From Symptoms to Diseases – Creating the Missing Link

Heiner Oberkampf[1;2], Turan Gojayev[1;3], Sonja Zillner[1;4], Dietlind Zühlke[5], Sören Auer[3;5], and Matthias Hammon[6]

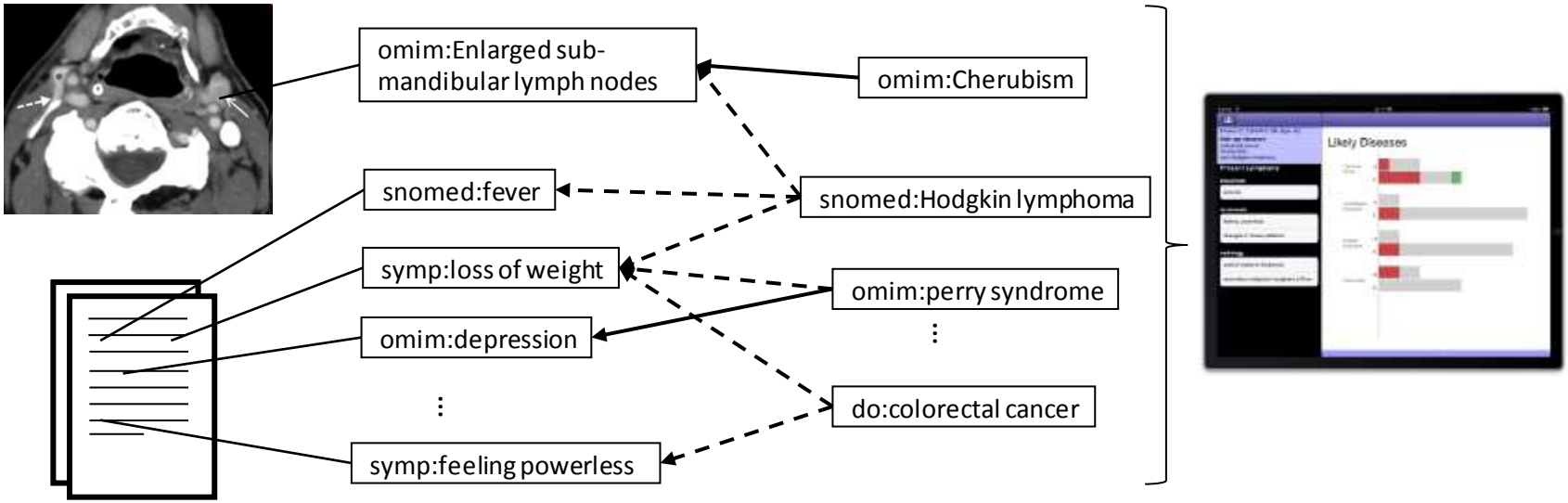
1Siemens AG, Corporate Technology, Germany
2Software Methodologies for Distributed Systems, University of Augsburg, Germany
3Institute for Applied Computer Science, University of Bonn, Germany
4School of International Business & Entrepreneurship, Steinbeis University, Germany
5Fraunhofer Institute for Intelligent Analysis and Information Systems, Germany
6Department of Radiology, University Hospital Erlangen, Germany

Disease Centric View on Patient Data

Universität Augsburg University

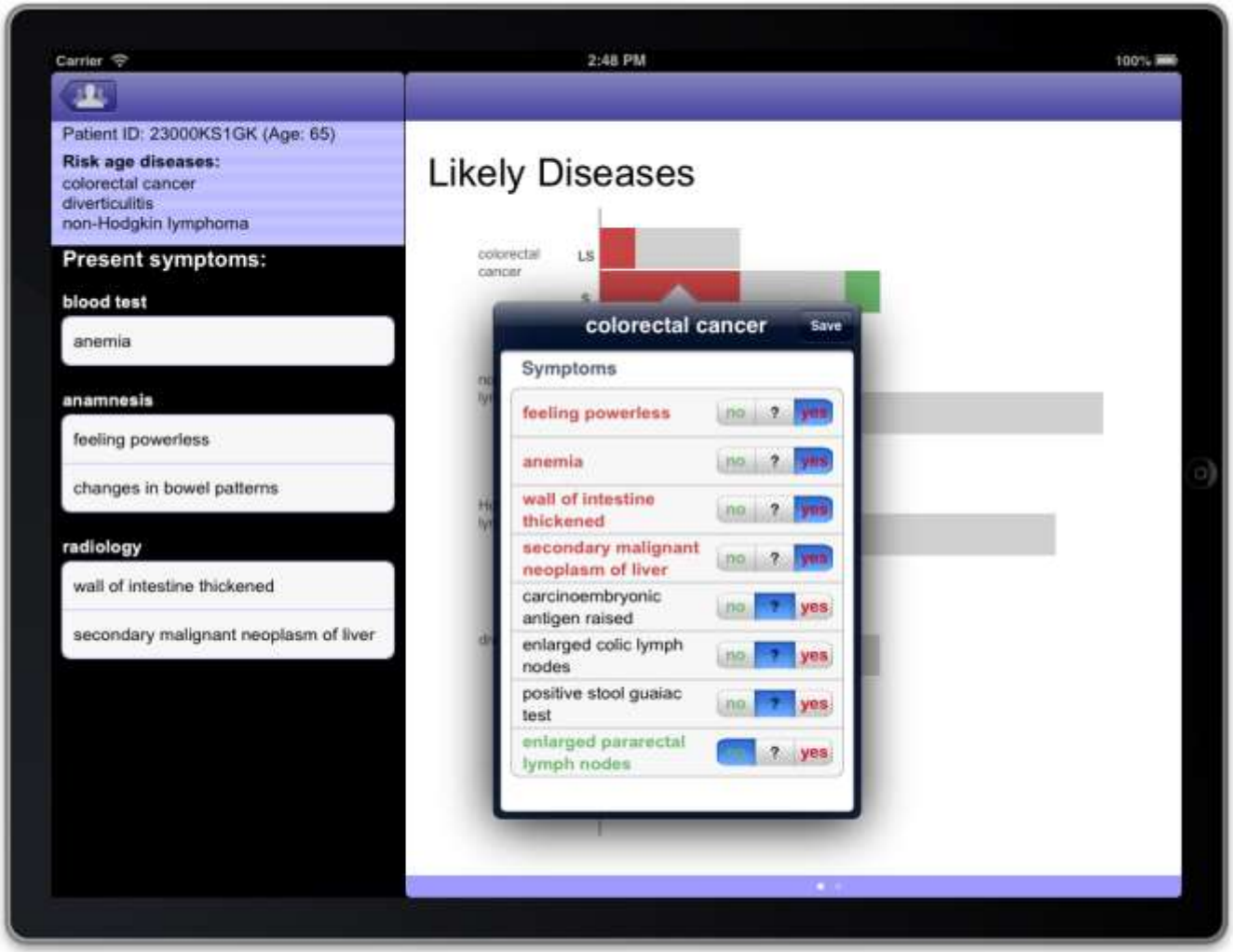
Software Methodologies for Distributed Systems

unstructured clincal data | **annotations with symptoms** | **related diseases** | **ranking of likely diseases**

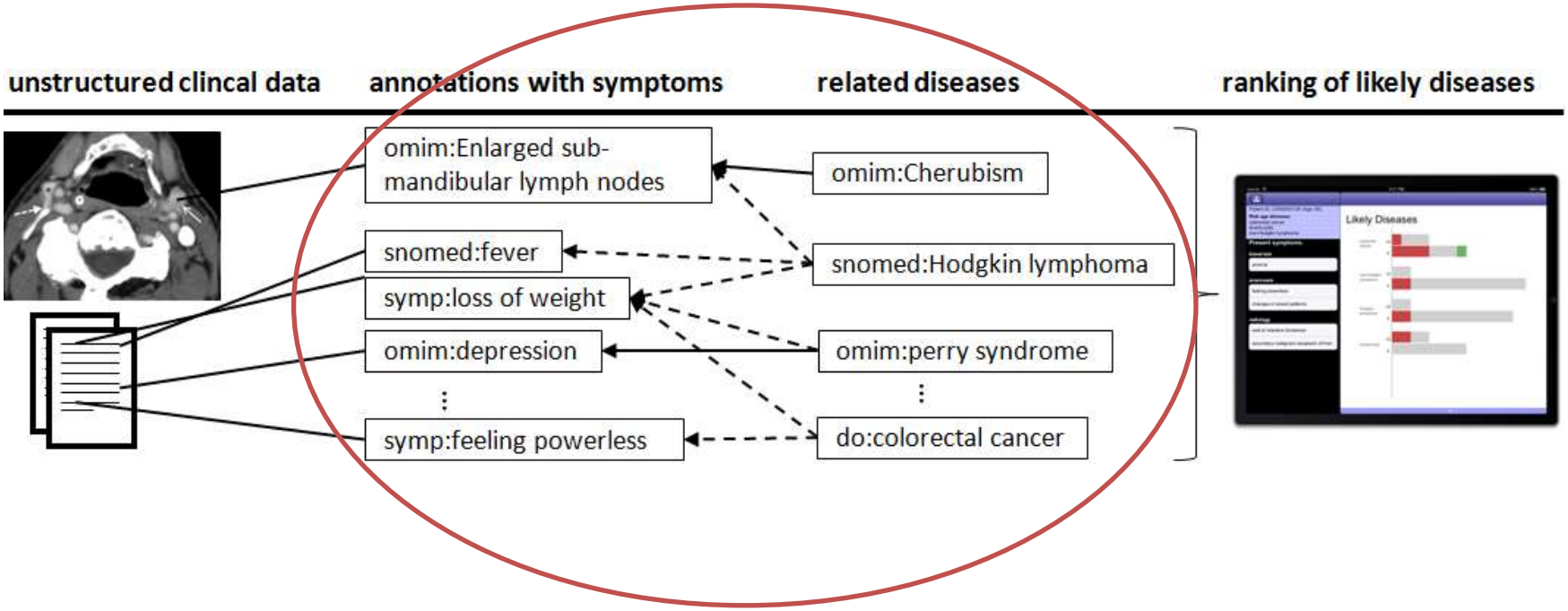


Disease Centric View on Patient Data

Disease Centric View on Patient Data



Challenge



- Disease and Symptoms *and* relations between them
- Only integrated information from *different ontologies* can provide the complete picture

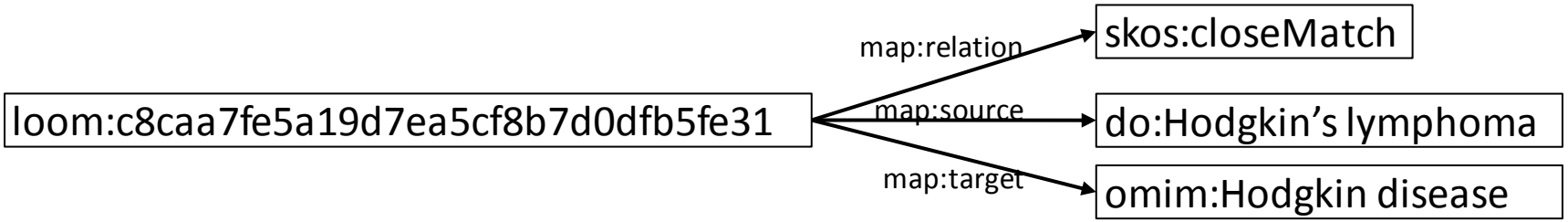
BioPortal[1]

- Ontologies
  - 400+ different ontologies
  - ~ 6 mio entities
  - 2600+ distinct object properties used in triples
- Mappings
- Metadata
- Service
  - REST-full API[2]
  - SPARQL endpoint[3]

1) http://bioportal.bioontology.org/
2) http://data.bioontology.org/
3) http://sparql.bioontology.org

BioPortal Mappings

- 6 Mapping Resources:
  - UMLS-CUI: Based on the UMLS Concept Unique Identifier
  - LOOM: Automatically generated mappings based on lexical similarity
  - OBO-XREF: Mappings for terms with same xref attribute
  - REST: User submitted mappings
  - …
- Can be used without additional preprocessing
- Mapping Types: skos:exactMatch, skos:closeMatch, skos:relatedMatch, owl:sameAs, rdfs:seeAlso



Human Disease Ontology[1]

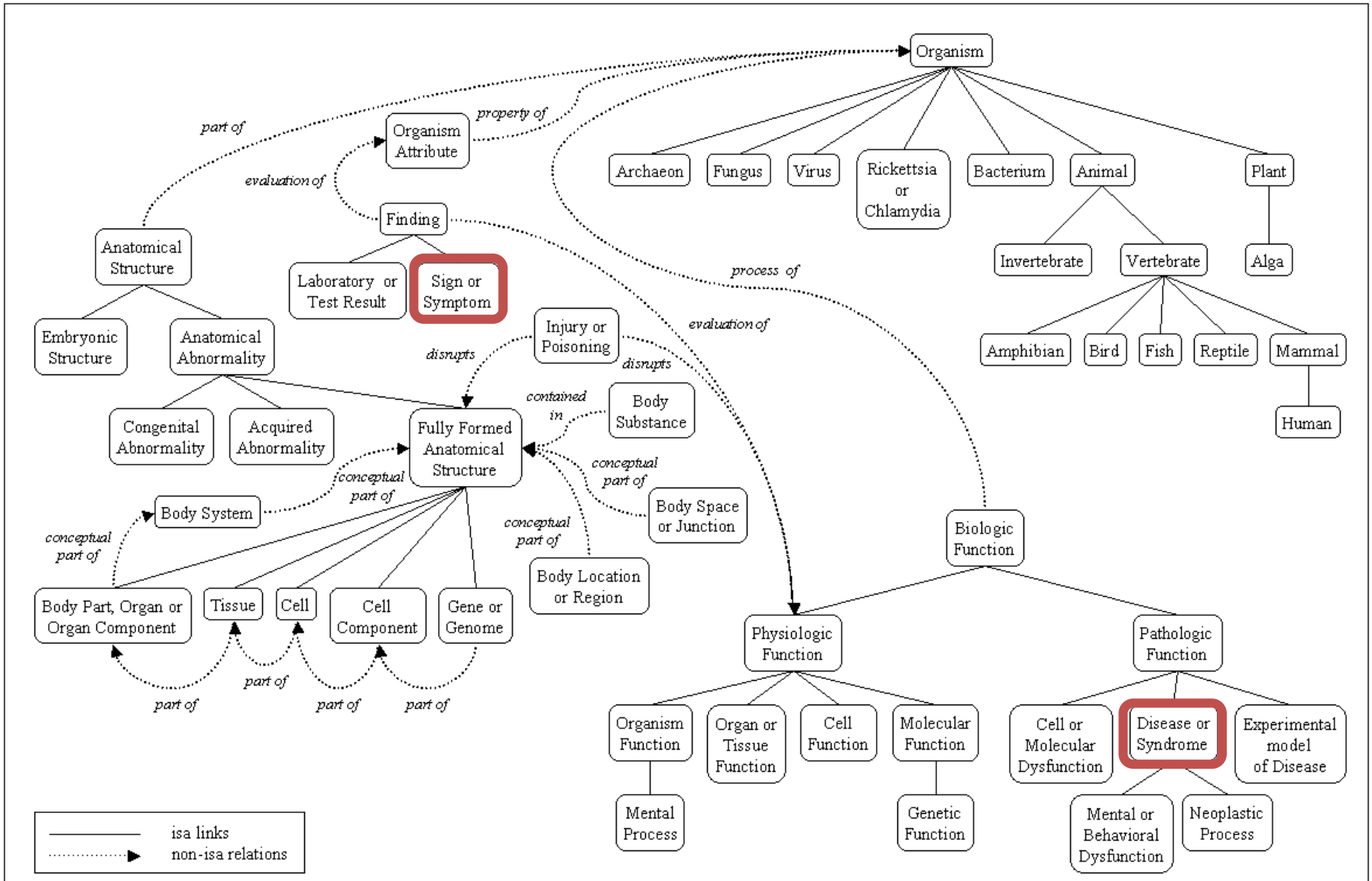
- 8,681 classes
- Utilized as reference by major biomedical databases
- Labels and textual definitions
- Part of the OBO Foundry[2]
- Available links to:
  - MeSH
  - ICD
  - NCIs thesaurus
  - SNOMED CT
  - OMIM

1) http://disease-ontology.org/
2) http://www.obofoundry.org/

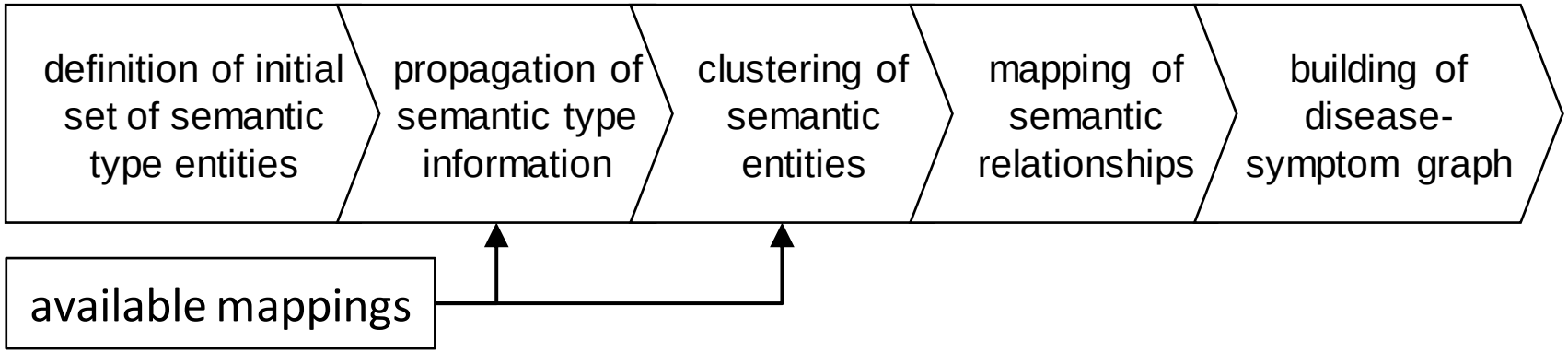
Unified Medical Language System (UMLS)

- Metathesaurus:
  - SNOMEDCT Systematized Nomenclature of Medicine – Clinical Terms
  - ICD10CM International Classification of Diseases, Version 10 – Clinical Modification
  - ICD10 International Classification of Diseases, Version 10
  - ICD9CM International Classification of Diseases, Version 9 – Clinical Modification
  - MDR Medical Dictionary for Regulatory Activities
  - RCD Read Codes, Clinical Terms Version 3 (CTV3)
  - OMIM Online Mendelian Inheritance In Man
  - ICPC2P International Classification of Primary Care – 2 Plus
  - CST Coding Symbols for a Thesaurus of Adverse Reaction Team
  - AIR Artificial Intelligence Rheumatology Consultant System Ontology
  - MEDLINEPLUS Medline Plus Health Topics
  - ICPC International Classification of Primary Care
  - MSH Medical Subject Headings
  - NDFRT National Drug File – Reference Terminology
  - …
- Semantic Network

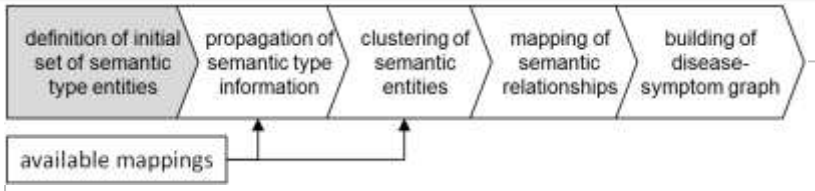
UMLS Semantic Network



Five Steps of Our Approach



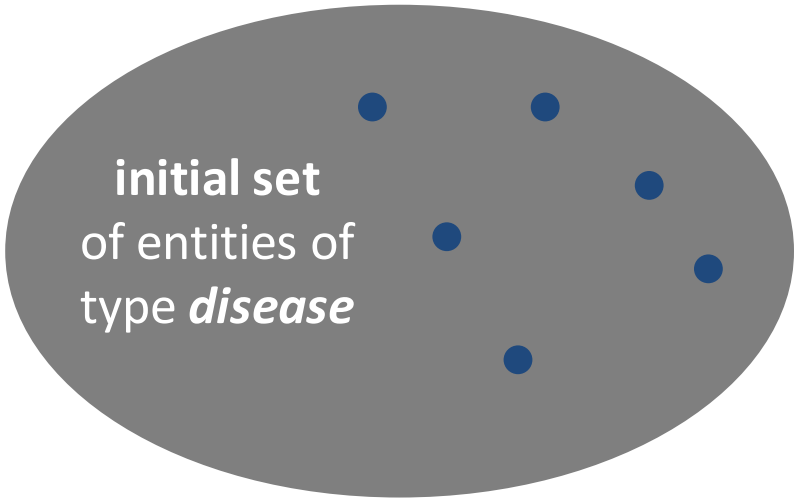
Definition of Initial Sets



Software Methodologies for Distributed Systems

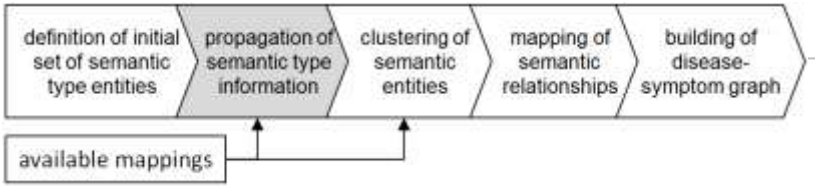
Disease entities:

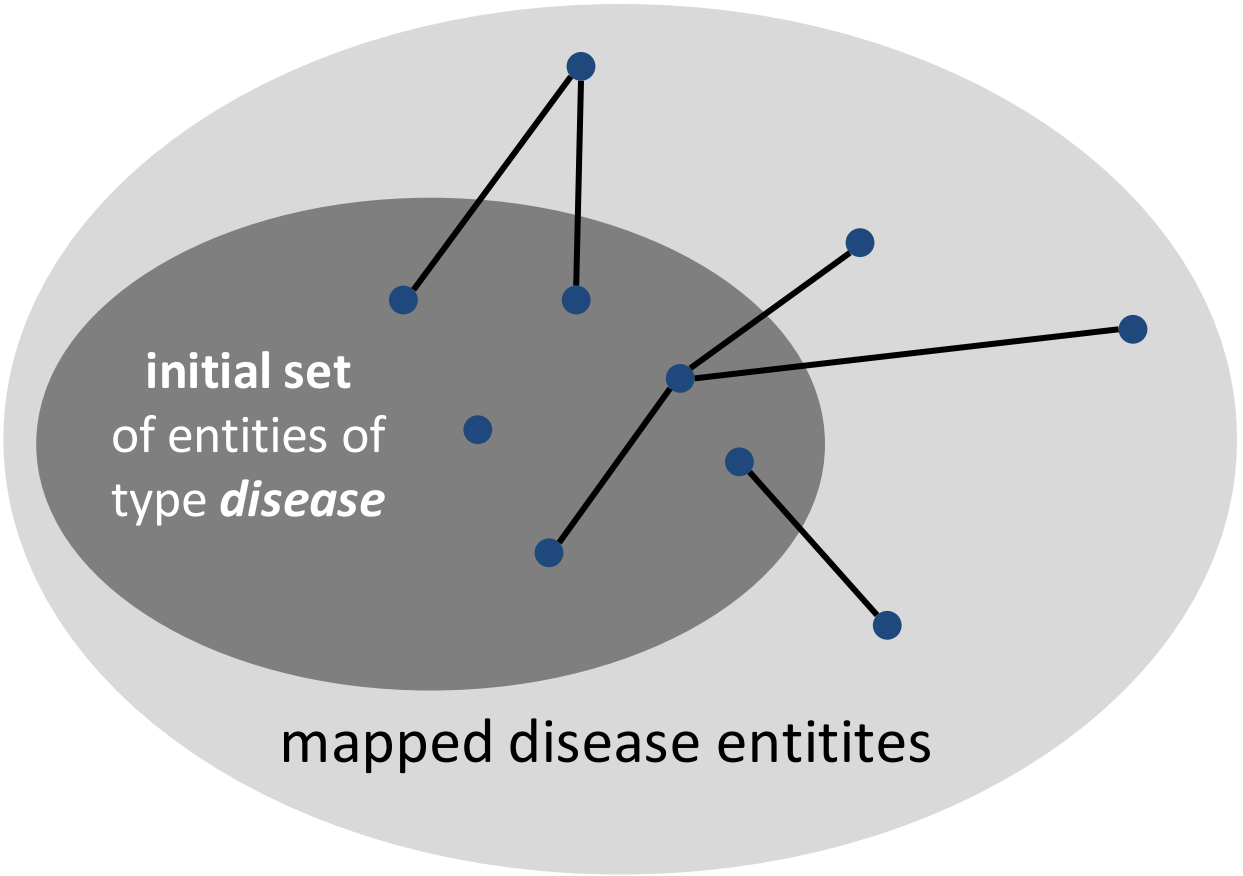
- UMLS entities of type T047 *disease or syndrome*
- Human Disease Ontology



For symptom entities we used UMLS entities of type „sign or symptom“ and the Symptom Ontology (SYMP) http://purl.obolibrary.org/symp.owl

Propagation of Semantic Types



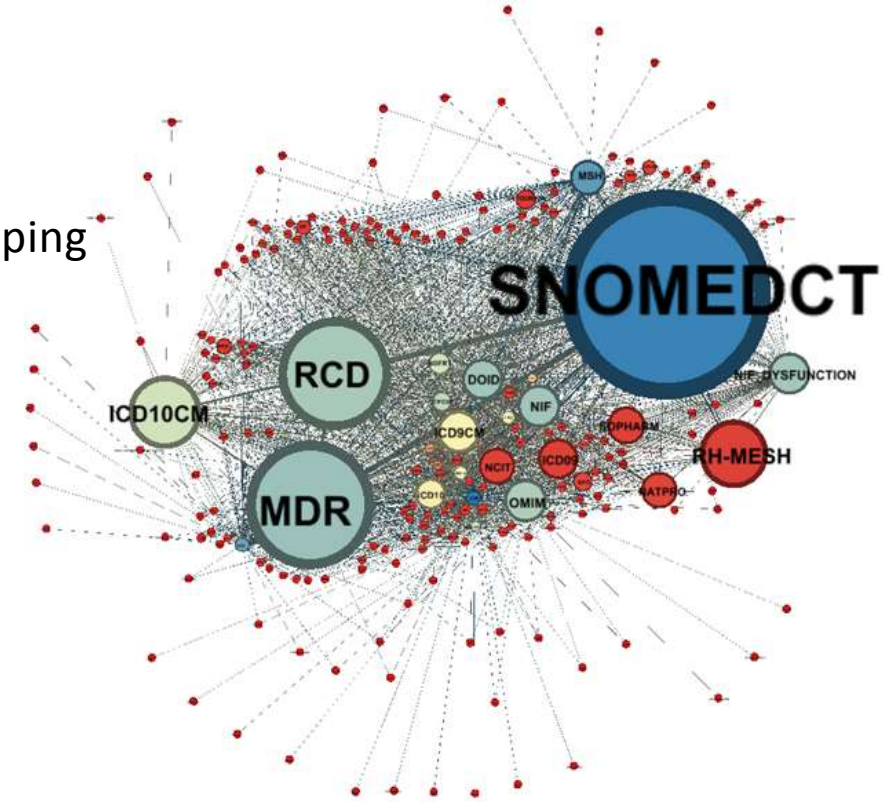


Disease Entities in BioPortal

Universität Augsburg University

Software Methodologies for Distributed Systems

- Start:
  - 18 ontologies
  - 153,223 initial disease entities
    - 123,736 have at least one mapping
- Result:
  - 219 ontologies
  - 247,683 disease entities

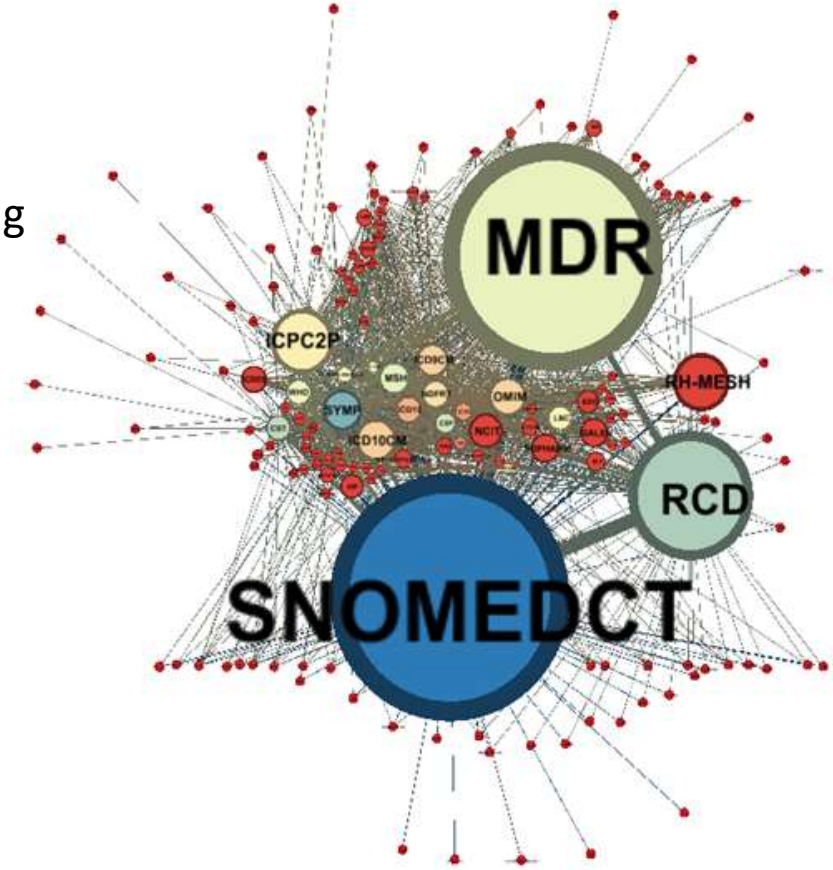


Symptom Entities in BioPortal

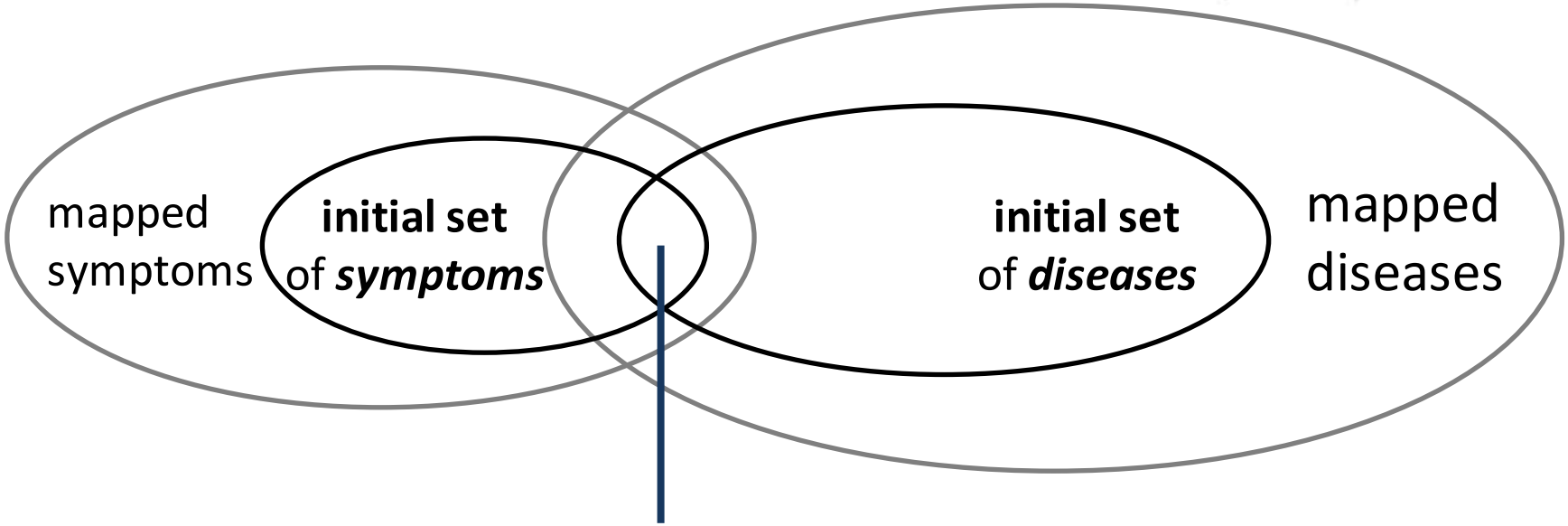
Universität Augsburg University

Software Methodologies for Distributed Systems

- Start:
  - 18 ontologies
  - 14,971 initial symptom entities
    - 11,882 have at least one mapping
- Result:
  - 161 ontologies
  - 34,088 symptom entities



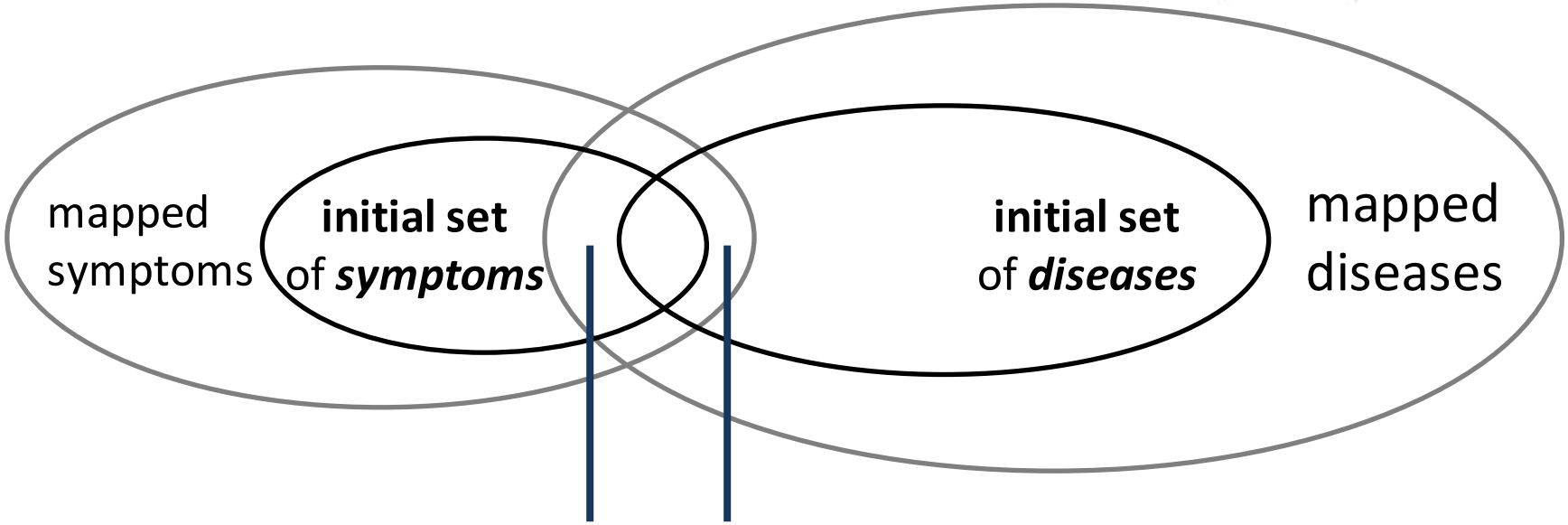
Overlap of Disease and Symptom Sets



- Manual classification of 471 classes in the overlap of initial sets by clinical expert:
  - 189 diseases: removed from symptom set (e.g. *migraine*)
  - 234 symptoms: remove from disease set (e.g. *dry mouth disorder*)
  - 48 both (not decidable): keep in both sets (e.g. *eating disorder*)

Complete classication results are available at http://goo.gl/CFgFVx

Overlap of Disease and Symptom Sets

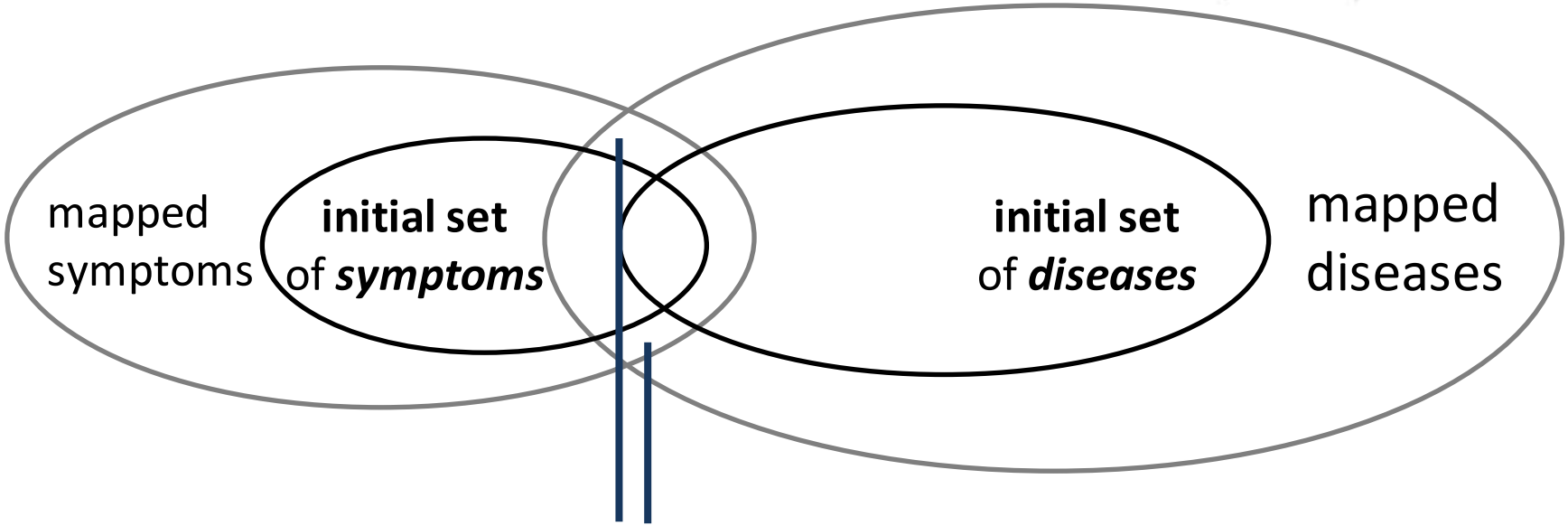


Classified as symptoms:
removed from diseases

Classified as diseases:
removed from symptoms

Complete classication results are available at http://goo.gl/CFgFVx

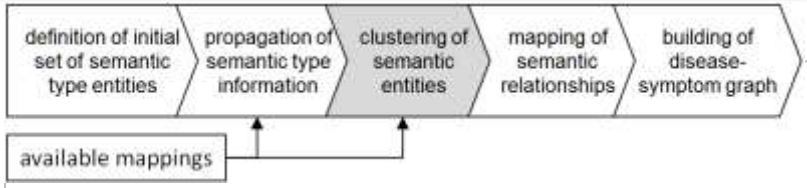
Overlap of Disease and Symptom Sets



- Automatic classification of 5139 classes in the overlap of mapped sets:
  - 2847 diseases: remove from symptom set
  - 2292 symptoms: remove from disease set

Complete classication results are available at http://goo.gl/CFgFVx

Clustering



- Initial approach: Maximal connected components
  - largest cluster with 70,000+ entities
- Adapted approach: Keep distinctions of source ontologies
  - largest disease cluster: 64 entities
  - largest symptom cluster: 57 entities
  - many 1-entity clusters

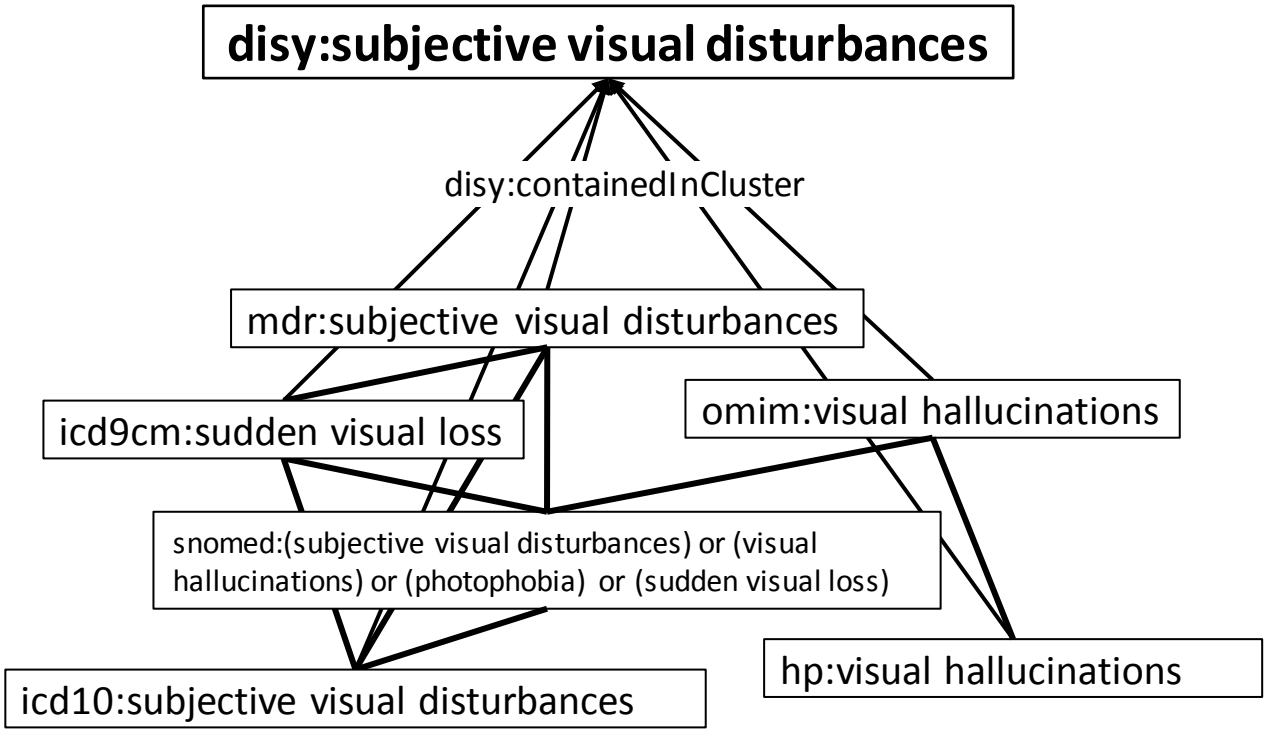
Diseases Clusters:

	UMLS	LOOM	All
clusters	167,970	113,165	102,990
max cluster size	20	53	64
1-entity-clusters	135,313	70,820	62,562

Symptom Clusters:

	UMLS	LOOM	All
clusters	16,416	13,000	11,530
max cluster size	18	53	57
1-entity-clusters	13,243	9,491	8,010

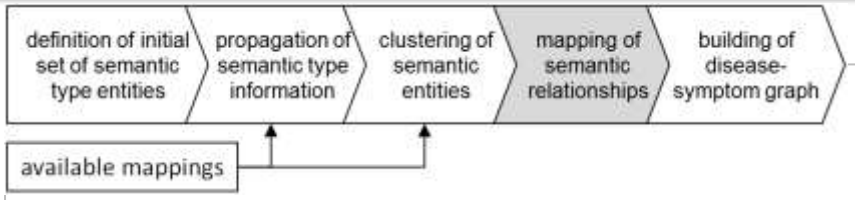
Example: Symptom Cluster

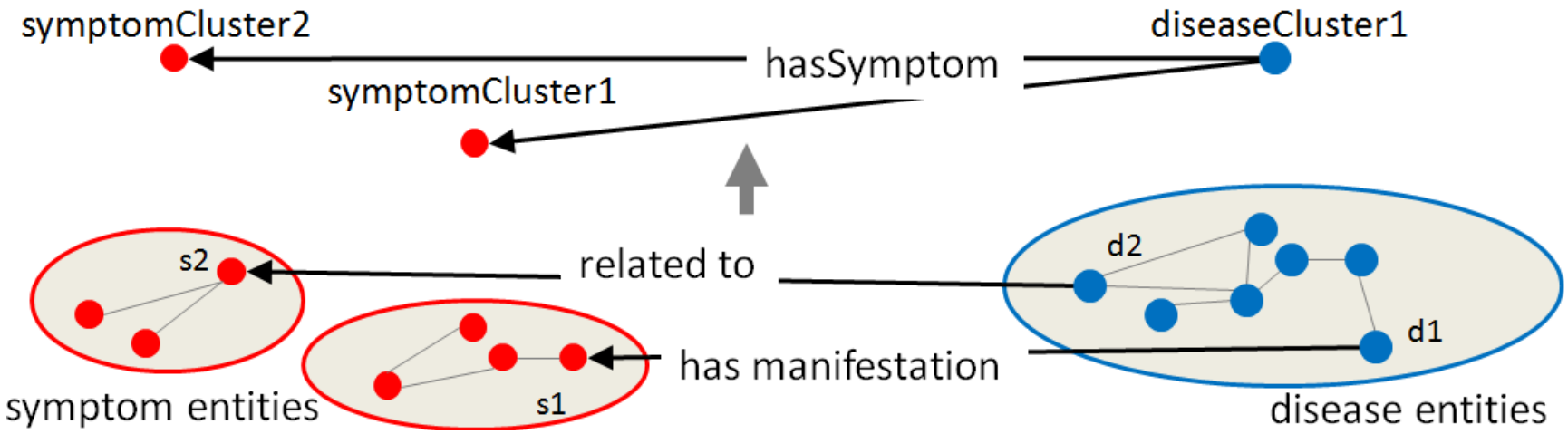


cluster

entities from different ontologies with mappings

Mapping Object Properties





Relations between Diseases and Symtpoms

Universität Augsburg University

Software Methodologies for Distributed Systems

2474	MDR:SIB
1463	rdfs:subClassOf
1346	RCD:SIB
1114	OMIM:has_manifestation
737	WHO:SIB
398	MSH:SIB
369	MEDLINEPLUS:SIB
330	ICD9CM:SIB
324	ICD10CM:SIB
268	MDR:classified_as
149	CSP:SIB
146	MSH:mapped_to
127	MDR:classifies
79	SNOMEDCT:may_be_a
66	WHO:RN
48	CSP:RN
47	MEDLINEPLUS:related_to
31	WHO:RB
28	MSH:RO
21	CSP:RB
18	CSP:RO
15	SNOMEDCT:associated_morphology_of
9	SNOMEDCT:same_as
8	SNOMEDCT:is_alternative_use
4	SNOMEDCT:replaces
3	SNOMEDCT:cause_of
3	SNOMEDCT:interprets
2	SNOMEDCT:replaced_by
2	ICPC2P:replaced_by
2	ICPC2P:replaces
1	SNOMEDCT:occurs_after
1	SNOMEDCT:associated_finding_of
1	SNOMEDCT:occurs_before

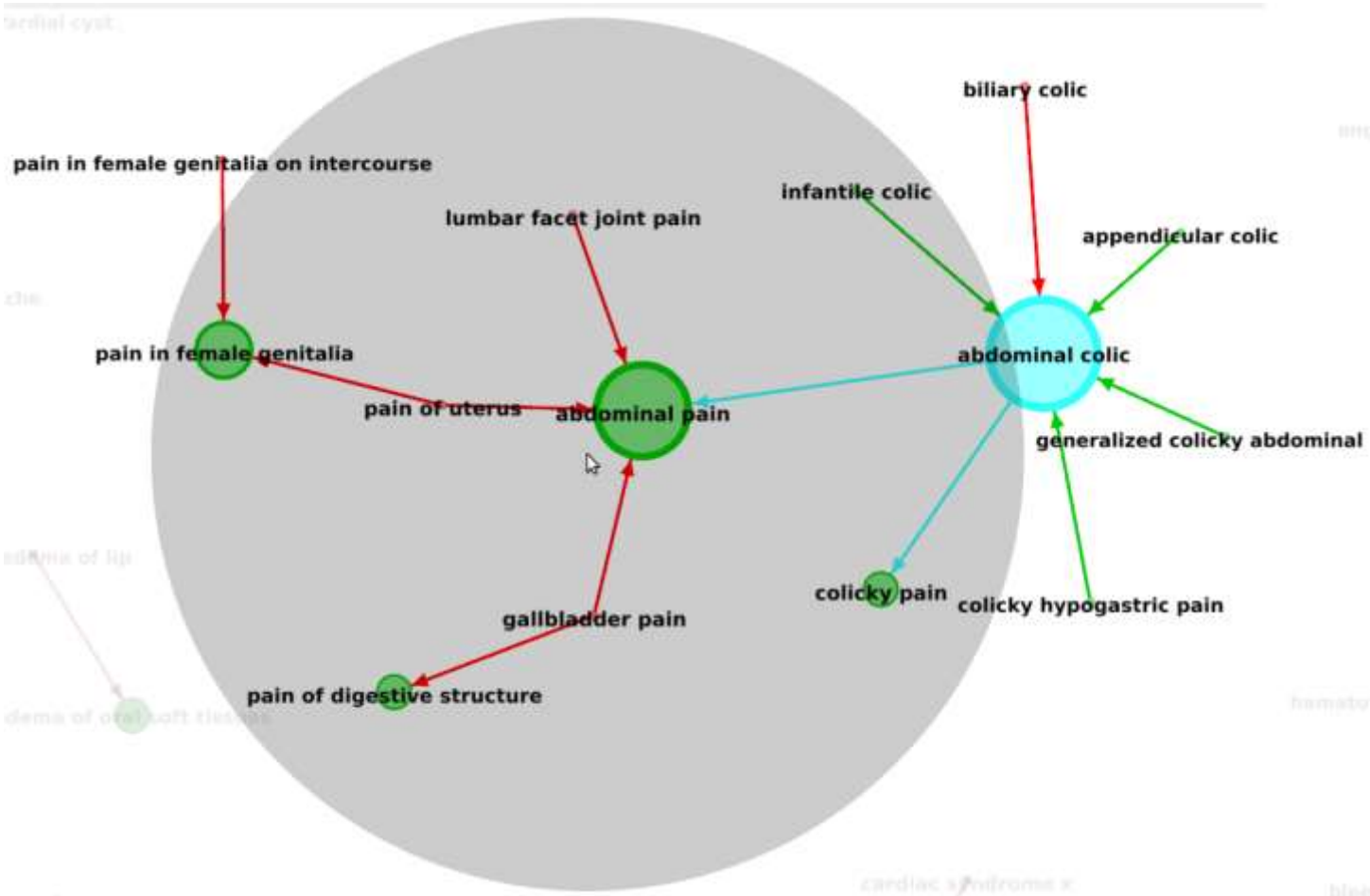
looked for triples

```
?disease ?p ?symptom .
```

obtained 33 properties

Subclass Relations

- Example: SNOMED CT



Mapping Object Properties

Universität Augsburg University

Software Methodologies for Distributed Systems

sibling

- MDR/SIB
- RCD/SIB
- WHO/SIB
- MSH/SIB
- MEDLINEPLUS/SIB
- ICD9CM/SIB
- ICD10CM/SIB
- CSP/SIB

hasSymptom

- OMIM/has_manifestation
- MEDLINEPLUS/related_to
- SNOMEDCT/cause_of

RN

- WHO/RN
- CSP/RN

rdfs:subClassOf

- WHO/RB
- CSP/RB

RO

- CSP/RO
- MSH/RO

skos:exactMatch

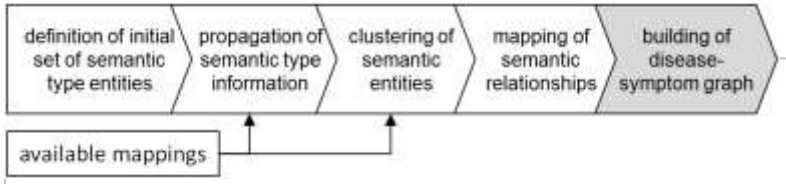
- SNOMEDCT/same_as
- MSH/mapped_to

replaces

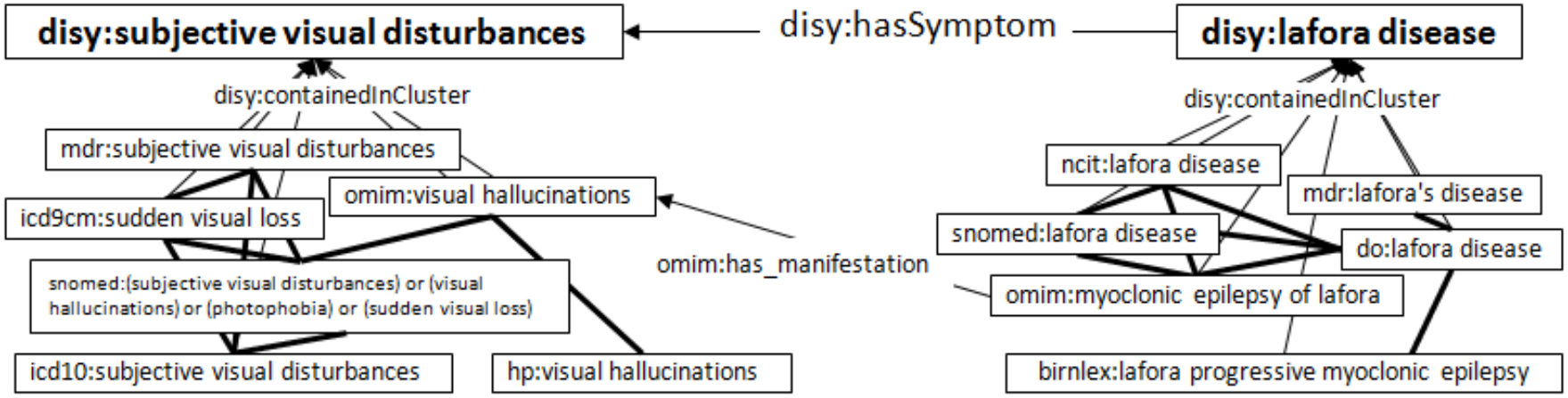
- SNOMEDCT/replaces
- ICPC2P/replaces

SNOMEDCT/replaced_by
SNOMEDCT/occurs_before
SNOMEDCT/occurs_after
SNOMEDCT/may_be_a
SNOMEDCT/is_alternative_use
SNOMEDCT/associated_finding_of
SNOMEDCT/associated_morphology_of
SNOMEDCT/interprets
MDR/classified_as
MDR/classifies
ICPC2P/replaced_by

Disease-Symptom Cluster Graph

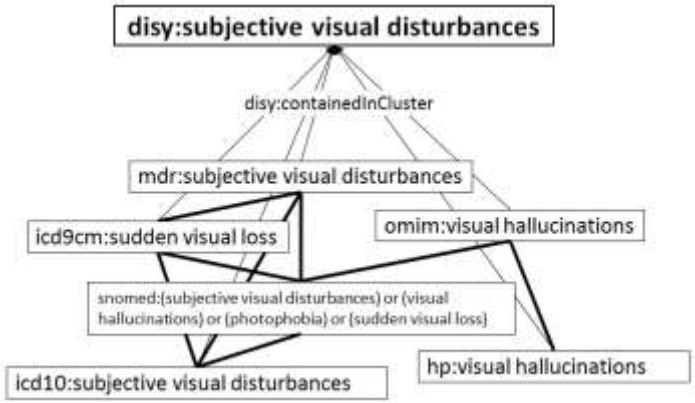


- **Initially:** 1114 distinct disease entities related to 345 symtom entities.
- **Cluster Graph:** 5960 diseases entities related to 3615 symtom entities.



Evaluation of Clusters

- Evaluated, whether all entities in a cluster are about the same disease (symptom) and thus correctly represented by the cluster.
- Checked the preferred labels of entities and cluster.



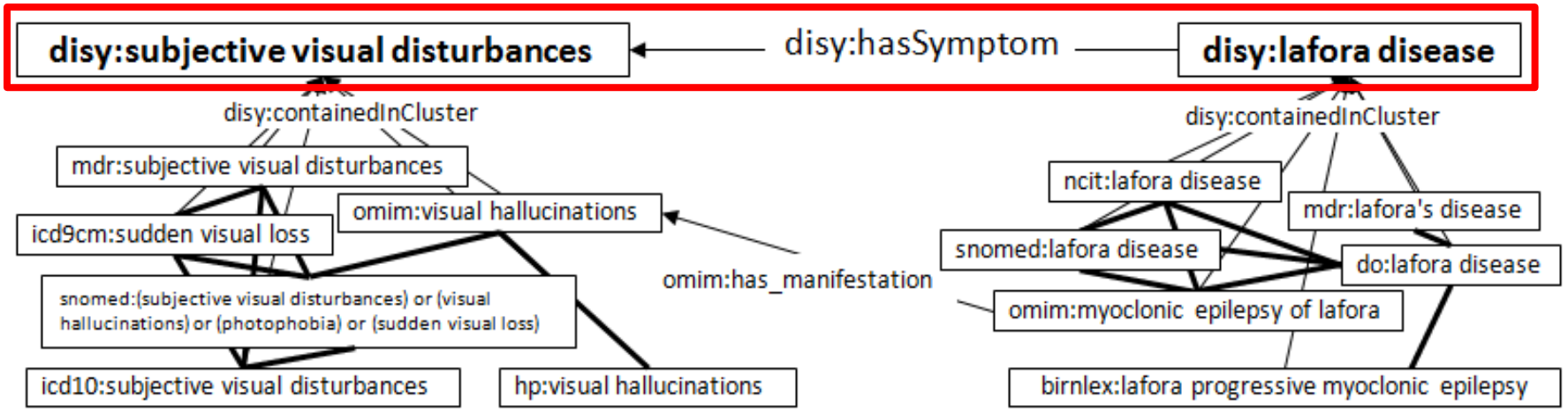
	disease	symptom
evaluated clusters	100	100
max. cluster size	28	36
max. distinct labels	15	18
correct clusters	**91**	**86**
clusters with wrong entities	**9**	**14**

Evaluation: Cluster-level Relations

Universität Augsburg University

Software Methodologies for Distributed Systems

- Randomly selected 500 cluster level relations out of all 2531 relations of the cluster graph.
- All were evaluated as correct.



Conclusion

- Clear picture on available disease-symptom relations
- Only few of the found disease-symptom relations
  - … and only few of them are useful are usefull for our application scenario.
- Clustering can be enhanced:
  - Allow overlapping clusters to reduce number of 1-entity clusters
  - use some clustering metric involving number and type of mappings
- Try the approach on other semantic types
- Cluster graph is a good resource for
  - identification of further disease-symptom relations by annotation of their textual definitions.
  - idetification of disease entities in other LOD resources

Questions?

Contact:

Heiner Oberkampf
OSTHUS GmbH
heiner.oberkampf@osthus.com
www.osthus.com