Stefan Schulz
University Medical Center, Freiburg, Germany
Elena Beisswanger
Language and Information Engineering Lab, Jena, Germany
Olivier Bodenreider
National Library of Medicine, Bethesda, MD, USA
László van den Hoek
Erik M. van Mulligen
Erasmus Medical Center, Rotterdam, The Netherlands

Alignment of the UMLS Semantic Network with BioTop

Methodology and Assessment

Ontology Alignment

- Linking two ontologies by detecting semantic correspondences between their representational units
- Types of correspondences:
  equivalence, subsumption, others
- Purpose of ontology alignment:
  - Creating interoperability between semantically annotated data
  - Enriching semantics
  - Cross-Validation of ontologies
- Requirements of ontology alignment:
  - comparable scope
  - comparable context
  - comparable semantic foundations

Outline

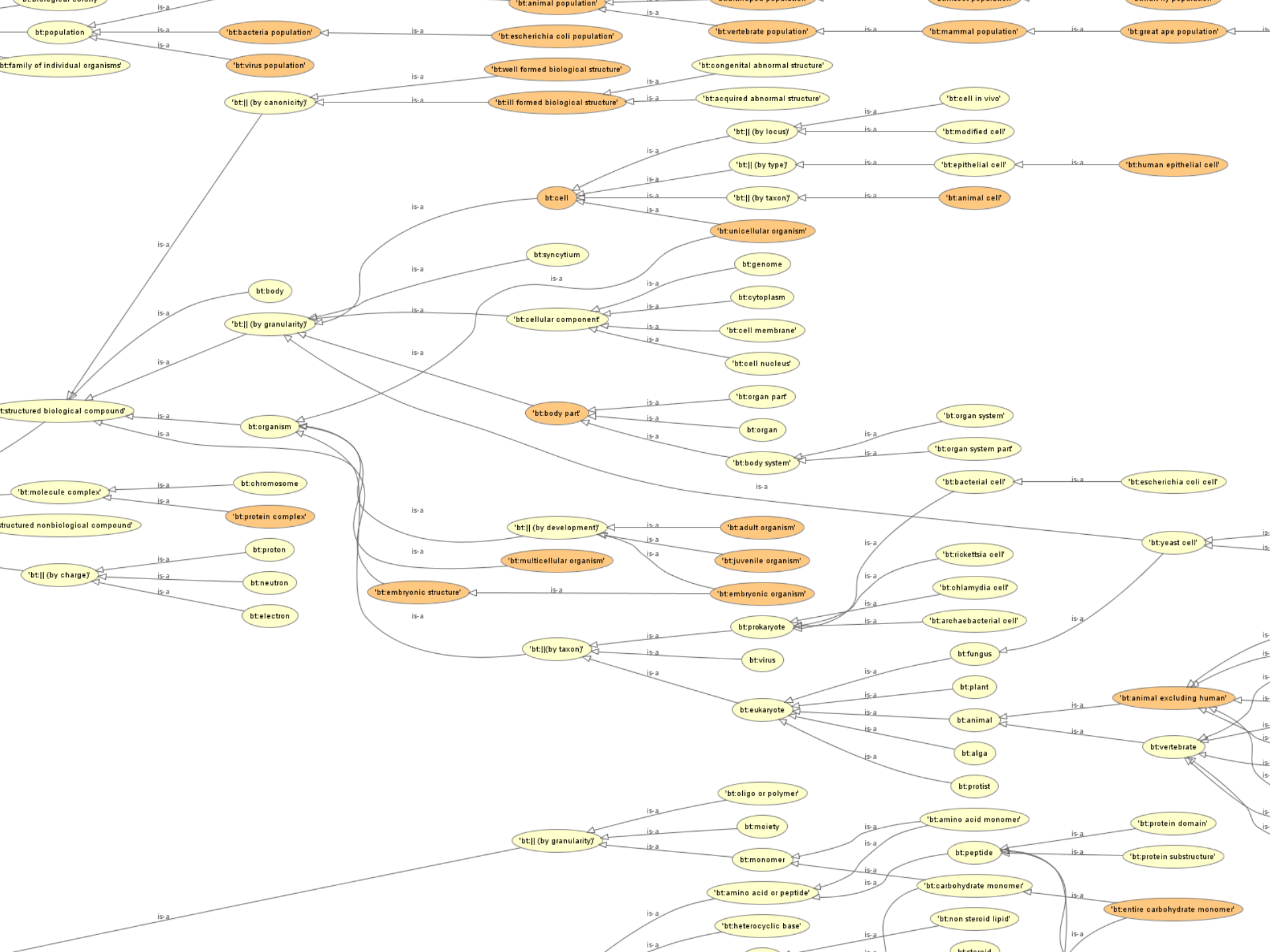
- Introduction
  - BioTop
  - UMLS SN
- Methodology
  - UMLS SN: formal redefinition
  - Interactive Mapping
- Assessment
  - Ontology Cross-Validation
  - NE co-occurrence validation
  - UMLS SN cluster consistency
- Conclusion

Outline

- **Introduction**
  - BioTop
  - UMLS SN
- Methodology
  - UMLS SN: formal redefinition
  - Interactive Mapping
- Assessment
  - Ontology Cross-Validation
  - NE co-occurrence validation
  - UMLS SN cluster consistency
- Conclusion

BioTop – a Life Science Upper Ontology

- Recent development (starting 2006, Freiburg & Jena)
- Goal: to provide formal definitions of upper-level types and relations for the biomedical domain
- Uses description logics (OWL-DL)
  - 339 classes, 60 relation types
  - 373 subclass axioms
  - 80 equivalent class axioms, 66 disjoint class axioms
- Compatible with BFO and DOLCE lite
- links to OBO ontologies
- downloadable from: http://purl.org/biotop

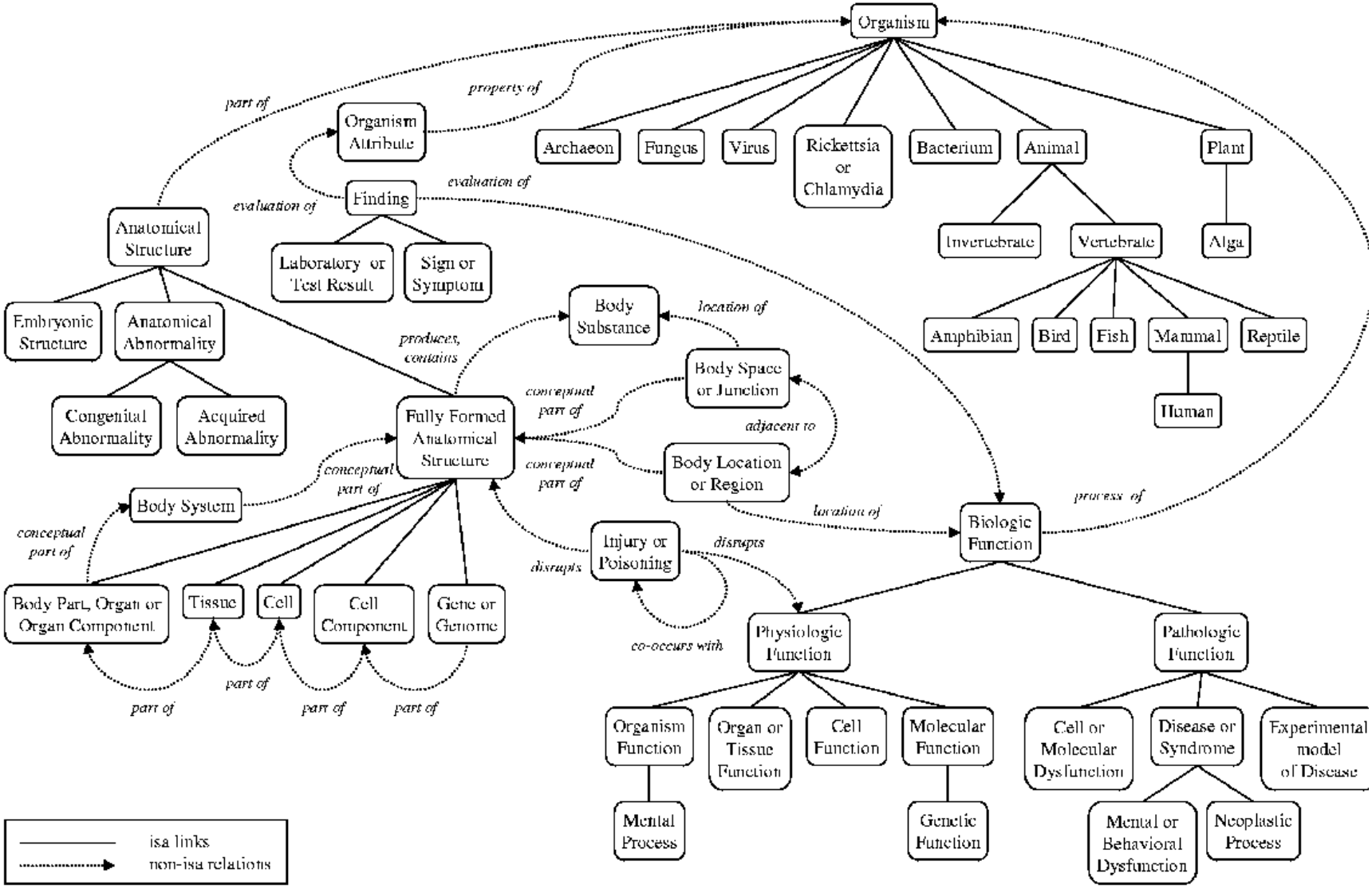

bt:population
'bt:bacteria population'
'bt:escherichia coli population'
'bt:virus population'
'bt:family of individual organisms'
'bt:animal population'
'bt:vertebrate population'
'bt:mammal population'
'bt:great ape population'
'bt:well formed biological structure'
'bt:congenital abnormal structure'
'bt:|| (by canonicity)'
'bt:ill formed biological structure'
'bt:acquired abnormal structure'
'bt:|| (by locus)'
'bt:cell in vivo'
'bt:modified cell'
'bt:|| (by type)'
'bt:epithelial cell'
'bt:human epithelial cell'
bt:cell
'bt:|| (by taxon)'
'bt:animal cell'
'bt:unicellular organism'
bt:syncytium
bt:genome
bt:cytoplasm
'bt:cellular component'
'bt:cell membrane'
'bt:cell nucleus'
bt:body
'bt:|| (by granularity)'
'bt:body part'
'bt:organ part'
bt:organ
'bt:body system'
'bt:organ system'
'bt:organ system part'
t:structured biological compound'
bt:organism
'bt:bacterial cell'
'bt:escherichia coli cell'
'bt:molecule complex'
bt:chromosome
'bt:protein complex'
structured nonbiological compound'
'bt:|| (by development)'
'bt:adult organism'
'bt:juvenile organism'
'bt:multicellular organism'
'bt:embryonic structure'
'bt:embryonic organism'
'bt:yeast cell'
'bt:rickettsia cell'
'bt:chlamydia cell'
'bt:archaebacterial cell'
bt:proton
'bt:|| (by charge)'
bt:neutron
bt:electron
'bt:prokaryote'
'bt:||(by taxon)'
bt:virus
bt:fungus
bt:plant
'bt:eukaryote'
'bt:animal excluding human'
bt:animal
bt:vertebrate
bt:alga
bt:protist
'bt:oligo or polymer'
'bt:|| (by granularity)'
bt:moiety
bt:monomer
'bt:amino acid monomer'
'bt:protein domain'
bt:peptide
'bt:protein substructure'
'bt:carbohydrate monomer'
'bt:amino acid or peptide'
'bt:entire carbohydrate monomer'
'bt:non steroid lipid'
'bt:heterocyclic base'
is-a

UMLS Semantic Network (SN)

- Upper-level semantic categorization framework for all (~1 M) concepts of the **UMLS Metathesaurus**
- Tree of 135 semantic types (e.g. *Tissue*, *Diagnostic_Procedure*)
- 53 associative relationships (e.g., *treats*, *location_of*)
- 612 relational assertions (triples), sanctioning the domain and range of relations {*Tissue*; *location_of*; *Diagnostic_Procedure*}
- mainly unchanged in the last 20 years

Unified Medical Language System (UMLS): Metathesaurus links over 100 biomedical vocabularies

UMLS Semantic Network (SN)



Comparison UMLS-SN - BioTop

	UMLS-SN	BioTop
Types / Classes	135	339
Relation Types	53	60 (object properties)
Axioms	612	509
Semantics	Implicit Frame-like Closed-world (?)	Explicit (description logics) Set-theoretic Open-world
Class subsumption $\sqsubseteq$	+	+
Relation subsumption $\sqsubseteq$	+	+
Domain / Range Restrictions	+	+
Relation Inheritance blocking	+	–
Full Definitions $\equiv$	–	+
Disjoint Partitions	–	+
Negations $\neg$	–	+
Existential Restrictions $\exists$	–	+
Value Restrictions $\forall$	–	+

Outline

- Introduction
  - BioTop
  - UMLS SN
- Methodology
  - UMLS SN: formal redefinition
  - Interactive Mapping
- Assessment
  - Ontology Cross-Validation
  - NE co-occurrence validation
  - UMLS SN cluster consistency
- Conclusion

Outline

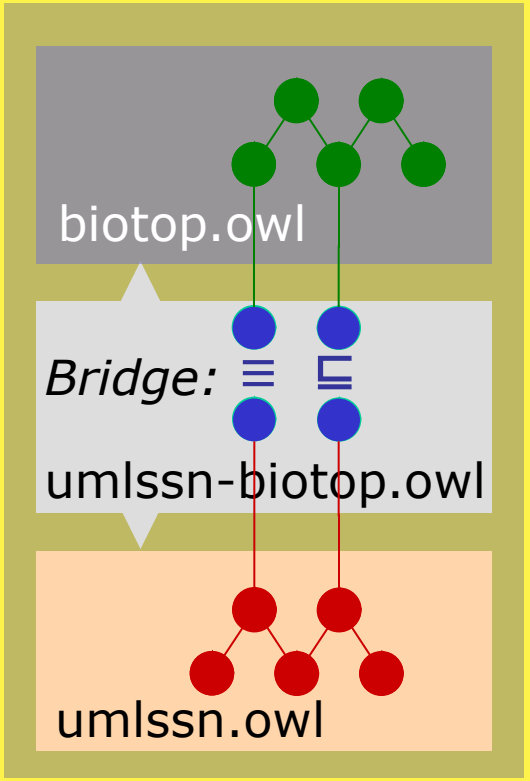
- Introduction
  - BioTop
  - UMLS SN
- **Methodology**
  - UMLS SN: formal redefinition
  - Interactive Mapping
- Assessment
  - Ontology Cross-Validation
  - NE co-occurrence validation
  - UMLS SN cluster consistency
- Conclusion

Methodology

1. Prerequisite: provide description logics semantics to the UMLS SN:

   umlssn.owl

2. Building a bridging ontology
   - Subsumption $\sqsubseteq$
   - Equivalence $\equiv$



Redefinition of UMLS SN semantics

Redefinition of UMLS SN semantics

- **Semantic Types**, e.g.: *Tissue*, *Diagnostic_Procedure*:
  - Types extend to classes of individuals
  - subsumption hierarchies = is-a hierarchies (every instance of a child is also an instance of each parent)
  - no explicit disjoint partitions
- **Semantic Relations**, e.g.: *treats*, *location_of*:
  - Reified as classes, **not** represented as OWL object properties
- **Triples**, e.g.: {*Tissue*; *location_of*; *Diagnostic_Procedure*}
  - domain and range restrictions = value restrictions on the roles *has-domain* and *has-range*

UMLS SN: Why SRs as classes …

and not OWL object properties? (I)

treats	Range	Disease	Person
Domain			
Drug		allowed	disallowed
Physician		disallowed	allowed

$$\mathit{TreatingPerson} \equiv \mathit{Action} \sqcap \exists\ \textbf{has\_domain}.\mathit{Physician} \sqcap \exists\ \textbf{has\_range}.\mathit{Person} \sqcap \forall\ \textbf{has\_domain}.\ \mathit{Physician} \sqcap \forall\ \textbf{has\_range}.\mathit{Person}$$

$$\mathit{TreatingDisease} \equiv \mathit{Action} \sqcap \exists\ \textbf{has\_domain}.\mathit{Drug} \sqcap \exists\ \textbf{has\_ range}.\mathit{Disease} \sqcap \forall\ \textbf{has\_domain}.\mathit{Drug} \sqcap \forall\ \textbf{has\_ range}.\ \mathit{Disease}$$

$$\mathit{Treating} \equiv \mathit{TreatingPerson} \sqcup \mathit{TreatingDisease}$$

UMLS SN: Why SRs as classes ..

and not OWL object properties? (II)

- Source Representation

Idea_or_Concept *conceptual_part_of* *Behavior*

"Defined not Inherited"

- Target Representation

$$Conceptual\_part\_of\_Domain\_Idea\_Or\_Concept\_Range\_\ Behavior\_Rest\_Class \sqsubseteq Conceptual\_part\_of \sqcap \forall has\_domain.\ Idea\_Or\_Concept\_Rest\_Class \sqcap \forall has\_range.\ Behavior\_Rest\_Class$$

$$Idea\_Or\_Concept\_Rest\_Class \equiv Idea\_Or\_Concept \sqcap \neg Temporal\_Concept \sqcap \neg Qualitative\_Concept \sqcap \neg Quantitative\_Concept \sqcap \neg Spatial\_Concept \sqcap \neg Functional\_Concept$$

$$Behavior\_Rest\_Class \equiv Behavior \sqcap \neg Individual\_Behavior \sqcap \neg Social\_Behavior$$

Representation of SRs and triples

- All triples including R are defined as subclasses of R

$$Affects\_Domain\_Cell\_Component\_Range\_Physiologic\_Function \sqsubseteq Affects \sqcap \forall has\_domain.\ Cell\_Component \sqcap \forall has\_range.\ Physiologic\_Function$$

- All parents are fully defined by the union of their children

$$Brings\_About \equiv Produces \sqcup Causes$$

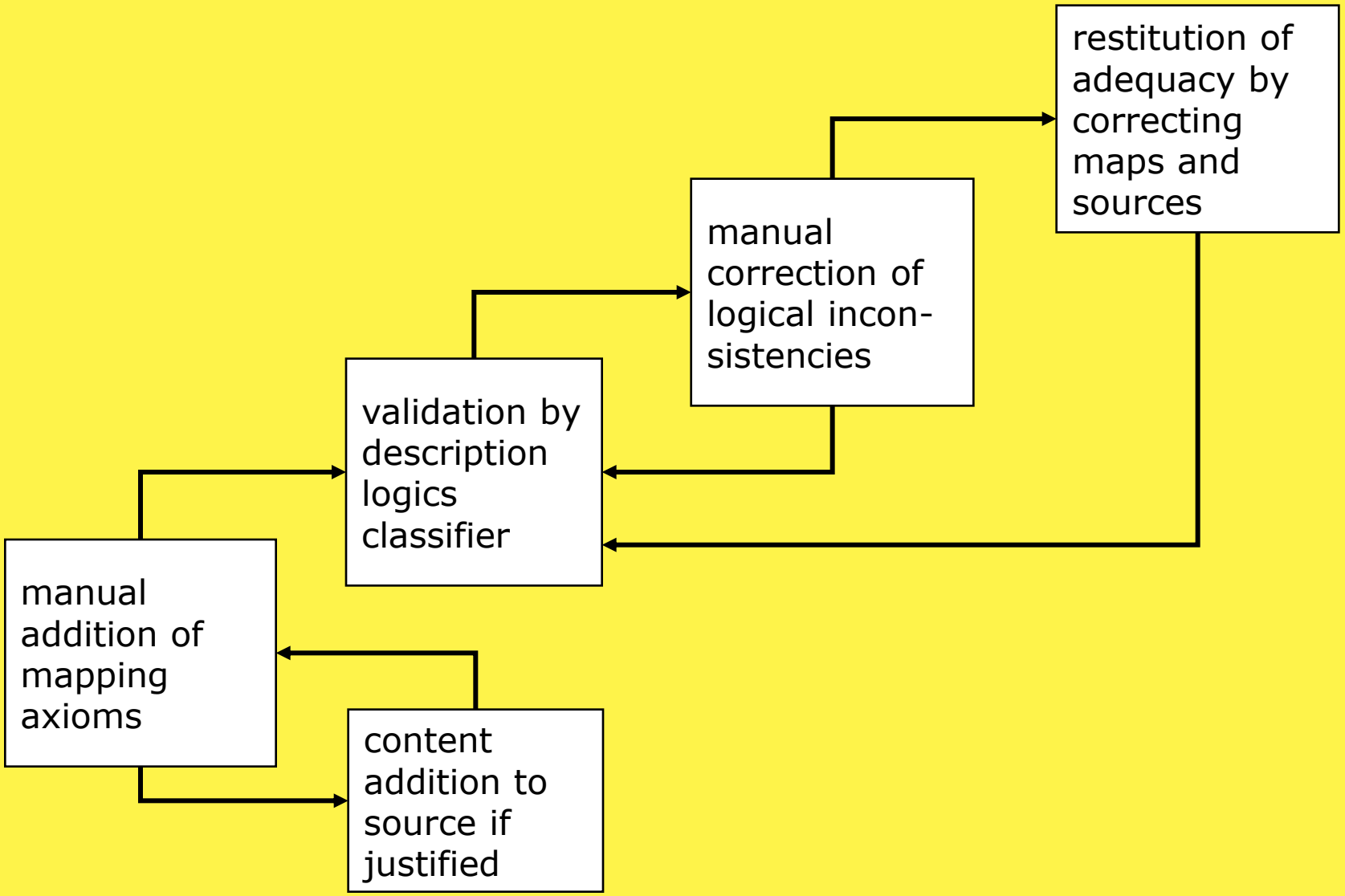
Mapping

Mapping

- Fully manually, using Protégé 4, consistency check with Fact++ and Pellet 1.5, supported by explanation plugin*
- Analyzing
  - UMLS SN hierarchies and free-text definitions
  - BioTop formal and free-text definitions
- Iterative check of
  - logic consistency (DL classifier)
  - domain adequacy (analysis of new entailments)

*(Horridge ISWC 2008)

Mapping workflow



Mapping of UMLS Types

- Direct Match (often after content addition to BioTop):

  sn:Plant ≡ bt:Plant

- Restriction mapping:

  sn:AnatomicalAbnormality ≡ bt:OrganismPart ⊓ ∃□ bt:bearerOf.bt:PathologicalCondition

- Union:

  sn:Gene_Or_Genome ≡ bt:Gene ⊔ bt:Genome.

- Out of scope

  sn:Daily_Or_Recreational_Activity ⊑ bt:Action ⊓ ∃□ bt:hasParticipant.bt:Human

- No mapping

  sn:Idea_or_concept

Mapping of UMLS Relations

- Mapping of domain and range

$$sn{:}hasDomain \equiv bt{:}hasAgent$$
$$sn{:}hasRange \equiv bt{:}hasPatient$$

- Mapping of (reified) SN relations

$$sn{:}Affects \equiv bt{:}Affecting$$

- Linkage of (reified) SN relations to BioTop relations by augmented restrictions:

$$sn{:}hasDomain \sqsubseteq \forall (bt{:}physicalPartOf \sqsubseteq \forall (ImmaterialPhysicalEntity \sqcup MaterialEntity)) \sqcap$$
$$sn{:}hasRange \sqsubseteq \forall (bt{:}hasPhysicalPart \sqsubseteq \forall (ImmaterialPhysicalEntity \sqcup MaterialEntity))$$

Outline

- Introduction
  - BioTop
  - UMLS SN
- Methodology
  - UMLS SN: formal redefinition
  - Interactive Mapping
- Assessment
  - Ontology Cross-Validation
  - NE co-occurrence validation
  - UMLS SN cluster consistency
- Conclusion

Outline

- Introduction
  - BioTop
  - UMLS SN
- Methodology
  - UMLS SN: formal redefinition
  - Interactive Mapping
- **Assessment**
  - Ontology Cross-Validation
  - NE co-occurrence validation
  - UMLS SN cluster consistency
- Conclusion

Assessment: Cross-evaluation

- Formative evaluation of BioTop: Mapping and subsequent classification unveils hidden problems in BioTop:
  - Faulty disjointness axioms (e.g. *bt:Organic Chemical* was disjoint from *bt:Carbohydrate*)
  - ambiguities: Sequence as information entity vs. sequence as molecular structure
  - granularity mismatches:
    e.g. Chromosome as molecule

Assessment: NE co-occurrences

- Named Entity tagging, UMLS concept pairs identified in 15 M PubMed abstracts

Semantic Type 1: UMLS ID	NE 1	**Semantic Type 2**: UMLS ID	NE 2
Enzyme:C0916840	superoxide reductase	**Organic_Chemical**:C0001992	aldehyde
Finding:C0883391	free testosterone index	**Laboratory_Procedure**:C0020980	immunoassay
Food:C1145642	sorghum	**Invertebrate**:C0009276	beetles
Functional_Concept:C0332240	idiopathic	**Pharmacologic_Substance**:C0011685	desipramine
Functional_Concept:C1510670	feeds	**Intellectual_Product**:C0023683	life table
Gene_or_Genome:C0087142	v-Jun	**Mammal**:C0025920	C3H
Gene_or_Genome:C0600449	essential gene	**Hazardous_or_Poisonous_Substance**:C0000511	4-nitroquinolone-1-oxide
Geographic_Area:C0027978	New Zealand	**Idea_or_Concept**:C0018741	health resources
Hazardous_or_Poisonous_Substance:C0036248	stx	**Organic_Chemical**:C0000967	acetal

- Expert rating with sample of co-occurrences: which are semantically related?

Assessment: NE co-occurrences

		Expert judgment: should be related (52)	Expert judgment: Should not be related (93)
matching against SN triplets	SN: sanctioned	31	22
	SN: unsanctioned	21	71
Description logics classification	SN-BioTop: accepted	52	90
	SN-BioTop: rejected	0	3

- Using SN alone: very low agreement with expert rating
- Using SN+BioTop: very few rejections (only 3)
- Reasons:
  - false-positive rate: Expert rating done on NE (e.g. *Superoxide reductase unrelated with Aldehyde*), but system judgments at type level: *sn:Enzyme* related to *sn:Organic Chemical*
  - few rejections: DL's open world semantics

Assessment: finding incompatible semantic types

- Each UMLS concept is categorized by one or more UMLS SN types
- 397 different SN type combinations
- Using UMLS-SN BioTop Bridge: 133 combinations inconsistent, affecting 6116 UMLS concepts
- Main reason: hidden ambiguities, e.g.

$$sn{:}Manufactured\ Object \sqcap sn{:}HealthCareRelatedOrganization$$

(e.g. *Hospital* as building vs. organization).

Outline

- Introduction
  - BioTop
  - UMLS SN
- Methodology
  - UMLS SN: formal redefinition
  - Interactive Mapping
- Assessment
  - Ontology Cross-Validation
  - NE co-occurrence validation
  - UMLS SN cluster consistency
- Conclusion

Outline

- Introduction
  - BioTop
  - UMLS SN
- Methodology
  - UMLS SN: formal redefinition
  - Interactive Mapping
- Assessment
  - Ontology Cross-Validation
  - NE co-occurrence validation
  - UMLS SN cluster consistency
- **Conclusion**

Conclusion

- Sucessful alignment between the (legacy) SN and the (novel) BioTop ontology
- Necessary: formal re-interpretation of SN
- Prospect: join large amount of data annotated by the SN with formal rigor of BioTop
- Strength: machine inference, consistency checking
- Challenge: Antagonize unwarranted effects of the open world semantics by making exhaustive use of disjoint partitions
- More use cases !

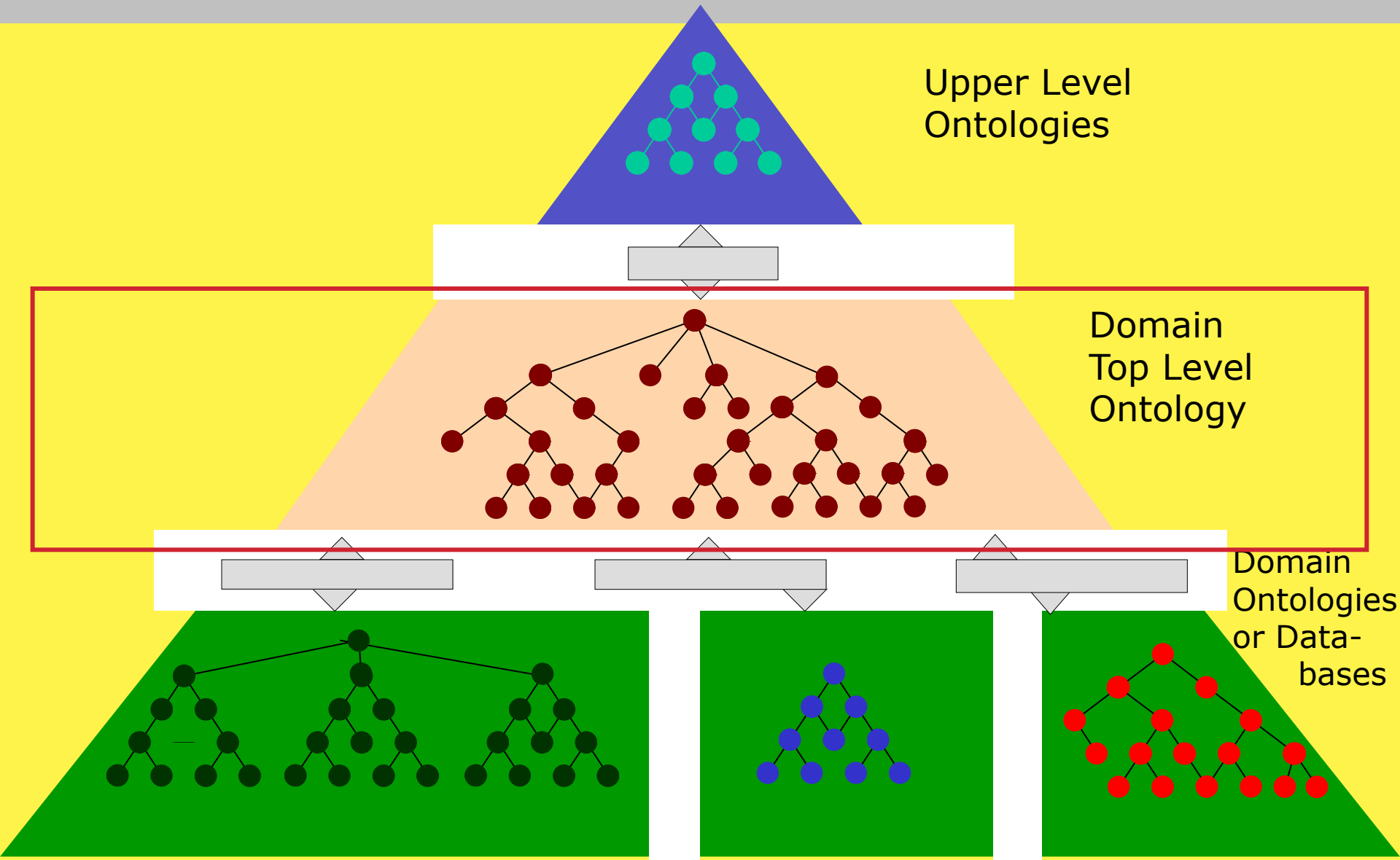
Acknowledgements

- EC STREP project “BOOTStrep” (FP6 – 028099)
- Intramural Research Program of the National Institutes of Health (NIH), US National Library of Medicine
- Martin Boeker (Freiburg)
- Holger Stenzhorn (Freiburg)
- Anonymous Reviewers

Stefan Schulz
University Medical Center, Freiburg, Germany
Elena Beisswanger
Language and Information Engineering Lab, Jena, Germany
Olivier Bodenreider
National Library of Medicine, Bethesda, MD, USA
László van den Hoek
Erik M. van Mulligen
Erasmus Medical Center, Rotterdam, The Netherlands

Alignment of the UMLS Semantic Network with BioTop *Methodology and Assessment*

Ontology Stack



The Semantic Network of the UMLS