

Interoperability standards for clinical text processing

Seminar at University of Manchester – January 11, 2023



Stefan Schulz

- Medical University of Graz, Austria
- Averbis GmbH, Freiburg, Germany



Computer Methods and Programs in Biomedicine 45 (1994) 75–78

computer methods
and programs
in biomedicine

The Galen Project[☆]

A.L. Rector*, W.A. Nowlan and the GALEN Consortium

Medical Informatics Group, Department of Computer Science, University of Manchester, Manchester M13 9PL, UK

Abstract

The GALEN project is developing language independent concept representation systems as the foundations for the next generation of multilingual coding systems. It aims to support the flexibility required to cope with the diversity amongst medical applications, while ensuring the coherence necessary for integration and re-use of terminologies. GALEN is developing a fully compositional and generative formal system for modelling concepts: the GALEN Representation and Integration Language (GRAIL) Kernel. Its goal is to overcome many of the problems with traditional coding and classification systems, in particular the combinatorial explosion of terms in enumerative systems and the generation of nonsensical terms in partially compositional systems. It will also provide a clean separation between the concept model and linguistic mechanisms which interpret that model (i.e., the words in a specific language, syntax, alternative phrasings, etc.) in order to allow the development of multilingual systems. GRAIL aims to be formally sound and produce models that are verifiable and contain no contradictions or ambiguities, with realistic human effort. A Coding Reference (CORE) Model of medical terminology covering is being developed which aims to represent the core concepts in for example pathology, anatomy and therapeutics, that have widespread applicability in medical applications. It should also provide the basis for specialist extensions according to the formal principles of GRAIL. The main results of GALEN will be delivered as a Terminology Server (TeS) which encapsulates and coordinates the functionality of the concept module, multilingual module, and code conversion module, and also provides a uniform applications programming interface and network services for use by external applications.

Key words: Medical; Terminology medical; Coding medical; Classification medical; Knowledge; Representation medical; Records



OWL 2 Web Ontology Language Manchester Syntax (Second Edition)

W3C Working Group Note 11 December 2012

This version:

<http://www.w3.org/TR/2012/NOTE-owl2-manchester-syntax-20121211/>

Latest version:

<http://www.w3.org/TR/owl2-manchester-syntax/>

Previous version:

<http://www.w3.org/TR/2012/NOTE-owl2-manchester-syntax-20121018/>

Authors:

[Matthew Horridge](#), University of Manchester
Peter F. Patel-Schneider, Nuance Communications

A [color-coded version of this document showing changes made since the previous version](#) is also available.

This document is also available in these non-normative formats: [PDF version](#).

Copyright © 2012 W3C® (MIT, ERCIM, Keio). All Rights Reserved. W3C [liability](#), [trademark](#) and [document use](#) rules apply.

Scope and focus



Role of ontologies and information models relevant for clinical documentation in the context of natural language

Biomedicine – area best covered by interoperability resources

■ Terminologies (not based on logic)

- Taxonomies, classifications, catalogues
 - International Classification of diseases (ICD-9, **ICD-10**, ICD-11)
 - ATC (Anatomical Therapeutic Chemical Classification System)
 - NCBI Taxonomy (biological species)
- Thesauri
 - MeSH – Literature indexing
 - MedDRA – Drug Regulation
 - NCIthesaurus (cancer documentation)
 - RxNorm (drugs)
 - **LOINC** (lab and other observables)

■ Information models

- **HL7 FHIR** ■ EN 13606
- openEHR

■ Ontologies (based on logic)

- Ontology-based terminology
 - **SNOMED CT** (electronic health records)
- Ontologies
 - Gene Ontology (activities, processes, sites)
 - Sequence Ontology (nucleotides, protein sequences)
 - ChEBi (chemical entities)
 - HPO (human phenotypes)
 - FMA (anatomy)

■ “Databases”

(large catalogues of similar entities with instance annotations by ontologies)

- UNIPROT (Proteins)
- Reactome (biological pathways)
- BRENDA (enzymes)

Purpose of biomedical interoperability resources / standards

- Healthcare and medicine:
 - **Routine coding, e.g. for reimbursement or controlling** (diagnoses, procedures)
 - Major source of bias: selective, coarse-grained, erroneous
 - **Mortality and Morbidity statistics** (e.g. ICD for diseases at WHO level)
 - **Clinical registries** (e.g. tumour documentation)
 - **Drug regulatory activities**
 - Standardisation of clinical data in electronic health records (EHRs)
 - Clinical decision support
 - Clinical research
- Biomedical Research
 - **Annotation of research papers**
 - Support interoperability of research data (FAIR criteria)
 - Machine support of biomedical data processing in AI scenarios

Most clinical information is contained in clinical narratives

... using the local natural language



Porto Alegre
Brazil



Graz
Austria

Paciente G1PO, IG de 38 sem 4 dia(s), TS A+, interna por bolsa roita há mais de 18hs, recebendo penicilina. Evolui para Parto Eutócico com episiotomia em 27/06/2007 22:24 hs. Nasce RN APGAR 10/10, MASC, 3060 G. Exames: Toxo IGG e IGM neg VDRL neg EQU neg UROC: ausência de crescimento bacteriano. Hemograma 198mil plaq; Hb 13,1; LT 12,5 (75% seg) Em condições de alta, amamentando, útero contraído, lóquios fisiológico, sinais vitais estáveis, FO com bom aspecto. Recebe as orientações abaixo. ORIENTAÇÕES NA ALTA: # AMAMENTAÇÃO EXCLUSIVA POR 6 MESES; # TOMAR AS MEDICAÇÕES PRESCRITAS (SULFATO FERROSO 300MG 3X/DIA POR 90 DIAS, LONGE DAS REFISÇÕES, COM SUCO DE LARANJA; PARACETAMOL 750 MG 6/6HS SE DOR); # ORIENTO ANTICONCEPÇÃO; # RETORNAR À EMERGÊNCIA DESTE HOSPITAL SE FEBRE, SANGRAMENTO AUMENTADO OU OUTRAS INTERCORRÊNCIAS. # NÃO É NECESSÁRIO RETIRAR OS PONTOS. # LAVAR FO 3X/DIA COM ÁGUA E SABÃO DE GLICERINA.

* Anamnese und klinische Symptomatik
Stat. Übernahme vom LKH Fürstenfeld wegen neuerlicher Dyspnoe bei bek. dil. CMP u hochgr. MINS zur CA und Mitraclip /erztransplant Evaluierung. Bei dem Patienten besteht der St.p. 2x Simdax Therapie im Okt 2013.
* Physikalischer Status
48 jähr.Patient, deutl. reduz. AZ, normaler EZ. Cor: Ht rh, nc, Systolikum mit p.max. über dem Erbschen Punkt mit Fortleitung in die Axila
Pulmo: VA bds., feuchte RGs re>li
Abdomen: BD weich, kein DS
Extremitäten: ausgeprägte Knöchelödeme bds.
Herr DI Max Mustermann wurde aufgrund einer neuerlichen Dyspnoesymptomatik bei bek. dilat. CMP und hochgrad. MINS zur weiteren Evaluierung stat. vom LKH Fürstenfeld übernommen.

Clinical language: compact, sloppy, contextualised

Phenomenon	Example	Elucidation
Telegram style	"left PICA stroke, presented to ED after fall"	Incomplete sentences, sketchy style
Colloquialisms	"pothole sign", "snorkel"	Milieu-specific sub-languages
Ad-hoc abbreviations	"infiltr"	Truncation ("infiltrated mucosa")
Ambiguous short forms	"RTA"	"Road traffic accident", "Renal-tubular acidosis"
Short forms of regional or local scope	"LDS Hospital" "St. p."	"Latter-Day-Saints Hospital" (and not "Leak Detection System") "Status post" = "History of"
Conventionalized Latin abbreviations	"V mors can dig V dext"	"Vulnus morsum canis digiti quinti dextri" (in some European languages)
Numeric codes	"45, 46 with crowns", "VI palsy", "2-2-2",	Tooth numbers, cranial nerves, dose frequencies
Spelling errors, typos	"Diabtes", "Astra-Seneca", "Hipotireose",	accidental (quick typing) or systematic (e.g. 2 nd language speakers)
Spelling variants	"Esophagus", "Oesophagus"	e.g. American vs. British English
Single noun compounds	"Ibuprofenintoxikation"	Non-lexicalized long words (in languages such as German, Swedish)
Anaphora	(i) "adenoCa rect pN+MX G2 (...). tumor excised in toto" (ii) "no blood in stomach (...). mult mucosal erosions "	(i) "Tumor" coreferential to adenocarcinom described in left context (ii) "mucosal erosions" refined to "erosions of gastric mucosa"
Negations	"No evidence of pneumonia" "Pulmones: nihil", "metastasenfrei"	non-standard, jargon-like
Epistemic contexts	"susp MI, DD lung embolism"	suspected diagnosis, differential diagnosis
Temporal contexts	"h/o Covid-19", "Streptokokkenangina 06/16"	"history of" Coarse-grained references to dates (mm/yy)
Other contexts	(i) father: pancreas ca" (ii) "refrained from resuscitation"	(i) family history (ii) plans not executed

Desideratum: narrative data → ontology-based data

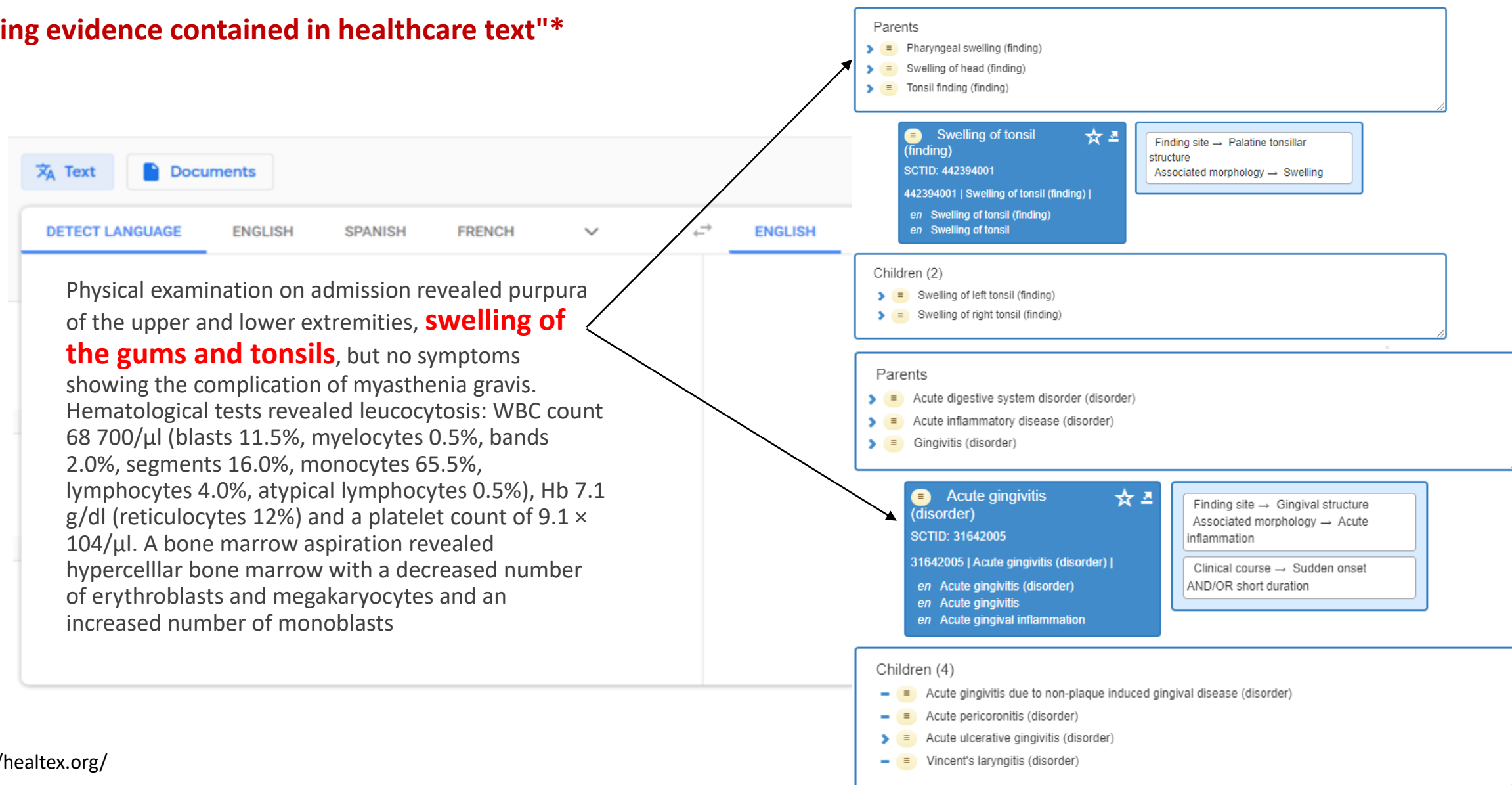
"Unlocking evidence contained in healthcare text"*

The screenshot displays the Healix web application interface. At the top, there are two tabs: 'Text' (selected) and 'Documents'. Below the tabs, there are two rows of language and ontology selection buttons. The first row includes 'DETECT LANGUAGE', 'ENGLISH', 'SPANISH', and 'FRENCH'. The second row includes 'ENGLISH', 'SPANISH', and 'SNOMED CT' (highlighted in red). The main content area is split into two columns. The left column contains a paragraph of medical text: 'Physical examination on admission revealed purpura of the upper and lower extremities, swelling of the gums and tonsils, but no symptoms showing the complication of myasthenia gravis. Hematological tests revealed leucocytosis: WBC count 68 700/μl (blasts 11.5%, myelocytes 0.5%, bands 2.0%, segments 16.0%, monocytes 65.5%, lymphocytes 4.0%, atypical lymphocytes 0.5%), Hb 7.1 g/dl (reticulocytes 12%) and a platelet count of 9.1 × 104/μl. A bone marrow aspiration revealed hypercellular bone marrow with a decreased number of erythroblasts and megakaryocytes and an increased number of monoblasts'. The right column displays a list of SNOMED CT codes corresponding to the entities in the text: 419620001 110714004 65124004 113279002 116223007 91637004 252275004 111583006 767002 [68700] 271040006 [11.5] 313696224 [0.5] 313696667 [2.0] 313696009 [16.0] 271037006 [65.5] 271036002 [4.0] 271036013 [0.5] 365809007 [7.1] 45995003 [12] 365632008 [91000] 49401003 76197007 14016003 420510009 103213002 53945006 35105006.

Text	SNOMED CT Codes
Physical examination on admission revealed purpura of the upper and lower extremities, swelling of the gums and tonsils, but no symptoms showing the complication of myasthenia gravis. Hematological tests revealed leucocytosis: WBC count 68 700/μl (blasts 11.5%, myelocytes 0.5%, bands 2.0%, segments 16.0%, monocytes 65.5%, lymphocytes 4.0%, atypical lymphocytes 0.5%), Hb 7.1 g/dl (reticulocytes 12%) and a platelet count of 9.1 × 104/μl. A bone marrow aspiration revealed hypercellular bone marrow with a decreased number of erythroblasts and megakaryocytes and an increased number of monoblasts	419620001 110714004 65124004 113279002 116223007 91637004 252275004 111583006 767002 [68700] 271040006 [11.5] 313696224 [0.5] 313696667 [2.0] 313696009 [16.0] 271037006 [65.5] 271036002 [4.0] 271036013 [0.5] 365809007 [7.1] 45995003 [12] 365632008 [91000] 49401003 76197007 14016003 420510009 103213002 53945006 35105006

Desideratum: narrative data → ontology-based data

"Unlocking evidence contained in healthcare text"*



Basic natural language processing (NLP) tasks

A 28-year-old female with a history of gestational diabetes mellitus diagnosed eight years prior to presentation and subsequent type two diabetes mellitus (T2DM), one prior episode of HTG-induced pancreatitis three years prior to presentation , associated with an acute hepatitis , and obesity with a body mass index (BMI) of 33.5 kg/m2 . She presented with a one-week history of polyuria , polydipsia , poor appetite , and vomiting . Two weeks prior to presentation , she was treated with a five-day course of amoxicillin for a respiratory tract infection . She was on metformin , glipizide , and dapagliflozin for T2DM and atorvastatin and gemfibrozil for HTG . She had been on dapagliflozin for six months at the time of presentation . Physical examination on presentation was significant for dry oral mucosa ; significantly , her abdominal examination was benign with no tenderness , guarding , or rigidity . Pertinent laboratory findings on admission were : serum glucose 111 mg/dl , bicarbonate 18 mmol/l , anion gap 20 , creatinine 0.4 mg/dL , triglycerides 508 mg/dL , total cholesterol 122 mg/dL , glycated hemoglobin (HbA1c) 10% , and venous pH 7.27 . Serum lipase was normal at 43 U/L . Serum acetone levels could not be assessed as blood samples kept hemolyzing due to significant lipemia . The patient was initially admitted for starvation ketosis , as she reported poor oral intake for three days prior to admission . However , serum chemistry obtained six hours after presentation revealed her glucose was 186 mg/dL , the anion gap was still elevated at 21 , serum bicarbonate was 16 mmol/L , triglyceride level peaked at 2050 mg/dL , and lipase was 52 U/L . The β -hydroxybutyrate level was obtained and found to be elevated at 5.29 mmol/L - the original sample was centrifuged and the chylomicron layer removed prior to analysis due to interference from turbidity caused by lipemia again . The patient was treated with an insulin drip for euDKA and HTG with a reduction in the anion gap to 13 and triglycerides to 1400 mg/dL , within 24 hours . Her euDKA was thought to be precipitated by her respiratory tract infection in the setting of SGLT2 inhibitor use . The patient was seen by the endocrinology service and she was discharged on 40 units of insulin glargine at night , 12 units of insulin lispro with meals , and metformin 1000 mg two times a day . It was determined that all SGLT2 inhibitors should be discontinued indefinitely . She had close follow-up with endocrinology post discharge .

Color codes: Patient problem, Test, Treatment

■ Entity recognition: identify spans in text and assign some semantic type

■ Entity normalization / term grounding: assign one or more codes from a CV

■ Disambiguation: chose the correct code for a mention in text

Not trivial in clinical language:

- "Amputation": Patient problem or treatment?
- "Potassium": Lab parameter or drug?
- "Emphysema" = "Lung emphysema"
- "Cancer" != "Lung cancer"
- "Diclophenac" = "Diclofenac", "Oesophagus" = "Esophagus"
- "Hepatectomy" != "Hepatotomy"
- "Type 2 diabetes" = "Type 2 diabetes mellitus"
- "Diabetes mellitus type 1 != Diabetes mellitus type 2"
- (beta blocker after) MI != (valve replacement due to) MI

- Creation and Maintenance of domain lexicons covering clinical jargon in the local natural language (interface terminologies)
- Linking interface terms to coding systems like SNOMED CT and ICD-10
- Support fuzzy term matching and disambiguation by algorithms and language models

The need for clinical interface terminologies

Clinical jargon != standard terminology

Frequency of SNOMED Preferred Terms and their translations

	Hits Google*
– English: "Secondary malignant neoplasm of liver"	100
– Swedish: "sekundär malign levertumör"	1
– German: "Sekundäre maligne Neoplasie der Leber"	1

Frequency of typical synonyms

– English: "liver metastases"	1,230,000
– Swedish: "levermetastaser"	217,000
– German: "Lebermetastasen"	204,000

Similar observations in clinical corpora / PubMed

– In a corpus with 30,000 German cardiology letters	
- "Electrocardiogram"	0
- "EKG"	0
– In Pubmed abstracts:	
- "phosphocholine transferase activity"	4
- "phosphocholine transferase"	36

Desideratum: narrative data → ontology-based data

TextDocuments

DETECT LANGUAGEENGLISHSPANISHFRENCH

↔

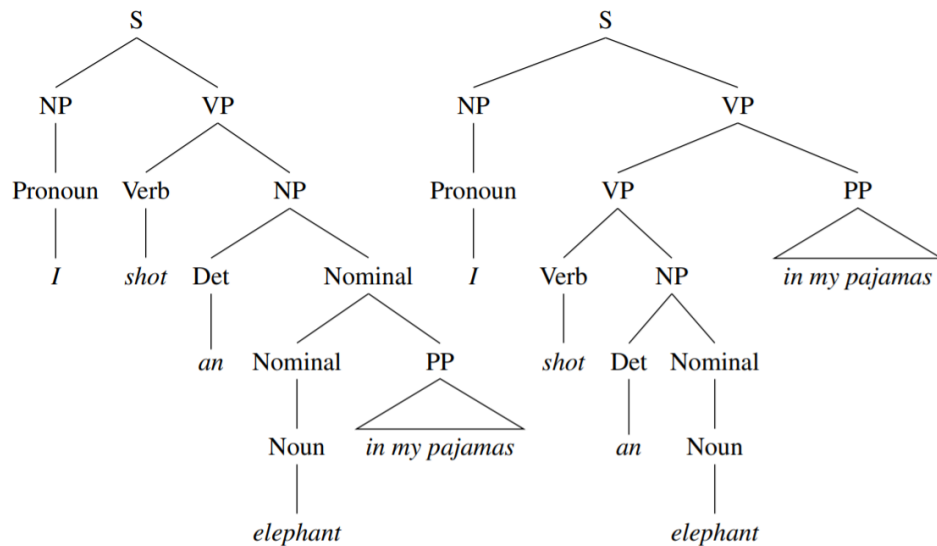
ENGLISHSPANISHSNOMED CT

Physical examination on admission revealed purpura of the upper and lower extremities, swelling of the gums and tonsils, but no symptoms showing the complication of myasthenia gravis. Hematological tests revealed leucocytosis: WBC count 68 700/ μ l (blasts 11.5%, myelocytes 0.5%, bands 2.0%, segments 16.0%, monocytes 65.5%, lymphocytes 4.0%, atypical lymphocytes 0.5%), Hb 7.1 g/dl (reticulocytes 12%) and a platelet count of 9.1×10^4 / μ l. A bone marrow aspiration revealed hypercellular bone marrow with a decreased number of erythroblasts and megakaryocytes and an increased number of monoblasts

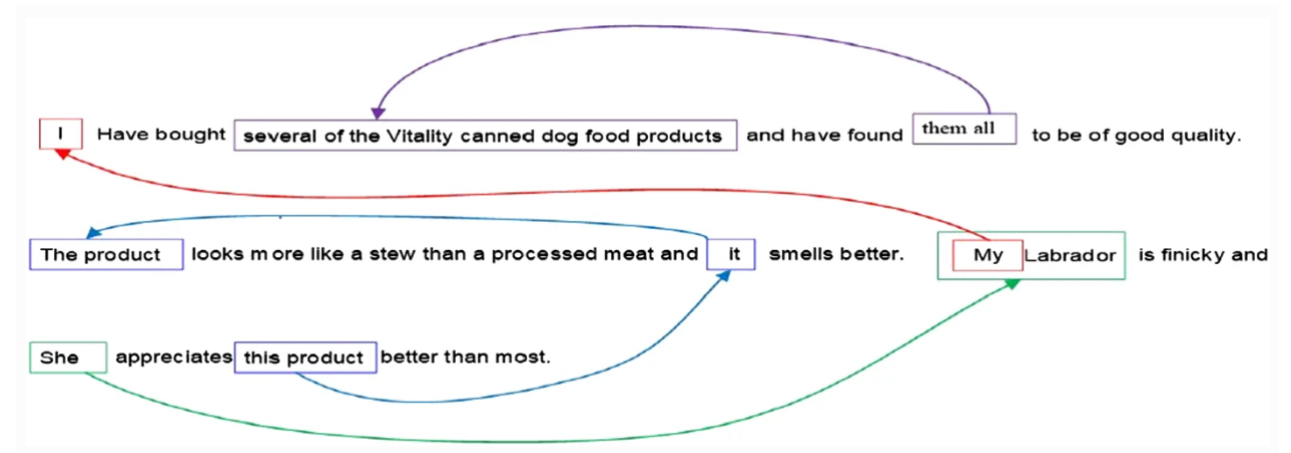
419620001 110714004 65124004 113279002 116223007
91637004 252275004 111583006 767002 [68700]
271040006 [11.5] 313696224 [0.5] 313696667 [2.0]
313696009 [16.0] 271037006 [65.5] 271036002 [4.0]
271036013 [0.5] 365809007 [7.1] 45995003 [12]
365632008 [91000] 49401003 76197007 14016003
420510009 103213002 53945006 35105006

Are sequences of ontology codes (and numeric values) really sufficient?

Human language is not linear



Syntax – Grammar



Discourse - Anaphora

Representation of narrative data as graphs rooted in ontologies


Text Documents

DETECT LANGUAGE ENGLISH SPANISH FRENCH

ENGLISH SPANISH SNOMED CT

Physical examination on admission revealed purpura of the upper and lower extremities, swelling of the gums and tonsils, but no symptoms showing the complication of myasthenia gravis. Hematological tests revealed leucocytosis: WBC count 68 700/ μ l (blasts 11.5%, myelocytes 0.5%, bands 2.0%, segments 16.0%, monocytes 65.5%, lymphocytes 4.0%, atypical lymphocytes 0.5%), Hb 7.1 g/dl (reticulocytes 12%) and a platelet count of 9.1×10^4 / μ l. A bone marrow aspiration revealed hypercellular bone marrow with a decreased number of erythroblasts and megakaryocytes and an increased number of monoblasts

QUERY



It is not sufficient to identify mentions and link them to codes:

Relation extraction and knowledge graph construction:

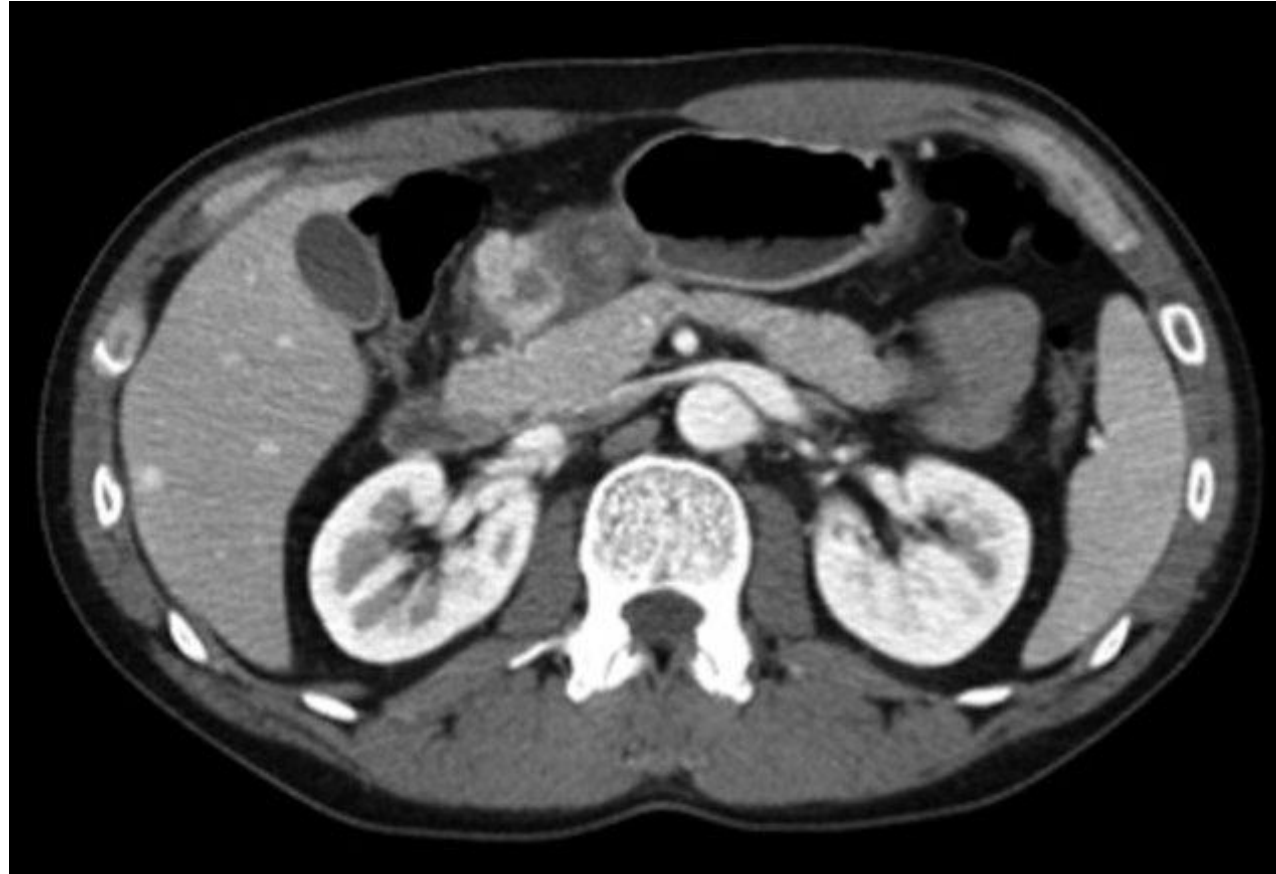
- Nodes represent the referents of mentions in the text ("entities") as instances of ontology concepts
- Relations from the same ontology are used to link the nodes

Knowledge graph construction by exploiting the axiomatic structure of the target ontology

Pylorus and superior duodenum:

Edematous thickening.

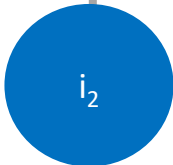
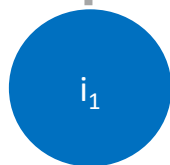
Diagnosis: ulcer.



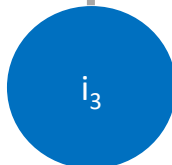
Leveraging entity
recognition and
normalization...

Anaphoric references (bridging anaphora)

Pylorus and superior duodenum:



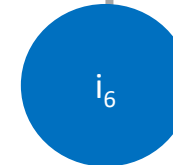
Edematous thickening.



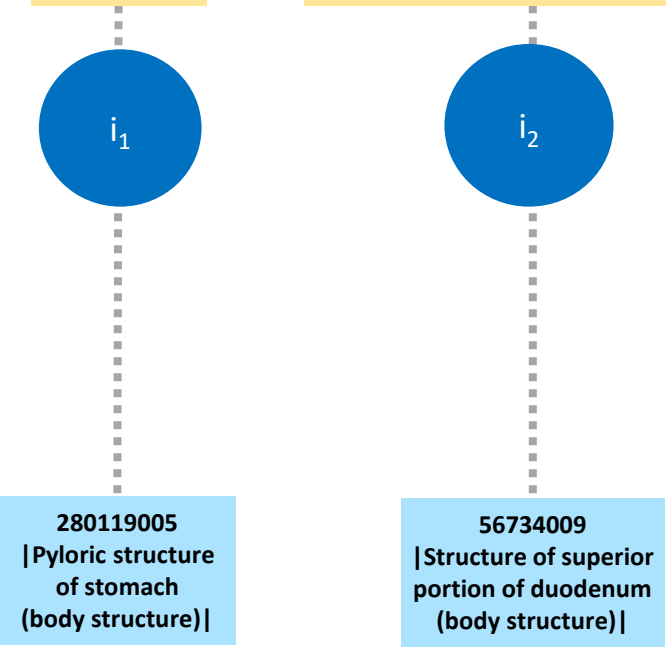
Diagnosis:



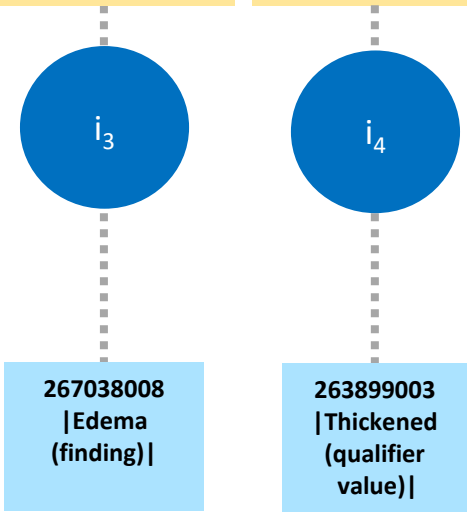
ulcer



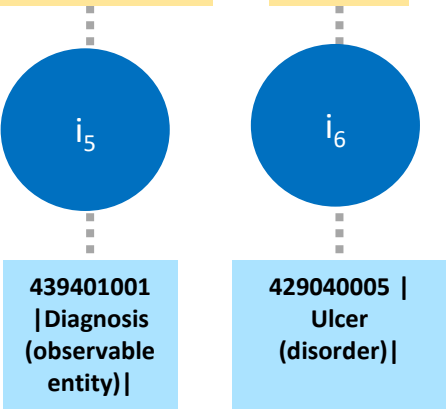
Pylorus and superior duodenum:

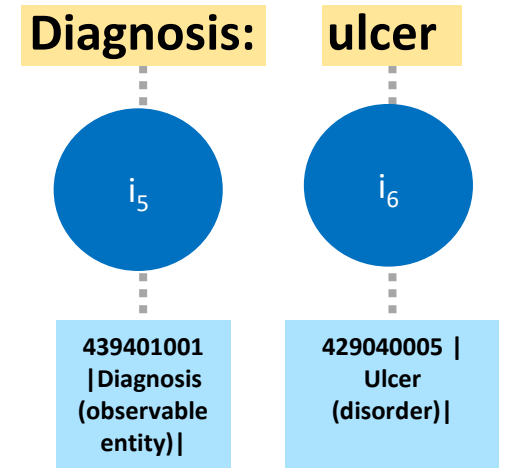
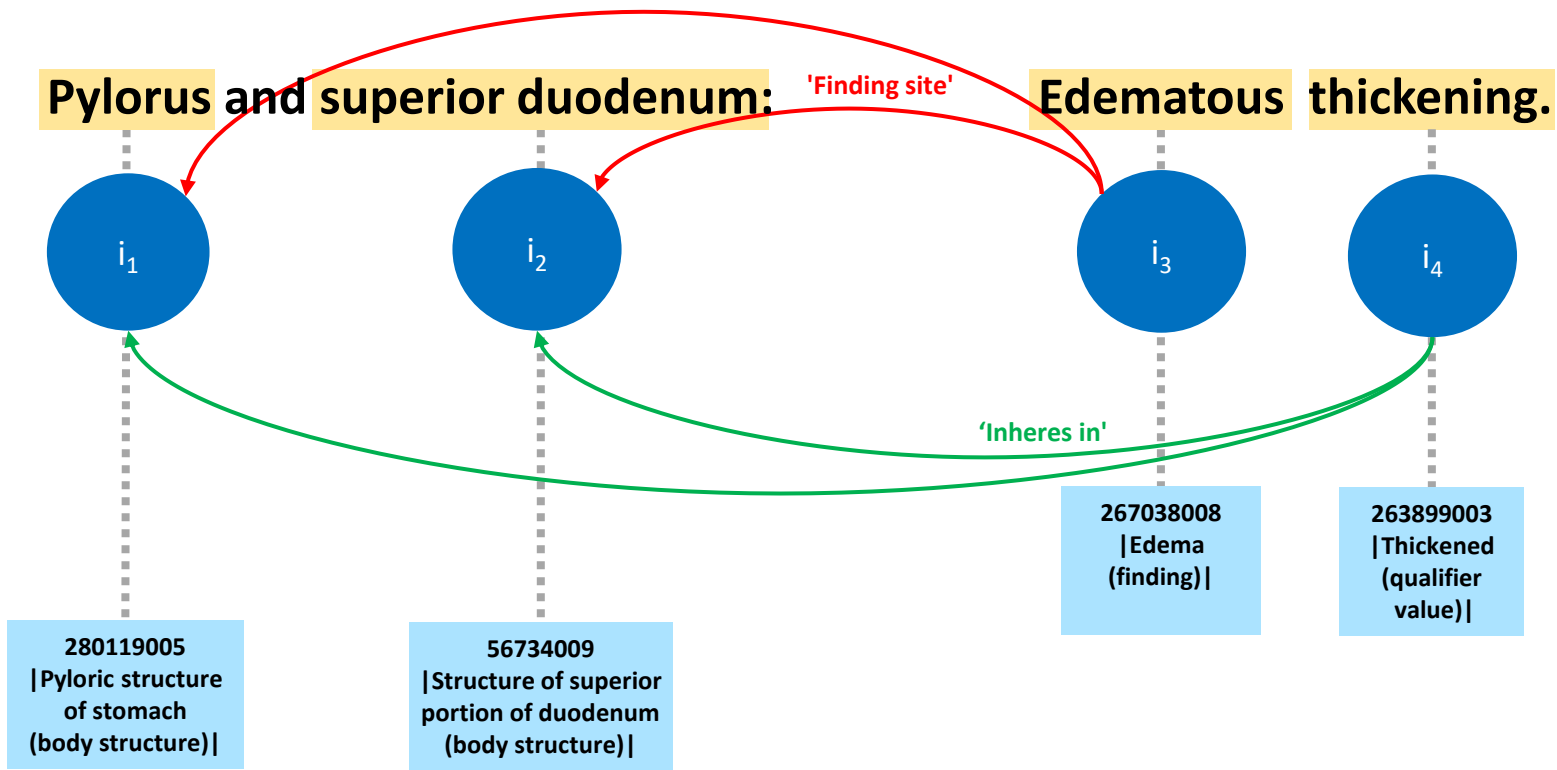


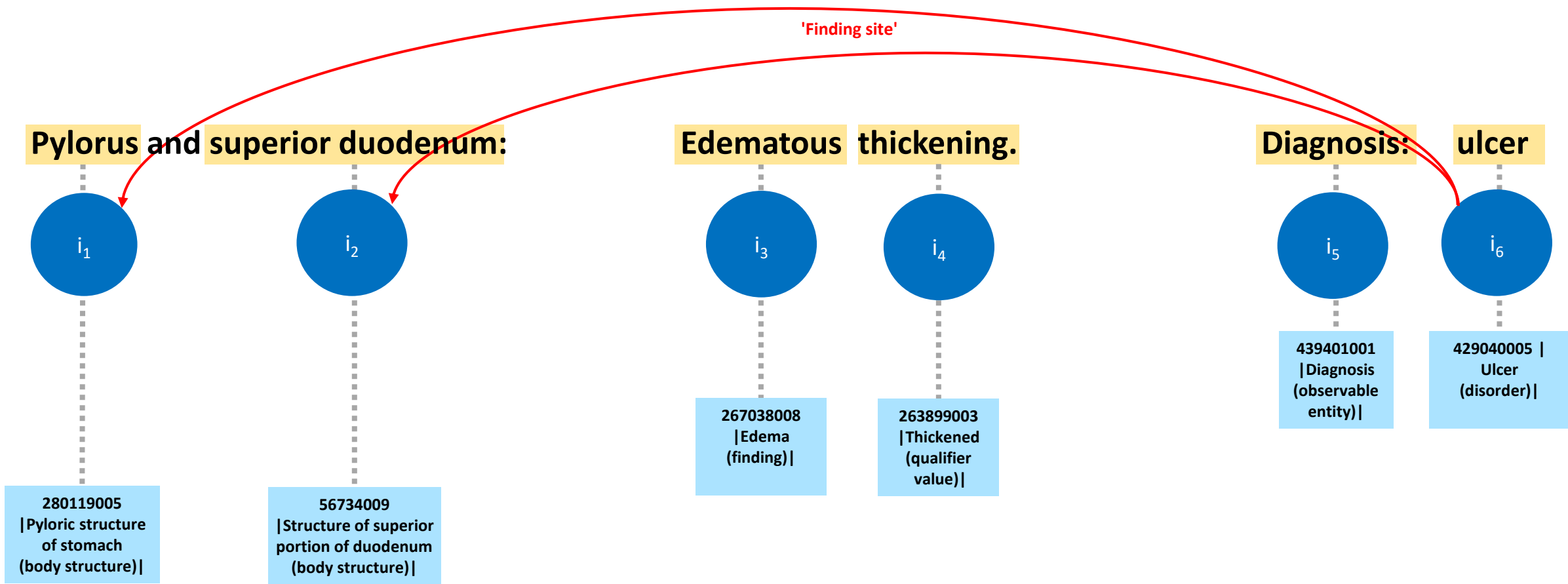
Edematous thickening.

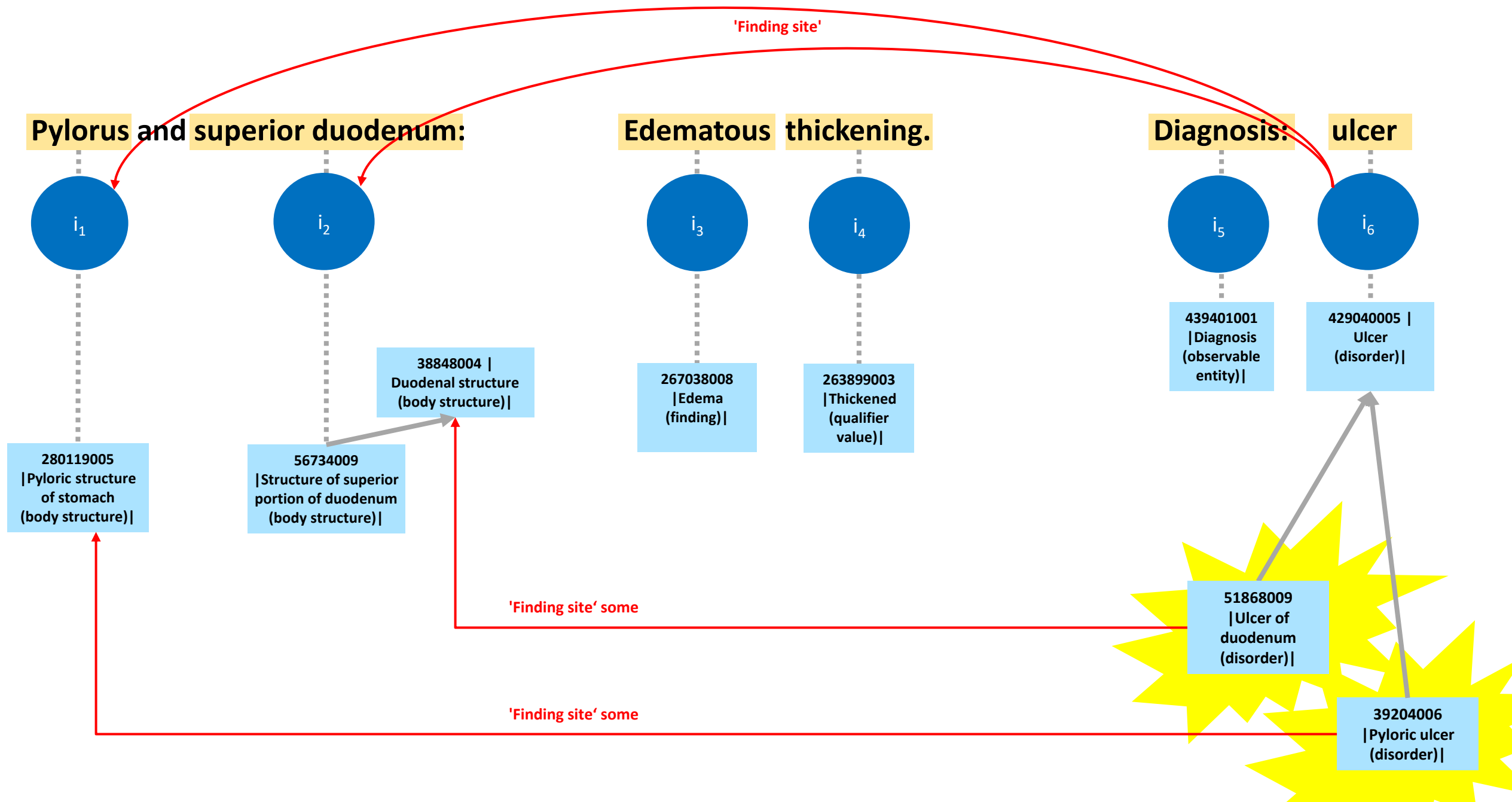


Diagnosis: ulcer

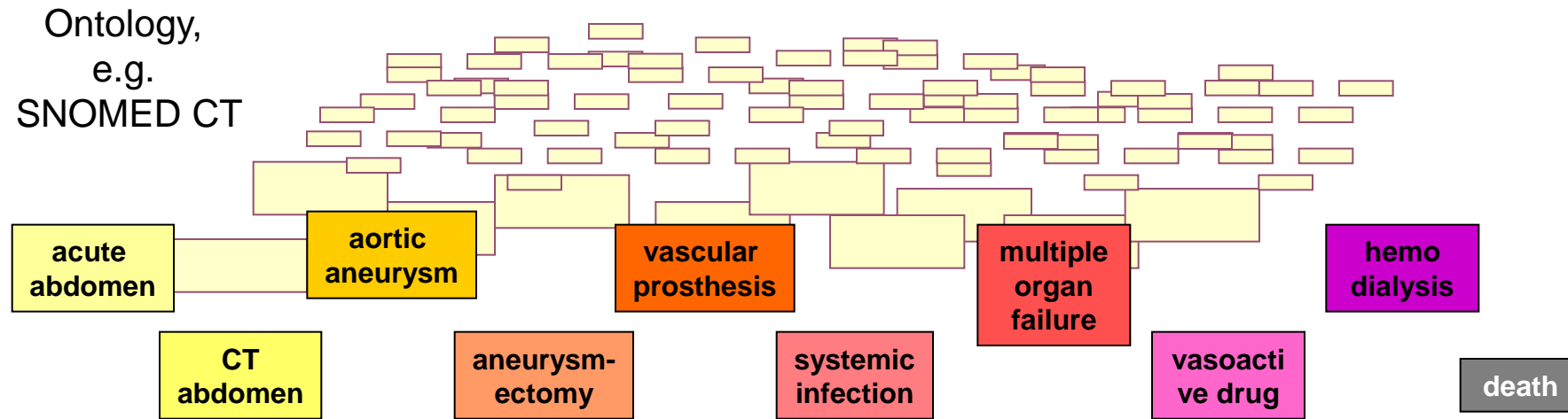








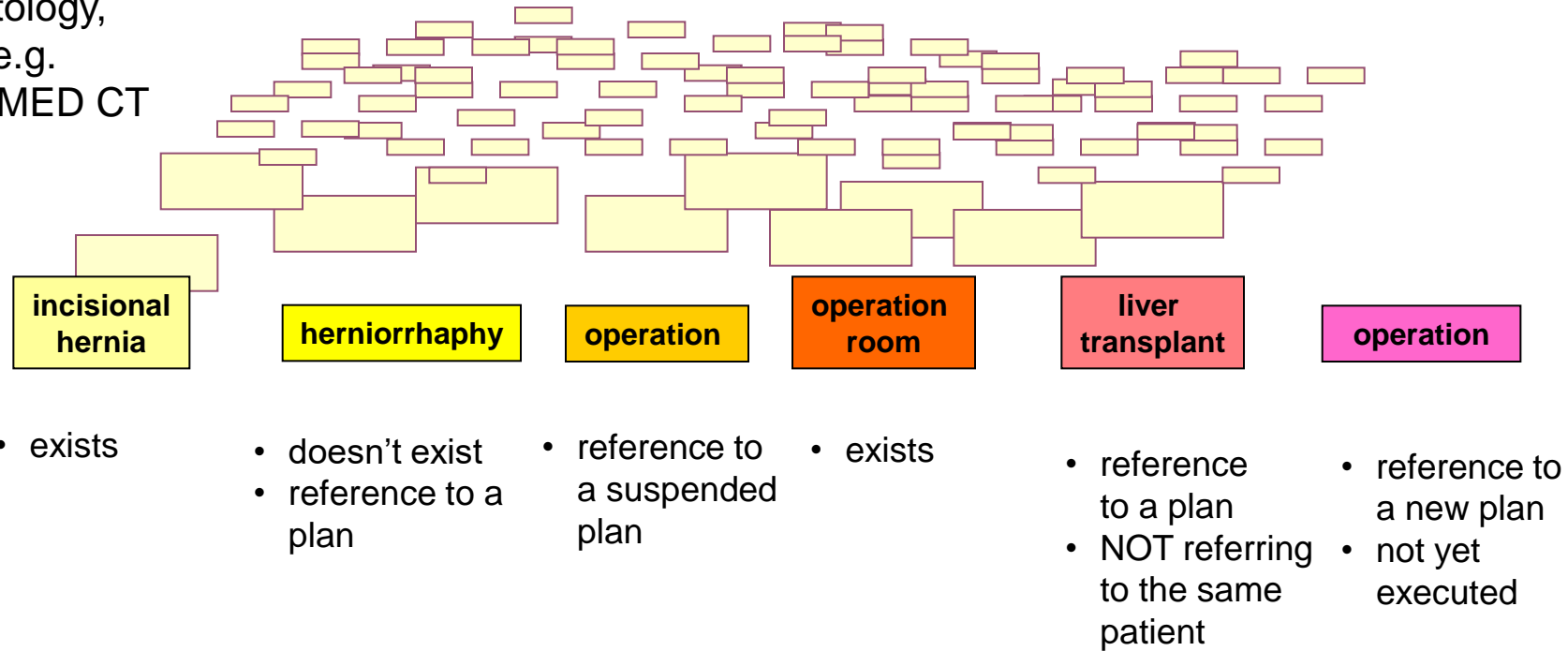
Ontology mapping by term recognition



Patient admitted with acute abdomen. Abdominal CT: leaking abdominal aortic aneurism. Emergency aneurysmectomy with prosthesis. Postoperative evolution with systemic inflammatory response syndrome, multiple organ failure and hemodynamic instability. Despite application of vasoactive drugs, volume replacement and hemodialysis, the patient's condition worsened evolving to death.

Ontology mapping is not enough!

Ontology,
e.g.
SNOMED CT

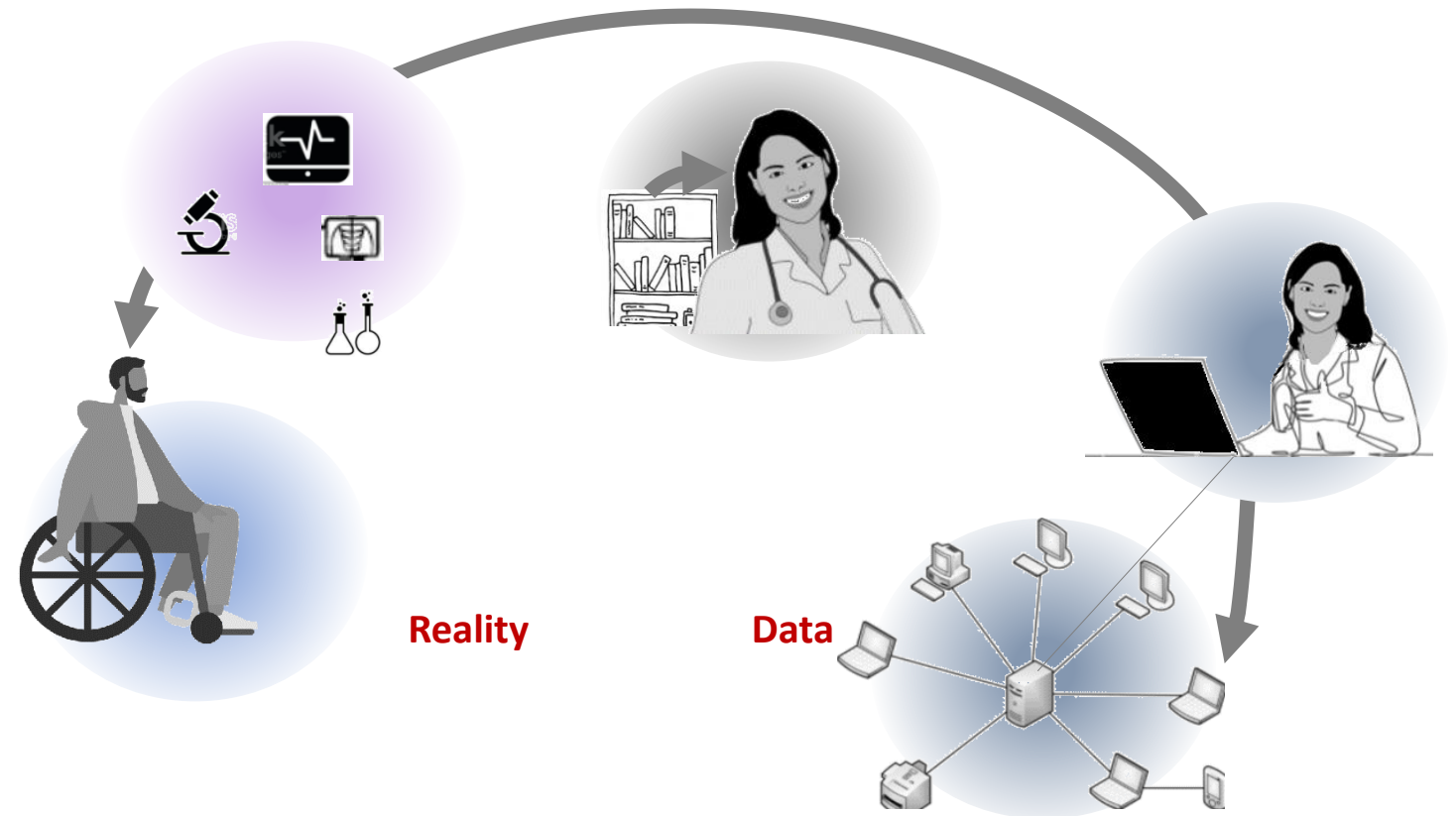


Patient with **incisional hernia** admitted for **herniorrhaphy**, but **operation** was suspended because **operation room** was urgently needed for **liver transplant**. Discharged with orientation and rescheduled **operation**.

Data vs. reality vs. context

Complex relationship between healthcare data and reality

- Mention of drug in EHR
 - Recommended by hospital doctor
 - Prescribed by general practitioner
 - Purchased by patient
 - Taken by patients
- Mention of disease
 - Suspected vs. confirmed
 - Disease or cause of death
 - Disease != Diagnosis !
 - There are undiagnosed diseases
 - There are wrong diagnoses



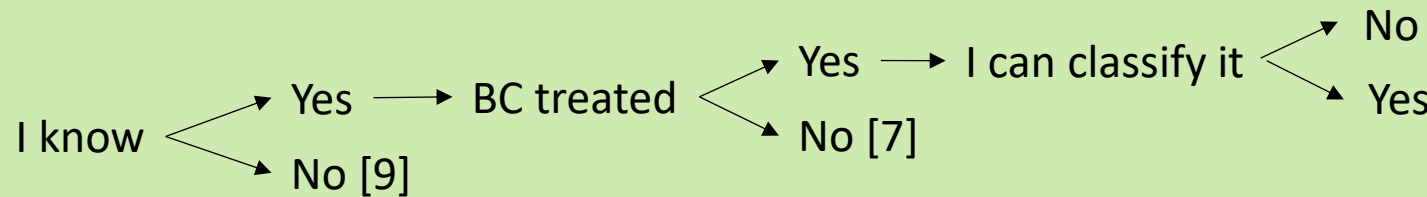
Ontology – Epistemology distinction

Often hidden in practical data management as well as in common speech

Typical checklist example:

"Breast cancer therapy"

[1] Operation [2] Radiotherapy [3] Antineoplastic [4] Hormone [5] Immunotherapy [6] Other [7] None [8] Unknown [9]



Epistemology:

- Knowledge / context
- Model of use

BC therapy [6]

BC Operation [1]
BC Radiotherapy [2]
BC Antineoplastic [3]
BC Hormone [4]
BC Immunotherapy [5]

Ontology:

- Entity types and their properties
- Model of meaning

Seminal papers

The Ontology-Epistemology Divide: A Case Study in Medical Terminology

Olivier BODENREIDER¹, Barry SMITH^{2,3}, Anita BURGUN⁴

¹ *US National Library of Medicine, Bethesda, Maryland, USA*

² *Institute for Formal Ontology and Medical Information Science, Saarbrücken, Germany*

³ *Department of Philosophy, University at Buffalo, New York, USA*

⁴ *Laboratoire d'Informatique Médicale, Université de Rennes I, France*

Abstract. Medical terminology collects and organizes the many different kinds of terms employed in the biomedical domain both by practitioners and also in the course of biomedical research. In addition to serving as labels for biomedical classes, these names reflect the organizational principles of biomedical vocabularies and ontologies. Some names represent invariant features (classes, universals) of biomedical reality (i.e., they are a matter for ontology). Other names, however, convey also how this reality is perceived, measured, and understood by health professionals (i.e., they belong to the domain of epistemology). We analyze terms from several biomedical vocabularies in order to throw light on the interactions between ontological and epistemological components of these terminologies. We identify four cases: 1) terms containing classification criteria, 2) terms reflecting detectability, modality, uncertainty, and vagueness, 3) terms created in order to obtain a complete partition of a given domain, and 4) terms reflecting mere fiat boundaries. We show that epistemology-loaded terms are pervasive in biomedical vocabularies, that the “classes” they name often do not comply with sound classification principles, and that they are therefore likely to cause problems in the evolution and alignment of terminologies and associated ontologies.

Binding Ontologies & Coding systems to Electronic Health Records and Messages

AL Rector MD PhD¹, R Qamar MSc¹ and T Marley MSc²

¹*School of Computer Science, University of Manchester, Manchester M13 9PL, UK*

²*Salford Health Informatics Research, University of Salford, Salford, UK*

ABSTRACT: *A major use of medical ontologies is to support coding systems for use in electronic healthcare records and messages. A key task is to define which codes are to be used where – to bind the terminology to the model of the medical record or message. To achieve this formally, it is necessary to recognise that the model of codes and information models are at a meta-level with respect to the underlying ontology. A methodology for defining a Code Binding Interface in OWL is presented which illustrates this point. It generalises methodologies that have been used in a successful test of the binding of HL7 messages to SNOMED-CT codes.*

Introduction

A major use of medical ontologies is to support medical terminologies and coding systems. A major use of medical terminology and coding systems is for electronic healthcare records and messages. Specifying the validation rules for how terminology and coding systems are to be used in electronic healthcare records and messages is, therefore, a key problem for medical ontologies.

We contend that electronic healthcare records messages are data structures and refer to their models as “information models”. By contrast we contend that the model of meaning or “ontology” is a model

Pragmatically, it is useful to decouple the coding system from the model of meaning so that reasoning about the model of meaning and model of coding system is always separated.

Using codes in messages and EHRs

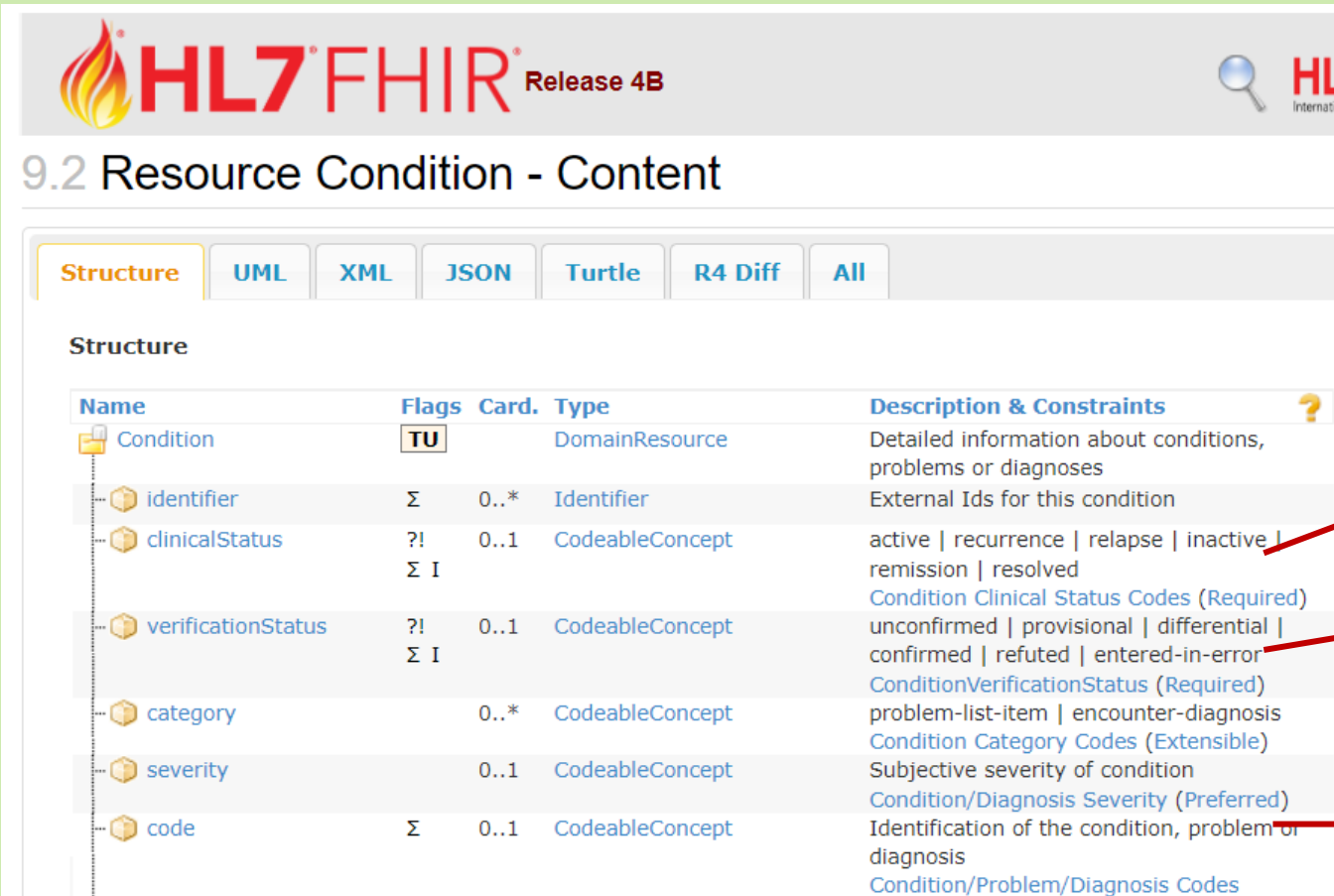
Our goal is to assist software developers in specifying information systems and the use of codes from coding systems within them. We seek to have specifications that are sufficiently precise that separately implemented systems will work together. To achieve this we need to be able to validate that the models themselves are self-consistent and that individual messages conform to the models.

Typically, we want to start with a generic information model such as the HL7 RIM¹ or the OpenEHR reference model². We then want to define progressively more specialised models in which each more specialised model is consistent with the next more generic model and ultimately the reference model. We want to use the models with separately developed coding systems – e.g. SNOMED, ICD, CPT, MEDRA, etc. Since we often want to use the same information model with more than one coding system, we want the “binding” between the

Interface between information models and ontologies

Example: "suspected active appendicitis"

- Epistemology:
- Knowledge / context
 - Model of use



HL7 FHIR Release 4B

9.2 Resource Condition - Content

Structure UML XML JSON Turtle R4 Diff All

Structure

Name	Flags	Card.	Type	Description & Constraints
Condition	TU		DomainResource	Detailed information about conditions, problems or diagnoses
identifier	Σ	0..*	Identifier	External Ids for this condition
clinicalStatus	?! Σ I	0..1	CodeableConcept	active recurrence relapse inactive remission resolved Condition Clinical Status Codes (Required)
verificationStatus	?! Σ I	0..1	CodeableConcept	unconfirmed provisional differential confirmed refuted entered-in-error ConditionVerificationStatus (Required)
category		0..*	CodeableConcept	problem-list-item encounter-diagnosis Condition Category Codes (Extensible)
severity		0..1	CodeableConcept	Subjective severity of condition Condition/Diagnosis Severity (Preferred)
code	Σ	0..1	CodeableConcept	Identification of the condition, problem or diagnosis Condition/Problem/Diagnosis Codes

- Ontology
- Entity types and their properties
 - Model of meaning
 - E.g. SNOMED CT codes
 - HL7 value sets

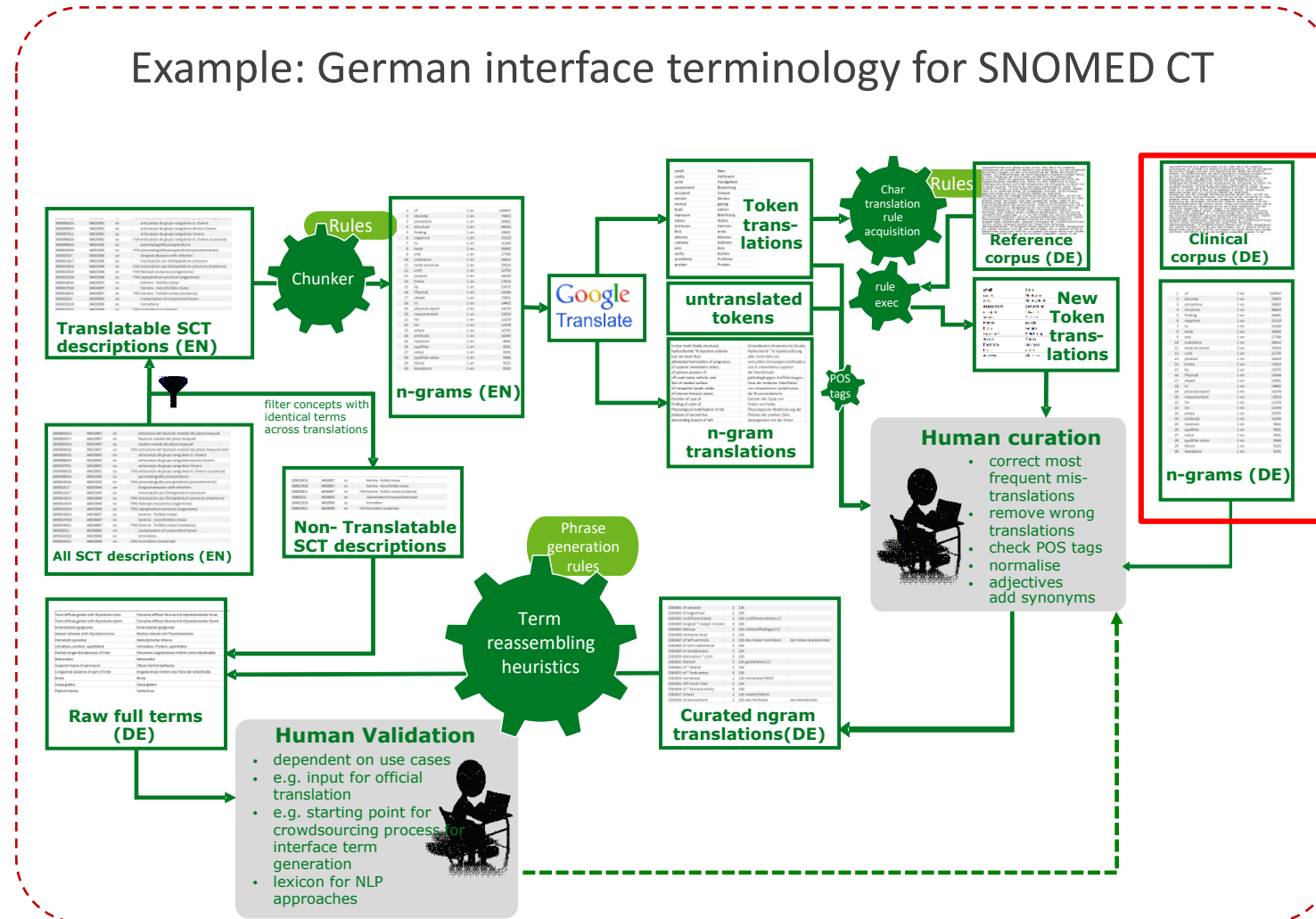
<https://hl7.org/fhir/codesystem-condition-clinical.html#condition-clinical-active>

<https://hl7.org/fhir/codesystem-condition-ver-status.html#condition-ver-status-unconfirmed>

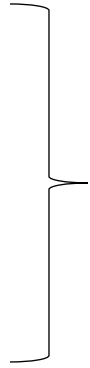
<http://snomed.info/id/74400008>

Which challenges have to be met
to promote ontology-based
data management
in biomedicine?

I – Building resources: language specific dictionaries and linking to ontologies



I – Building resources: language specific dictionaries and linking to ontologies

- Manual creation / maintenance
 - Language productivity / compositionality: impossible to collect all variations in a dictionary, as well as all ambiguous readings
 - Community processes (crowdsourcing)
- The potential of machine learning
 - Synonym detection
 - Machine translation
 - Spelling correction
 - Short form resolution
 - Word sense disambiguation

Clinical training data required
Still require human review
- Safe entity recognition and normalization still a long way to go

II – Building resources: annotated domain corpora, particularly clinical corpora

Annotation guideline for annotating clinical narratives according to SNOMED CT and FHIR

Akhila Naz Kuppassery, Alexander Beger, Larissa Hammer, Markus Kreuzthaler, Stefan Schulz

Version 20220628

Rationale

Annotated clinical corpora are the "fuel" for the successful use of text mining and AI for interpreting clinical texts and converting their content into an interoperable format such as [as given](#) by SNOMED CT and FHIR.

Guideline draft (to be discussed and enriched with examples)

[BROWSER] As a reference, the [SNOMED CT browser](#) (English) is used to find the correct codes. Only active content should be used. The decision for a code should be made according to the

- Wording of the Fully Specified Name of a concept
- The concept's text definition (if available)
- Its formal axioms
- Its parents and children

In case two concepts fit equally well, they can be added with an OR
In case the meaning of two concepts need to be combined, they should be added with an AND

Always copy and paste the ConceptID and the term

[PRECOORDINATION vs. POSTCOORDINATION]

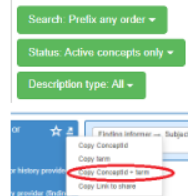
Always use pre-coordinated concepts if they represent the meaning of a text passage.

[PREFERENCES]

We use for our primary annotations concepts from the following hierarchies

- Clinical Conditions (SNOMED findings / disorders / events) related to morphology, anatomy, devices, organisms
- Procedures related to anatomy, devices, Procedures also used for measurement (as long as there are no observables available)
- Observables, together with qualitative values or numbers (+ units)

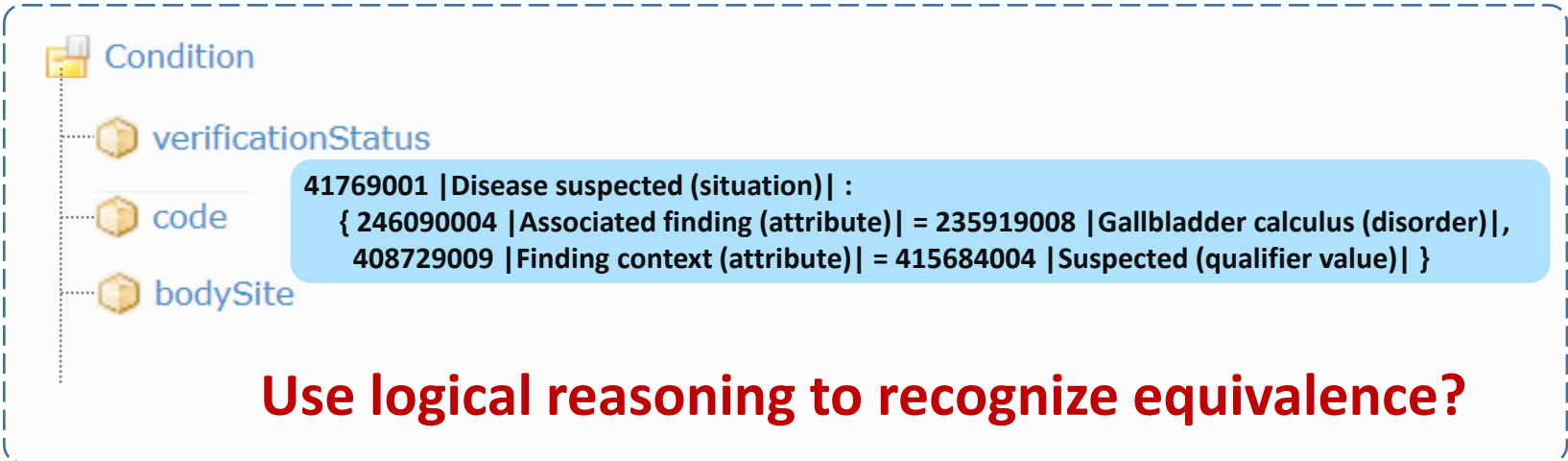
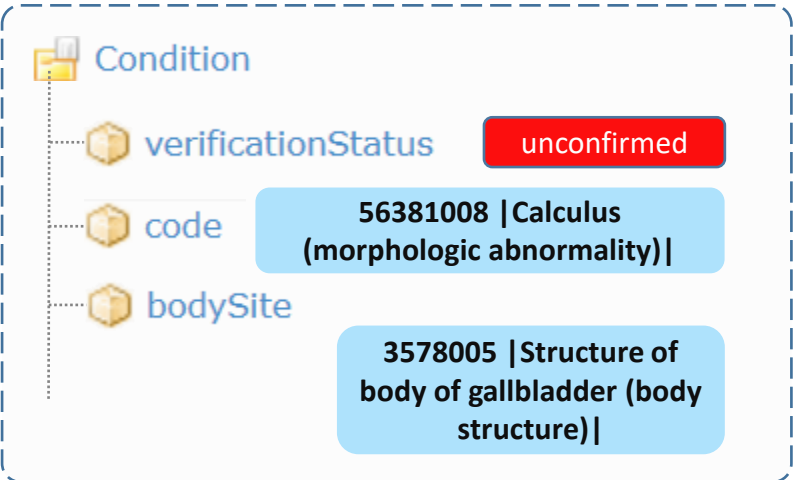
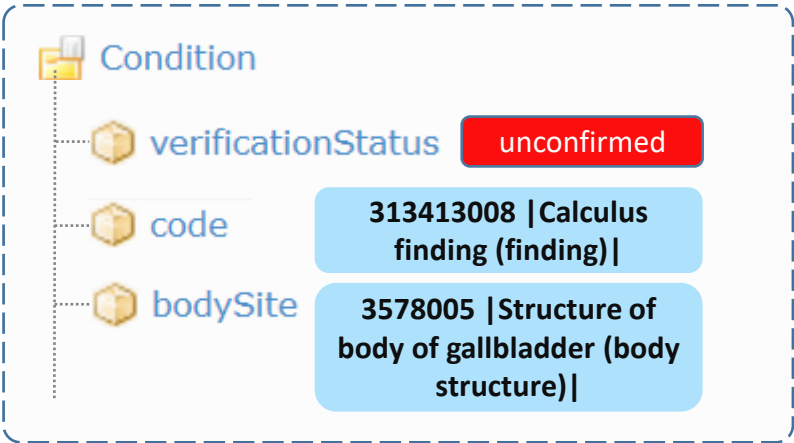
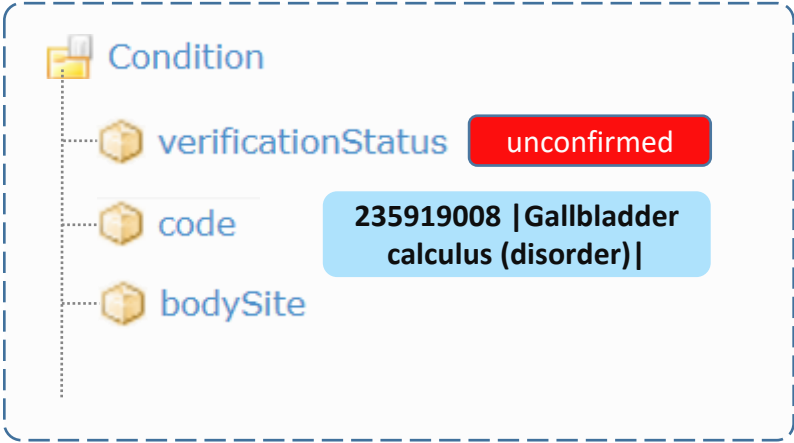
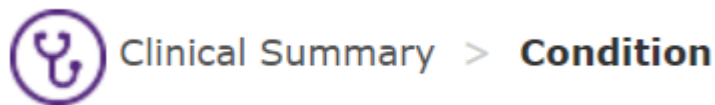
Others are only used in those cases where they are clinically important and not expressible with the above hierarchies, and when the interpretation of other parts of the text depends on it.



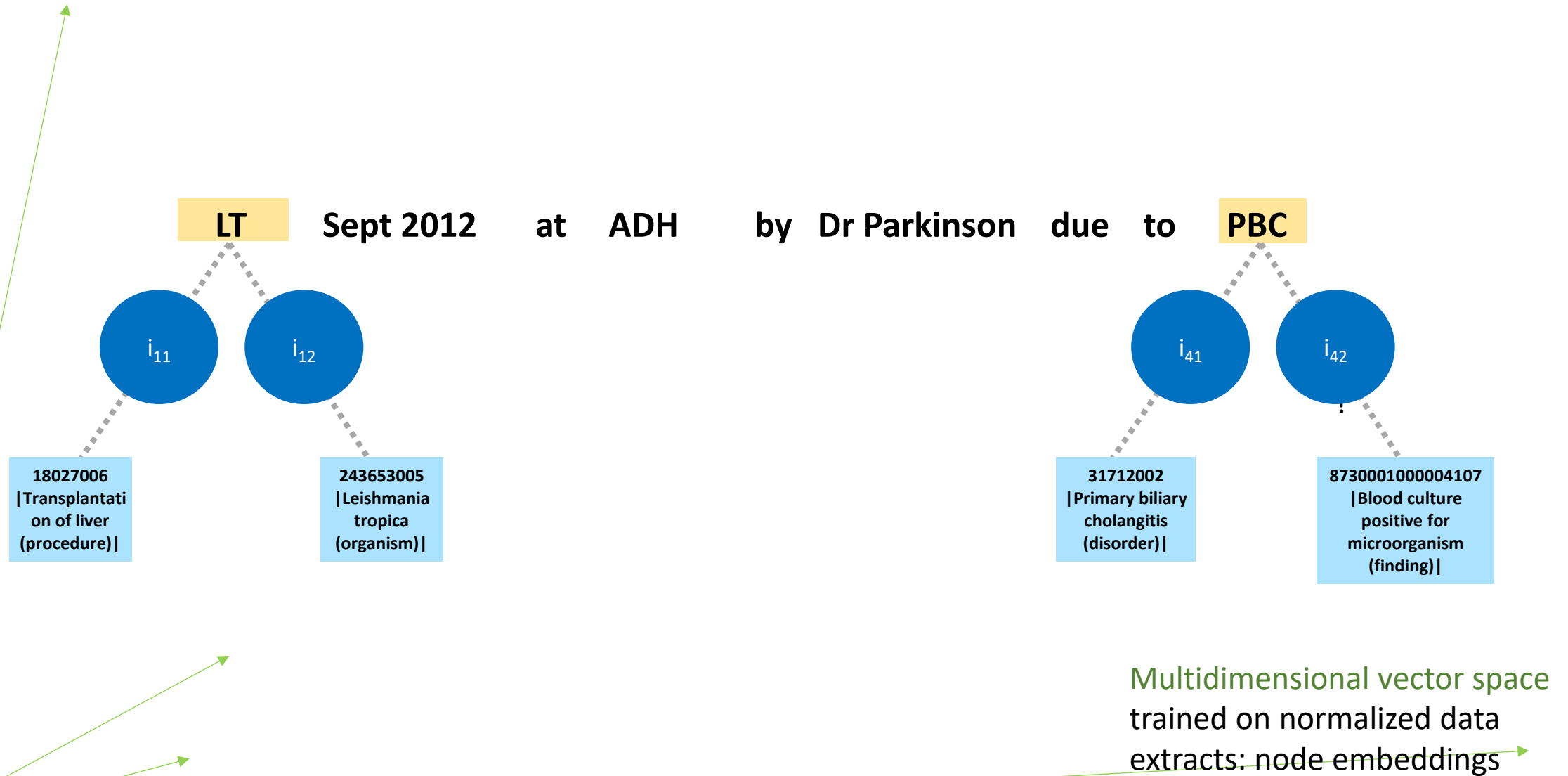
A screenshot of the INCEpTION interface. The top bar shows "INCEpTION" and navigation links for "Projects", "Dashboard", "Help", "Administration", "admin", "Log out", and a timer "29 min". The main area displays a clinical narrative titled "admin: Clinical Cases/Jadassohn.txt" with 21-30 / 48 lines. The narrative text is: "Sonografie vom 23.3.23: Homogene Leber, keine Raumforderungen, keine gestauten Gallenwege. Tumor sonografisch nicht darstellbar. Gefäße soweit beurteilbar frei. Nierenzyste links. Kontroll-Rektoskopie vom 23.3.23: Normaler Ruhe- und Kneiftonus. Anastomose unauffällig, kein Anhalt". The text is annotated with SNOMED CT concepts: "Diagnostic ultrasonography" (Finding method), "Normal" (value), "Structure of parenchyma of liver" (value), "refuted" (verification status), "Mass" (value), "refuted" (verification status), "Cholestasis" (value), "Follow-up visit" (Finding method), "Proctoscopy" (Finding method), "AND" (Finding method), "Anal tone normal" (value), "Anal sphincter squeeze tone" (value), "Normal" (value), "Anastomosis" (value), "Normal" (value), "refuted" (verification status), and "recurrence" (verification status). The right sidebar shows the "Layer" set to "SNOMED CT" and a list of "Relations: Finding method" including "Raumforderungen" and "gestauten Gallenwege". The bottom status bar shows "Technische Universität Darmstadt -- Computer Science Department -- INCEpTION -- 22.5 (2022-03-12 19:54:59, build a81eecf0)" and a "Warnings" icon.

Annotations should use the same semantic resources as expected for the processing of real data: in our case SNOMED CT and FHIR

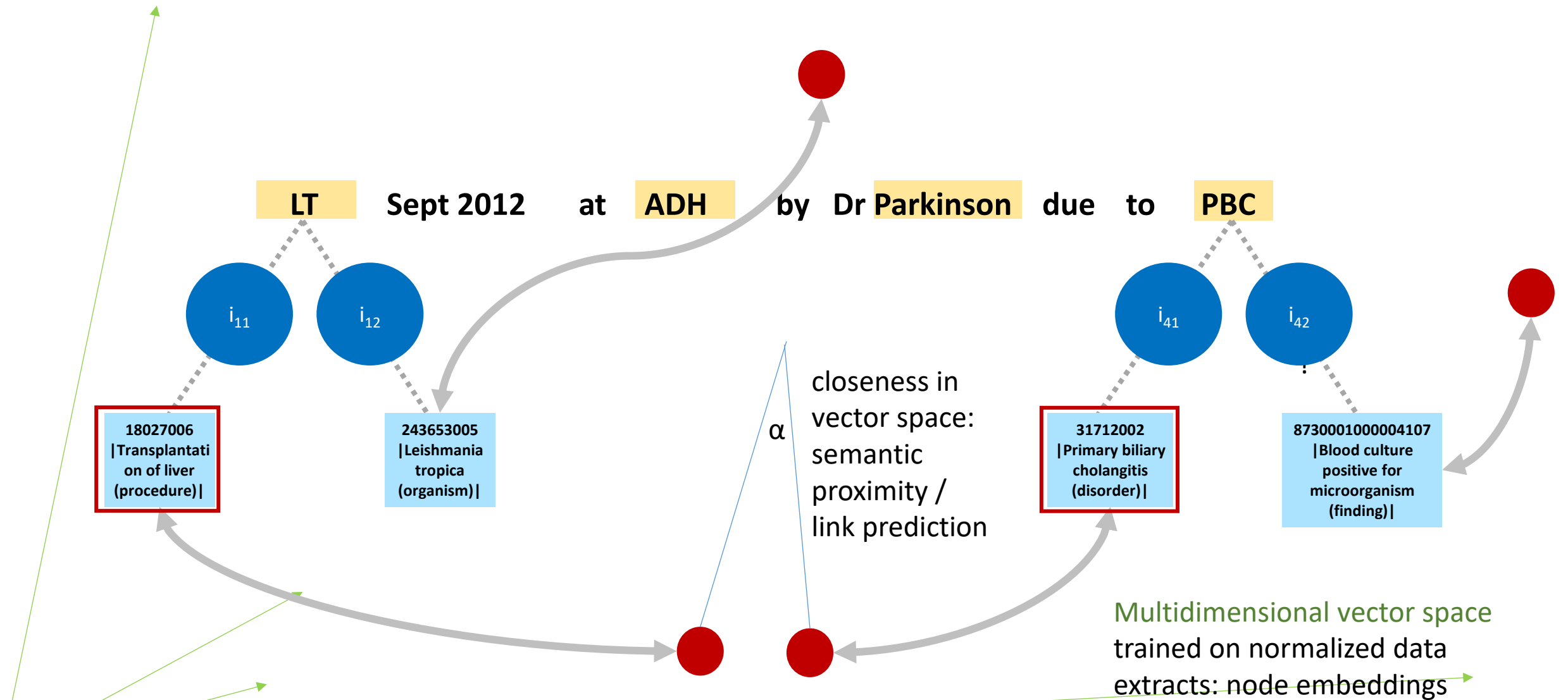
Resolve competing semantic representations



Word sense disambiguation via contextualised embeddings



Word sense disambiguation via contextualized embeddings



Conclusion – in a nutshell

- The medical domain is rich of semantic resources, but heterogeneous
- Biomedical ontologies are rich in axioms, e.g. SNOMED CT and OBO ontologies
- Most relevant information is in clinical narratives
- Clinical language is particularly hard to normalize and disambiguate
- Much needed:
 - Multilingual terminology resources
 - Annotated corpora for training models and benchmarking implementations
 - Scientific challenges for comparing and validating
 - Safe access to clinical data extracts for research
- Combination of symbolic with probabilistic / neural methods: to explore

Questions?

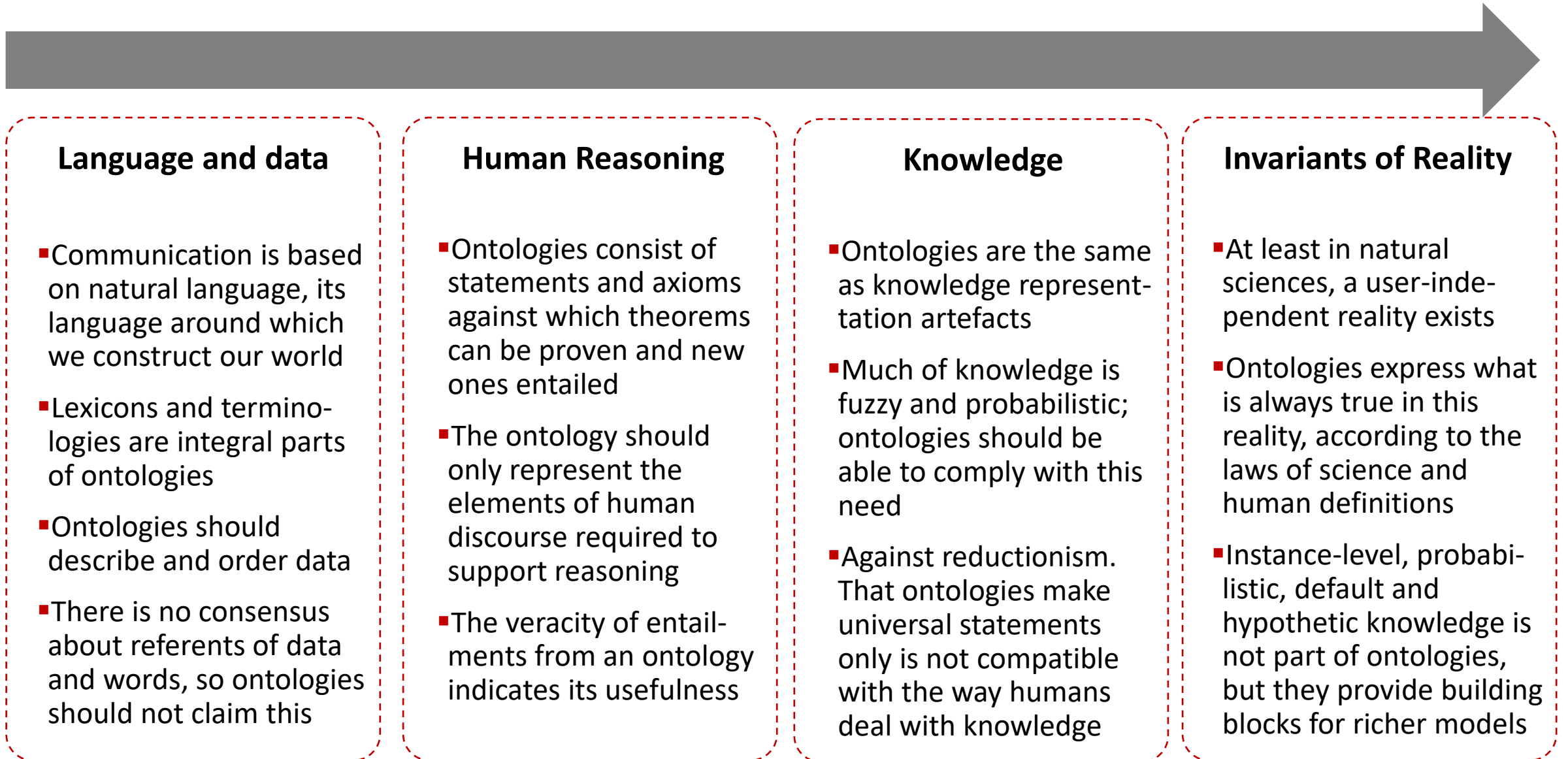
Stefan Schulz:



stefan.schulz@medunigraz.at

<http://purl.org/steschu>

What should ontologies represent?



Which logic is appropriate for biomedical ontologies?

No logic at all

- Meaning of domain terms is fuzzy and context-dependent, it cannot be generalised, otherwise meaning is arbitrary
- The bigger the system the more difficult to maintain it consistent
- Domain experts struggle with formality anyway
- The future is deep learning from data

Inexpressive description logics (OWL EL)

- Only a simple logic can be expected to be applied by domain experts in a consistent way
- Only inexpressive description logic offers the performance needed for automatic reasoning
- Sufficient for the really important jobs, such as reasoning with Aristotelian definitions

Expressive description logics (OWL DL)

- The domain is overly complicated, many domain concepts need to be defined by negation and concrete domains
- Current limits of reasoners may be overcome by new optimisation techniques, together with improved memory and computing power

First order logic

- Ontologies as standards do not allow definition gaps. Particularly the restriction of description logics to binary predicates is not acceptable.
- Important concepts are process-like (occurents / perdurants), therefore time as a third argument is indispensable for representation and reasoning

Do we need foundational (upper-level) ontologies ?

Not at all

- Will not be understood anyway, delays development workflows
- Burden that restricts freedom of the modeller
- The bigger the system the more difficult to maintain it consistent
- Domain terms are too ambiguous, which would conflict with upper-level constraints

Only if use cases require

- Foundational decisions have to be made with use cases in mind (e.g. whether an entity is a class or an instance)
- Modellers will only use them as external sets of constraints if there is a measurable benefit in terms of quality and productivity

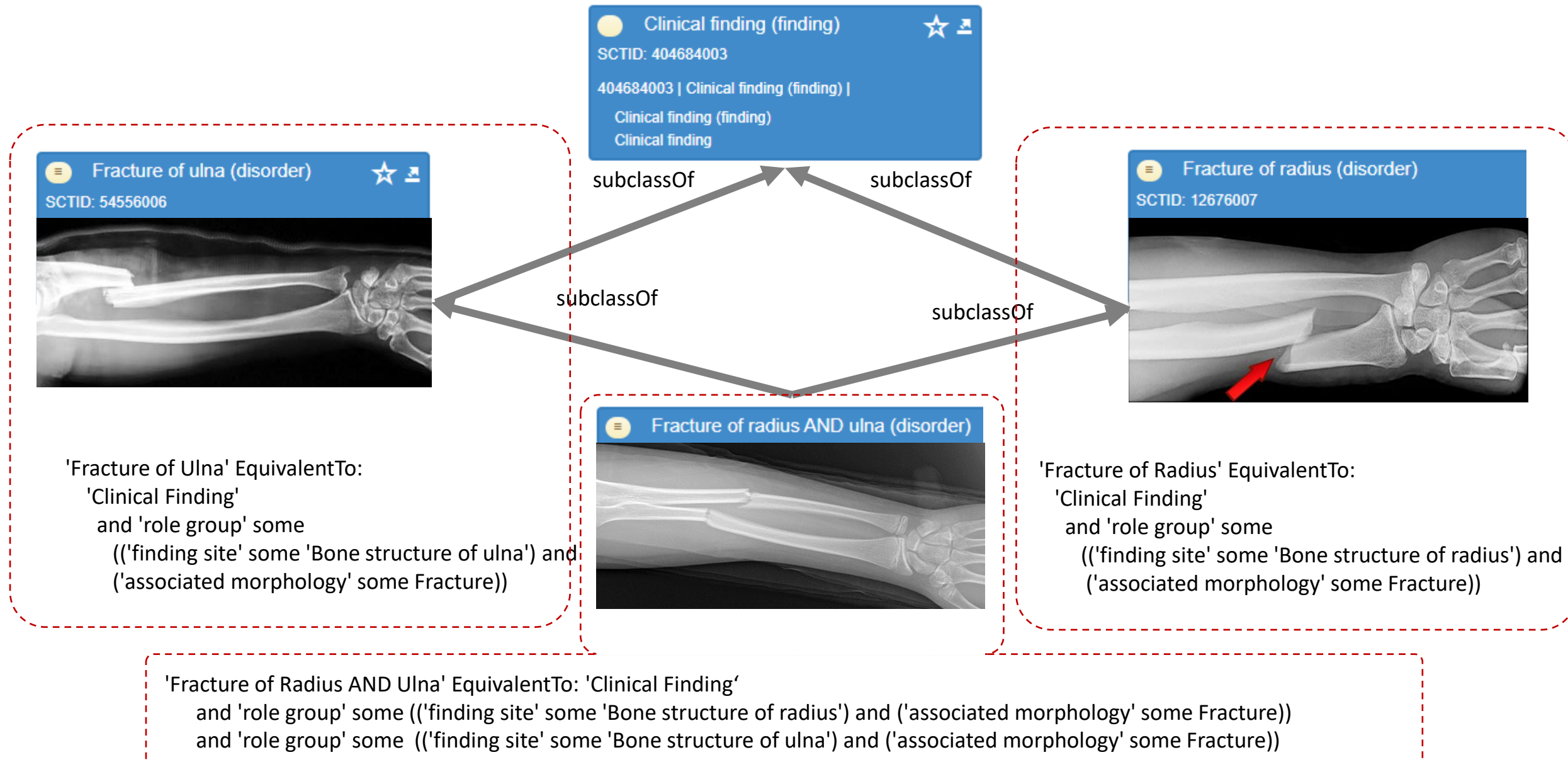
For each ontology

- Each ontology should have its own, well thought out system of upper-level classes
- An appropriate upper level supports modelling discipline, reduces arbitrary modelling decisions
- Apart from upper-level classes, relations with domain and range restrictions are necessary

For the whole domain

- Only if domain ontologies are modelled under a foundational level, interoperability between different levels can be achieved
- The creation of foundational ontologies demands high efforts, integration of many stakeholders and a strong anchoring in metaphysics
- Should become standards

Example - what is a clinical finding in SNOMED CT?



How reliable are ontology-based text annotations?

- Context: ASSESS-CT: EU support action on the fitness of **SNOMED CT** as a EU core reference terminology
- Experts annotate 60 clinical documents with SNOMED CT codes
- Support: Annotation guidelines, Webinars
- 1/3 of documents annotated twice for inter-annotation agreement

Nitroglycerin pump spray as required	387404004;385074009;225761000
Amantadine bds	372763006;229799001
Allopurinol 300 ½ tablet every other day (last dose on 20091130)	387135004;385055001;225760004
Mefenamic acid 500 mg up to 3x daily for pain in conjunction with	387185008;258684004;229798009;22253000
simultaneous administration of a drug to protect the stomach e. g.	79970003;416118004;373517009;69695003
Pantoprazole 40mg.	395821003;258684004
Torsemide bds	318034005;229799001
Melperone 50 mg p. m.	442519006;258684004;422133006
§ 7 Intact teeth are in the mouth.	11163003;245543004;123851003
Fractures are visible on the medians of Mandible and Maxilla	263172003;263156006;260528009
the fragments are dislocated.	123735002

Concept coverage [95% CI]



86% [82-88 %]

Term coverage (EN) [95% CI]



68 % [.64-70 %]

Inter annotator agreement
Krippendorff's Alpha [95% CI]



37% [33-41 %]

What is the right strategy for entity recognition and normalization?

Terminology engineering

- Collecting terminology used by clinicians and researchers
- Assigning ontology IDs
- Problem: terminology constantly changing and increasing
- Constant maintenance cost
- Acronyms and other short forms are ambiguous
- Spelling variants
- Risk of content explosion

Rule-based systems

- Recurrent rules for term formation, e.g. in case of medication statements, lab results
- Capitalising on domain knowledge by experts more cost-effective than large amounts of training data

Machine learning

- State-of-the-art: neural network embeddings: mapping semantics to vectors in a multidimensional space
- Compute similarity, e.g. of language expressions: resolution of synonyms in context
- Require huge amounts of training data: see power of Google translate
- Problem: privacy: restricts the use not only of data but also of models

Hybrid approaches?

- Speeding up construction of annotated corpora: pre-annotations, then correction by humans
- Large terminologies to create variations and therefore more training data instances
- Safe de-identification techniques and IRB approval to use clinical narratives for training language models