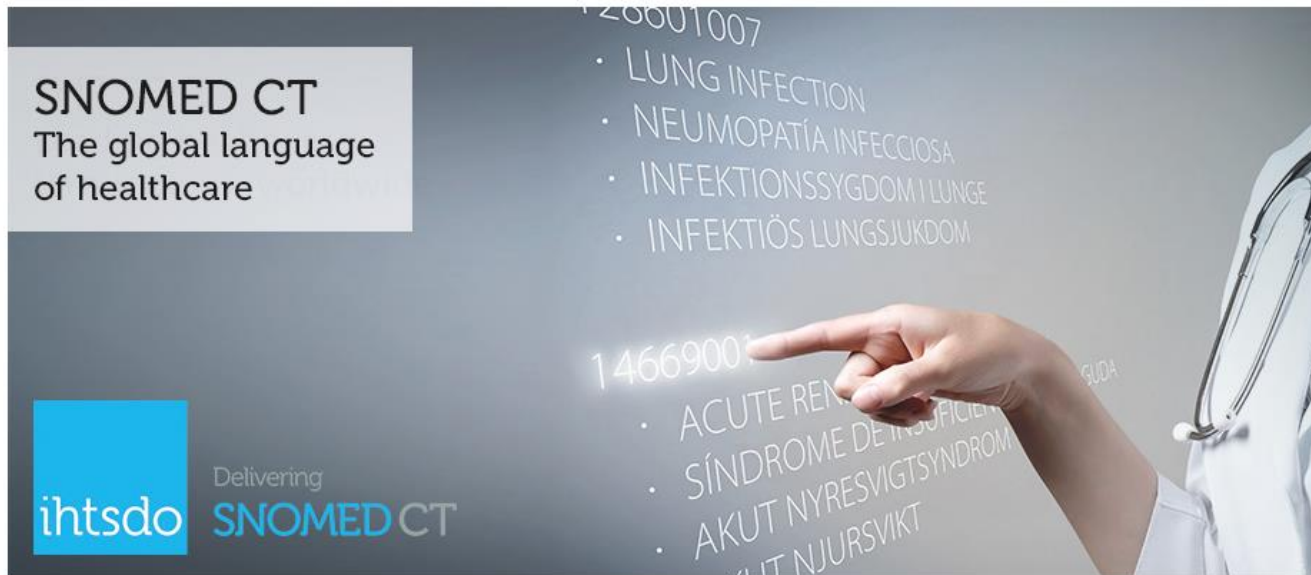


Home

IHTSDO

SNOMED CT

Participate



Welcome to IHTSDO

The International Health Terminology Standards Development Organisation determines global standards for health terms, an essential part of improving the health of humankind.

<http://www.ihtsdo.org/>

SNOMED CT as terminology



Myocardial infarction (disorder)



SCTID: 22298006

22298006 | Myocardial infarction (disorder) |

Code +
Fully Specified Name

Myocardial infarction (disorder)

Cardiac infarction

Heart attack

Infarction of heart

MI - Myocardial infarction

Myocardial infarct

Myocardial infarction

Synonyms

SNOMED CT as ontology

Parents

- ≡ Ischemic heart disease (disorder)
- ≡ Myocardial disease (disorder)
- ≡ Myocardial necrosis (finding)
- ≡ Necrosis of anatomical site (disorder)

≡ Myocardial infarction (disorder)

SCTID: 22298006

22298006 | Myocardial infarction (disorder) |

Multiple subclass
hierarchies (is-a)

Associated morphology → Infarct
Finding site → Myocardium structure

Relations (OWL object properties):

e.g.

Associated morphology

Associated procedure

Finding site

Ontology axioms:

$C_1 - Rel - C_2$ triples interpreted as:

(FOL) $\forall x: instanceOf(x, C_1) \Rightarrow$
 $\exists y: instanceOf(C_2) \wedge Rel(x, y)$

(DL) C_1 subclassOf **Rel** some C_2

Open Biomedical Ontology (OBO) Foundry



Open Biomedical Ontology (OBO) Foundry

- Suite of orthogonal interoperable reference ontologies in the biomedical domain



<u>Title</u>	<u>Domain</u>	<u>Prefix</u>
Biological process	biological process	GO
Cellular component	anatomy	GO
Chemical entities of biological interest	biochemistry	CHEBI
Molecular function	biological function	GO
Ontology for biomedical investigations	experiments	OBI
Phenotypic quality	phenotype	PATO
Plant Ontology	anatomy and development	PO
PRotein Ontology (PRO)	proteins	PR
Xenopus anatomy and development	anatomy	XAO
Zebrafish anatomy and development	anatomy	ZFA

Open Biomedical Ontology (OBO) Foundry

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)

Upper Level Ontologies

- Strict categorization through limited set of top classes and relations
- Examples: DOLCE, BFO, SSIO, UFO, GFO, SUMO, BioTopLite

Classes

- ▶ Disposition
- ▶ Function
- ▶ Immaterial object
- ▶ Information object
- ▶ Material object
- ▶ Process
- ▶ Quality
- ▶ Role
- ▶ Temporal region
- ▶ Value region

Relations

- ▶ at some time
- ▶ includes
 - ▶ has part
 - ▶ has boundary
 - ▶ has granular part
 - ▶ has component part
- ▶ is bearer of
- ▶ causes
 - ▶ has realization
- ▶ precedes
- ▶ has condition
- ▶ projects onto
- ▶ has participant
 - ▶ has agent
 - ▶ has patient
 - ▶ has outcome
- ▶ is life of
- ▶ is referred to at time
- ▶ represents

3. Contribute to develop existing standards / specifications

- Join communities that use common terminology / ontology specifications
- Contribute to development / maintenance
- Ontologies
 - objective descriptions of a domain and not as application-specific knowledge bases (scientific realism*)
 - Only express what is universally true
- Examples
 - SNOMED CT
 - OBO Foundry
 - Upper-level ontologies (BFO, DOLCE, BioTop)



Adaptation of existing standards / specifications

- Create extensions of existing semantic resources
 - Additional subclasses, interface terms
- Address specific use cases / contexts
 - Add additional upper-level orderings, e.g. "Indication", "Phenotype", "Clinical Problem", "Target", orthogonal to existing top-level
 - Refine ambiguous classes like *Animal*, *Tree*, *Heart*
 - animal (biological) vs. animal (legal)
 - tree (morphology) vs. tree (taxonomic) vs. tree (growth pattern)
 - heart (anatomical) vs. heart (surgical)

Conclusion

- Semantic resources for Life Sciences: Large number, large heterogeneity (context, quality, formalisms)
- How to make best use of them?
 - Linked Data / "little semantics" large-scale re-use only where low precision is tolerable
 - Else: Building on a limited number of high-quality terminology standards / specification efforts, join communities, custom additions / refinements
- Refrain from building "yet another" ontology
- Value semantic interoperability

Thank you



Stefan Schulz

(Univ.-Prof. Dr. med.)

Institut für Medizinische Informatik, Statistik und Dokumentation
Medizinische Universität Graz, Auenbruggerplatz 2/V
8036 Graz (Austria)

<http://www.medunigraz.at/imi>

<http://g.co/maps/aqedt>

0043 316 385 16939

0043 316 385 13201

<http://purl.org/steschu>

<mailto:stefan.schulz@medunigraz.at>

Skype: stschulz