Joint Conference
ICBO - ONTOBRAS
2023

Towards principles of ontology-based annotation of clinical narratives

Stefan Schulz[1,2]
Warren Del-Pinto[3]
Markus Kreuzthaler[1]
Lifeng Han[3]
Sareh Aghaei[1]
Goran Nenadic[3]

[1]Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Austria

[2]Averbis GmbH, Freiburg, Germany

[3]Department of Computer Science, University of Manchester, UK

Med Uni Graz

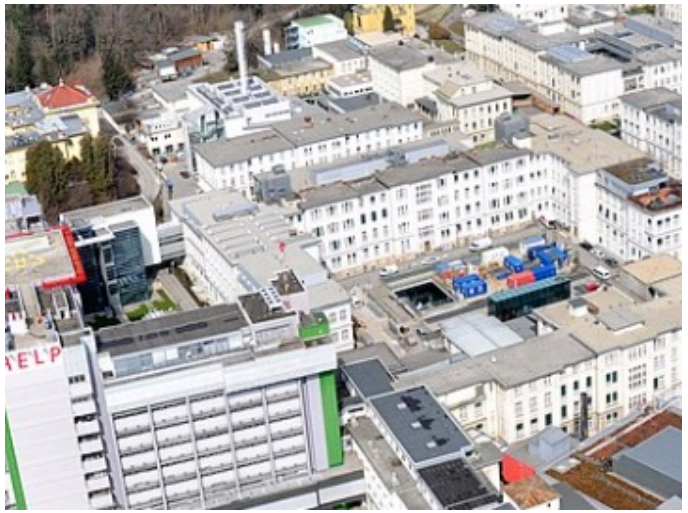
averbis
text analytics

MANCHESTER 1824
The University of Manchester

Most information in health records is “locked”in narratives

… in the local languages / sociolects

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

- Works well for expert-to-expert communication

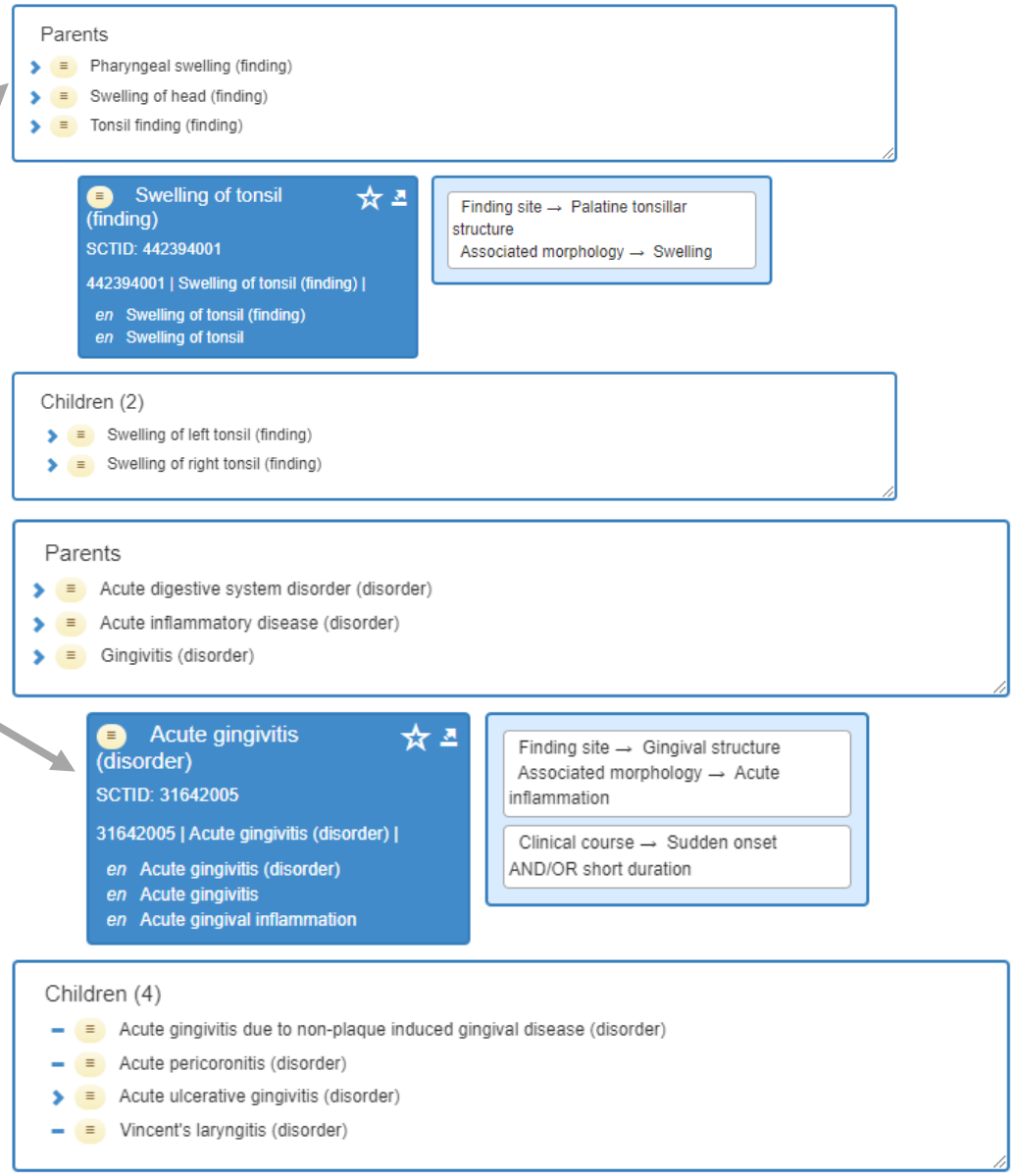
Phenomenon	Example	Elucidation
Telegram style	“left PICA stroke, presented to ED after fall”	Incomplete sentences, sketchy style
Colloquialisms	“pothole sign”, “snorkel”	Milieu-specific sub-languages
Ad-hoc abbreviations	“infiltr”	Truncation (“infiltrated mucosa”)
Ambiguous short forms	“RTA”	“Road traffic 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. 2nd 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

- Major interoperability bottleneck for machine processing

Desideratum: making unstructured health record data interoperable

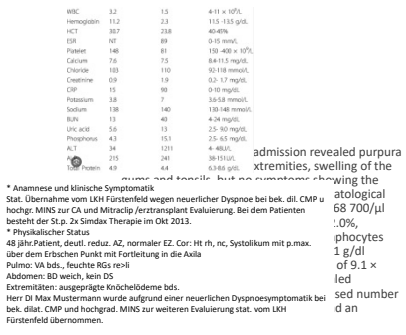
- Using international standards
- Rooted in Applied Ontology principles
- Information extraction via NLP
  (Natural language processing)

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 hypercelllar bone marrow with a decreased number of erythroblasts and megakaryocytes and an increased number of monoblasts



SNOMED CT large clinical ontology
(350k concepts, > 1M English terms)

Clinical knowledge graph as canonical content representations



admission revealed purpura
xtremities, swelling of the
wing the
atological
68 700/µl
:.0%,
phocytes
1 g/dl
of 9.1 ×
led
sed number
d an

* 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 Axilla
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.

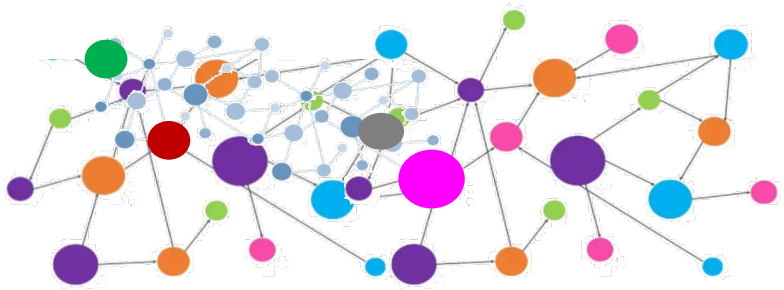
Heterogeneous clinical data

NLP

Tools & resources

„Models of meaning“* SNOMED CT

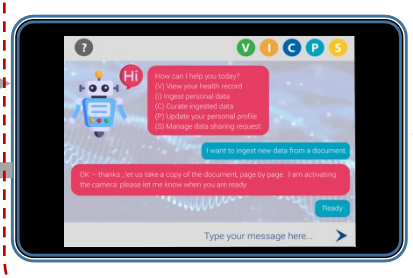
„Models of use“* HL7® FHIR®



Standardised and contextualised data

raw

curated



Health data curators

Curation platform

- Primary use:
  - Document Retrieval
  - Question Answering
  - Content summarization
  - Information extraction
  - Input for decision support
  - Data visualization, navigation

Query

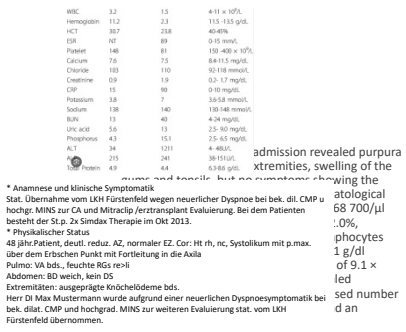
aggregated knowledge graph

Query

- Secondary use:
  - Retrospective research
  - Patient recruiting
  - Training of predictive models
  - Training of decision support systems
  - Quality improvement

* A. L. Rector, R. Qamar, T. Marley, Binding ontologies and coding systems to electronic health records and messages, Applied Ontology 4 (2009) 51–69.

Natural Language Processing : clinical narratives → knowledge graphs



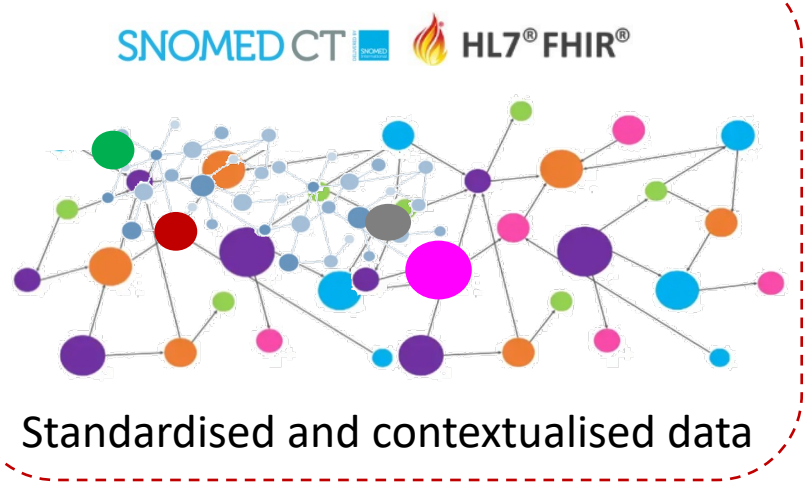
admission revealed purpura
xtremities, swelling of the
owing the
atological
68 700/µl
:.0%,
phocytes
1 g/dl
of 9.1 ×
led
sed number
d an

* 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.

Heterogeneous clinical data

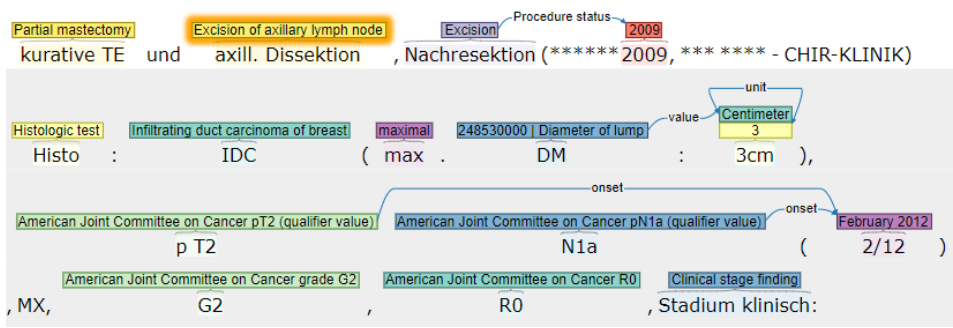
NLP

Tools & resources



Standardised and contextualised data

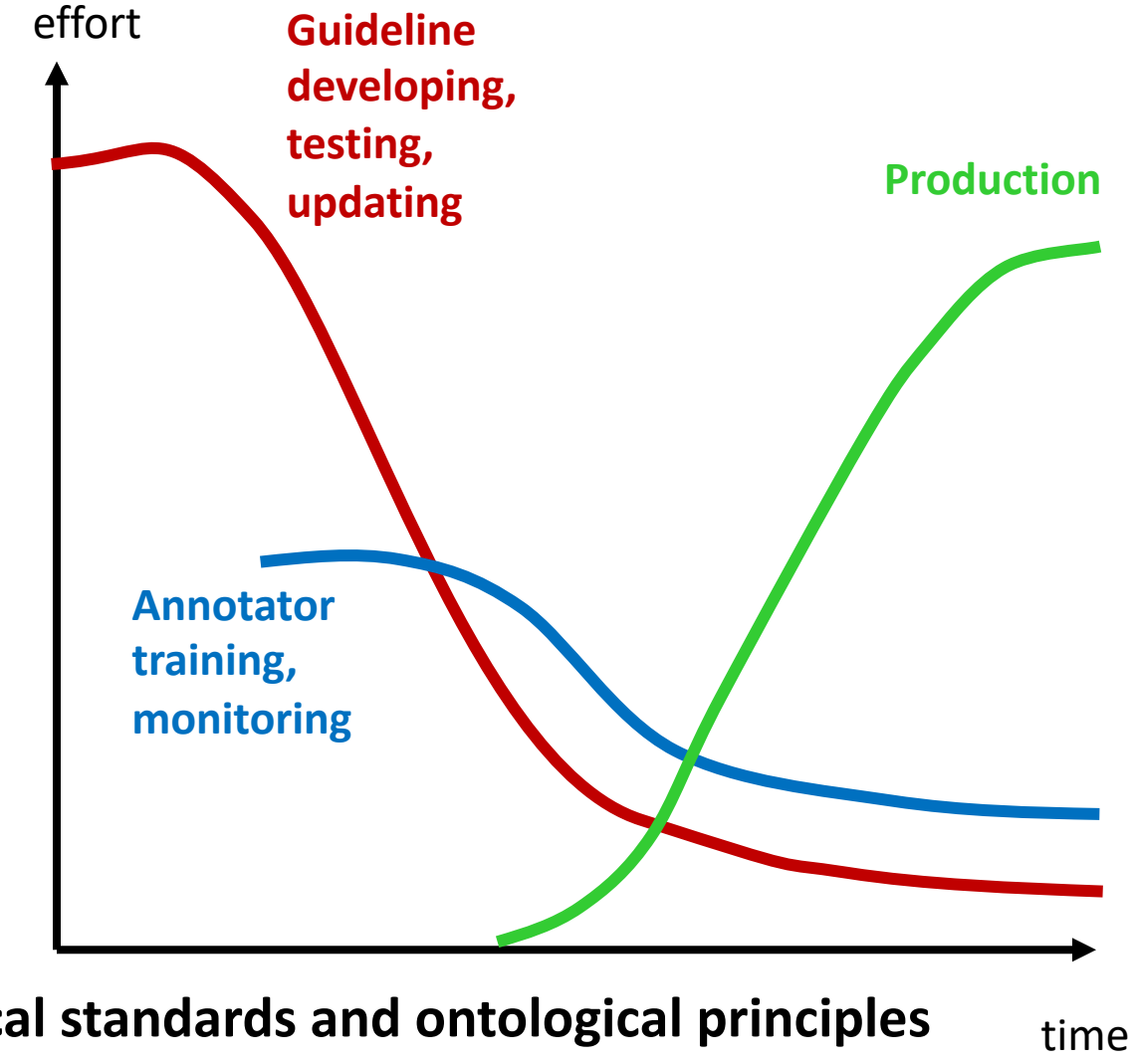
- NLP methods
  - Classical pipelines (sequential processing steps), rules, lexicons
  - End-to-end approaches: single architecture directly maps input text to the desired output: Deep learning, large language models
- **Annotated clinical corpora: central resource**
  - Training, model fine-tuning
  - Benchmarking of NLP systems



Document annotation – knowledge acquisition bottleneck

- Requirements
  - Domain expertise
  - Extensive training
  - Motivation
- Problems
  - Inter-annotator variability
  - Annotation fatigue
  - Ambiguities
  - Time constraints
- Success factors:
  - Good tooling
  - Repeated training sessions
  - Adjudication between annotators
  - Quality checks (inter-annotator agreement)
  - Good communication channels
  - **Rigorous annotation guidelines based on clinical standards and ontological principles**

medical students



Annotation guideline browser, annotation tool

Annotation guideline for semantic annotations of clinical narratives based on SNOMED CT and FHIR

Stefan Schulz[1,2], Akhila Naz Kuppassery[1], Alexander Beger[1], Sareh Aghaei[1], Daniel Dür[1], Larissa Hammer[1], Kristian Kankainen[3], Markus Kreuzthaler[1], NN, NN...

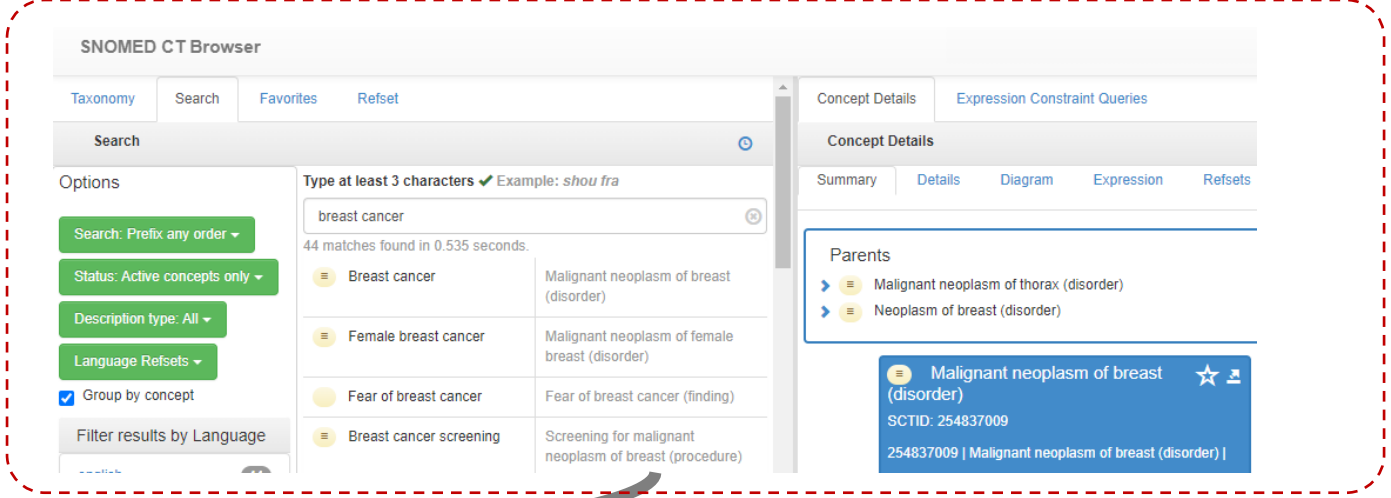
[1]Medical University of Graz, Austria
[2]Averbis GmbH, Freiburg, Germany
[3]Tallin University of Technology, Estonia

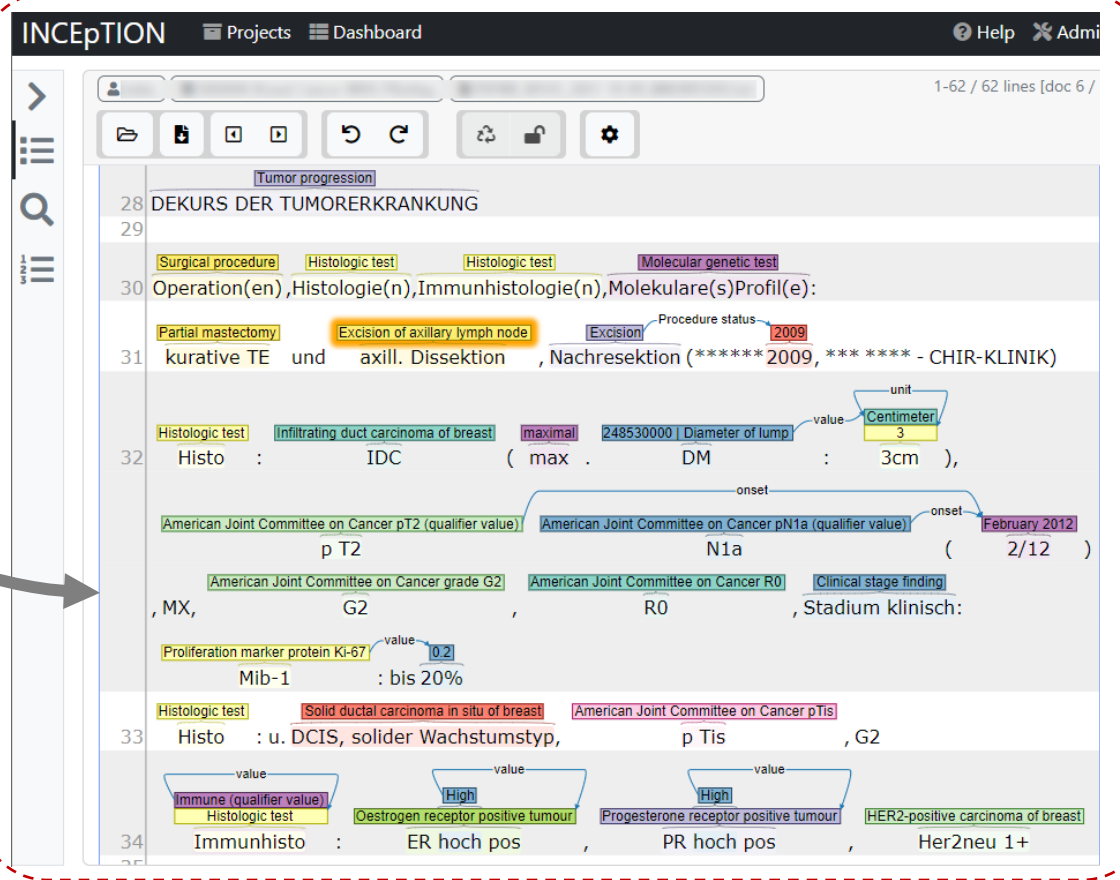
Version 2023-06-14

Guests are welcome! Feel free to drop comments.
E-mail contact: stefan.schulz@medunigraz.at

In progress: consolidation of practical content for insertion into AIDAVA D4.2

https://bit.ly/3X4McGC





Annotation guideline for semantic annotations of clinical narratives based on SNOMED CT and FHIR

Stefan Schulz[1,2], Akhila Naz Kuppassery[1], Alexander Beger[1], Sareh Aghaei[1], Daniel Dür[1], Larissa Hammer[1], Kristian Kankainen[3], Markus Kreuzthaler[1], NN, NN…

[1]Medical University of Graz, Austria
[2]Averbis GmbH, Freiburg, Germany
[3]Tallin University of Technology, Estonia

Version 2023-06-14

Guests are welcome! Feel free to drop comments.
E-mail contact: stefan.schulz@medunigraz.at

In progress: consolidation of practical content for insertion into AIDAVA D4.2

https://bit.ly/3X4McGC

General annotation principles

- Semantic annotation only (no POS, syntactic relations etc.)
- Annotation at two levels
  - Text spans (“entities”) with codes and literals
  - Binary relations with user-friendly predicates (hide complexity from annotators): (i) semantic relations (ii) co-reference annotations
- Annotation vocabularies linked to ontology-based standards
  - SNOMED CT, FHIR, HPO, RxNorm …
  - multilingual, well-curated, free, ontology-based, compositional (post-coordination)
- Annotation vocabulary determines
  - Annotation spans (subword to multiword): longest match preference
  - Granularity and scope
- Close-to-text annotation
  - no interpretation by annotators

Specific annotation principles (SNOMED CT + FHIR annotation)

- "Core" hierarchies:
  - *Clinical finding*, *Event*, *Observable entity*, *Pharmaceutical / biologic product*, *Procedure*, *Specimen*
  - High proportion of fully defined concepts
- "Supportive" hierarchies
  - *Substance*, *Organism*, *Body structure*, *Physical object, Qualifier*
  - Primitive concepts
  - HL7-FHIR values sets mapped to SNOMED concepts
- Annotation predicates (binary relations) link "core" concepts to "supportive" concepts, grounded in
  - SNOMED CT object properties or chains thereof,
  - Relational chains of FHIR elements
  - both

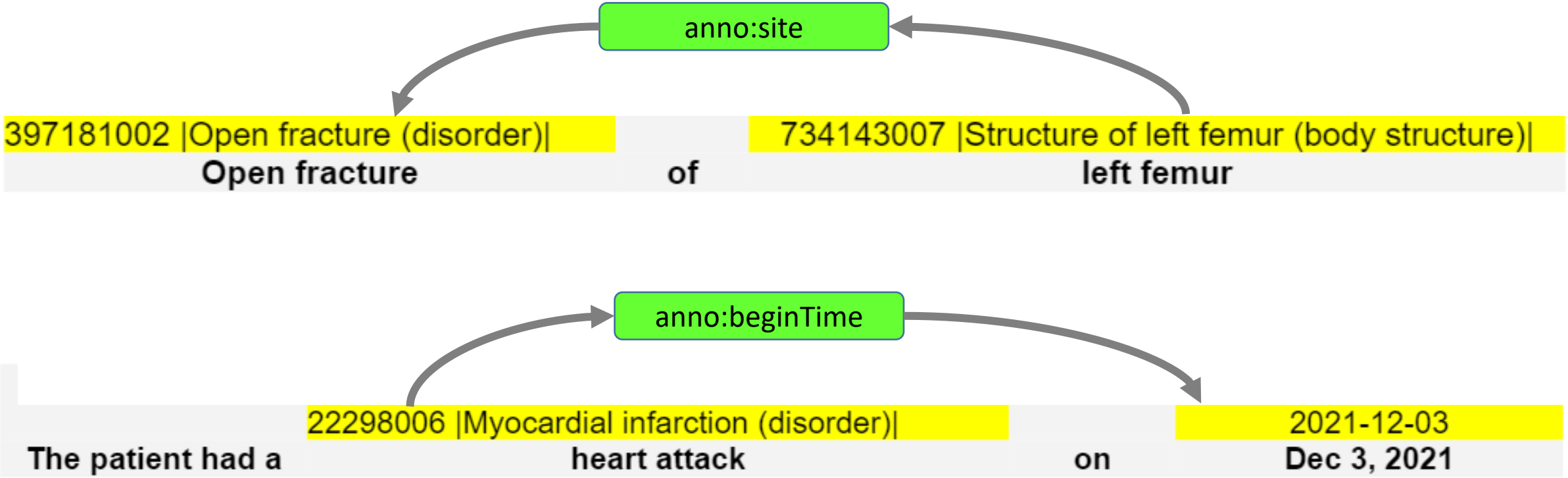
Relation annotation vocabulary based on SNOMED CT and FHIR

- Close-to-user predicates
- Mapped to relations or relation chains in underlying standards

anno:	Domain	Target path	Range
site	‘sct:Clinical finding’	[a]‘sct:Finding site’ [b] INV(fhir:Condition.code) || fhir:Condition.body	‘sct:Body structure’
site	‘sct:Procedure’	[a]‘sct:Procedure - Direct’ [b] INV(fhir:Procedure.code) || fhir:Procedure.body	‘sct:Body structure’
inFamily	‘sct:Clinical finding’	[b] INV(fhir:FamilyMemberHistory.condition) || fhir:FamilyMemberHistory.relationship [a] INV(‘sct:Associated finding’) || ‘sct:Subject relationship context’	‘sct:Person’
verification status	‘sct:Clinical finding’	[b] INV(fhir:Condition.code) || fhir:Condition.verificationStatus [a] INV(‘sct:Associated finding’) || ‘sct:Finding context’	‘sct:| Qualifier value’ (cf. Tab. 1)

Example 1: “Two level” annotation

1. Text spans, annotated with codes or literals
2. Linkage of text spans by binary predicates



Example 2: Deep annotation

Annotations exploit the whole depth of the annotation vocabulary

No “entity-type” annotation

1223620100011910 3 |Conjunctivitis of right eye (disorder)|
Unilateral conjunctivitis right

Not:

Disorder
Unilateral conjunctivitis right

Example 3: Flexible annotation spans / longest match principle

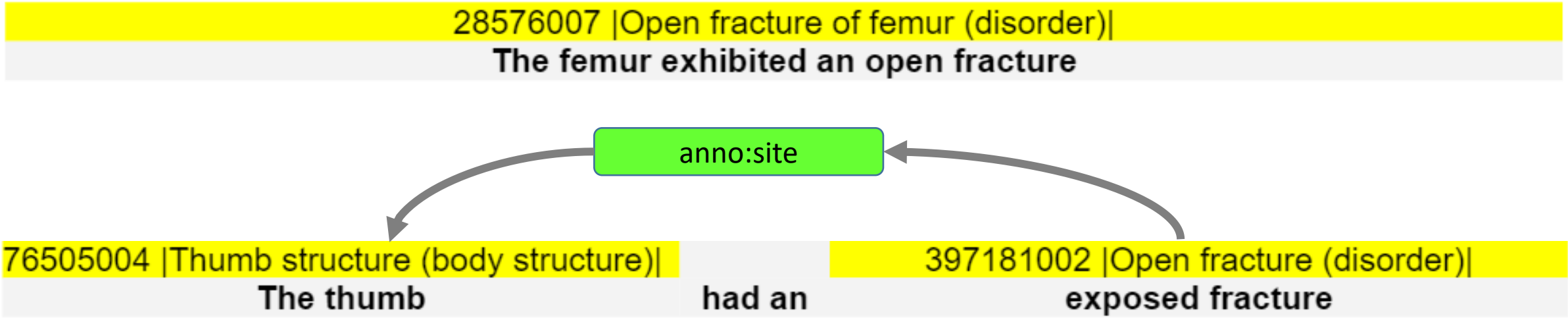
- Annotations spans determined by annotation vocabulary
- Preference given to longest match (precoordinated concepts)
- No determination of spans by NER before annotation

28576007 |Open fracture of femur (disorder)|

The femur exhibited an open fracture

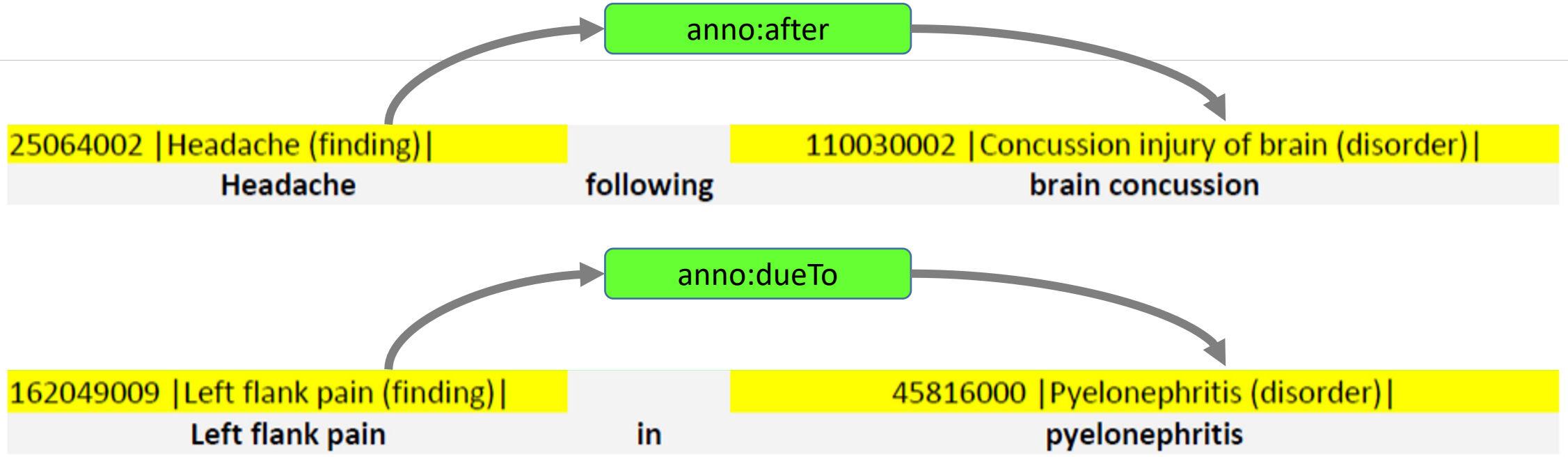
Example 3: Flexible annotation spans / longest match principle

- Annotations spans determined by annotation vocabulary
- Preference given to longest match (precoordinated concepts)
- No determination of spans by NER before annotation



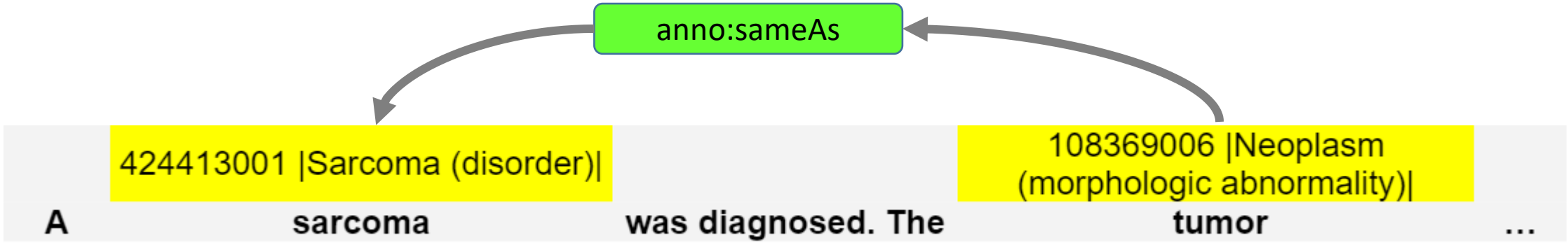
Example 4: Close-to-text: no interpretation of content

- Only annotate what is explicitly stated, not what might be medically plausible



Example 5: Coreference annotations

- Nominal anaphora



From text to canonical representation

Text

Words
phrases
numeric expressions

Anntotations

Codes for classes
Annotation predicates
literals

Knowledge graph

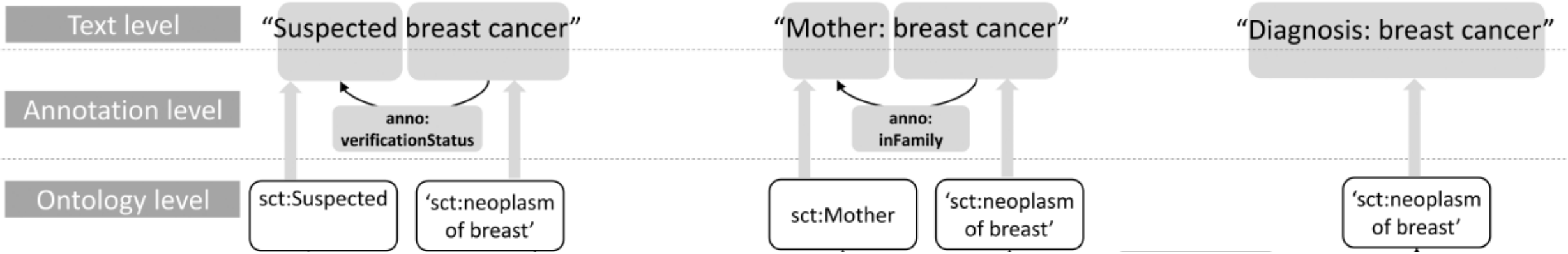
Semantic interpretations
Classes vs. Individuals
“Isosemantic” reporesentations
(e;g; SNOMED only vs. SNOMED + FHIR)

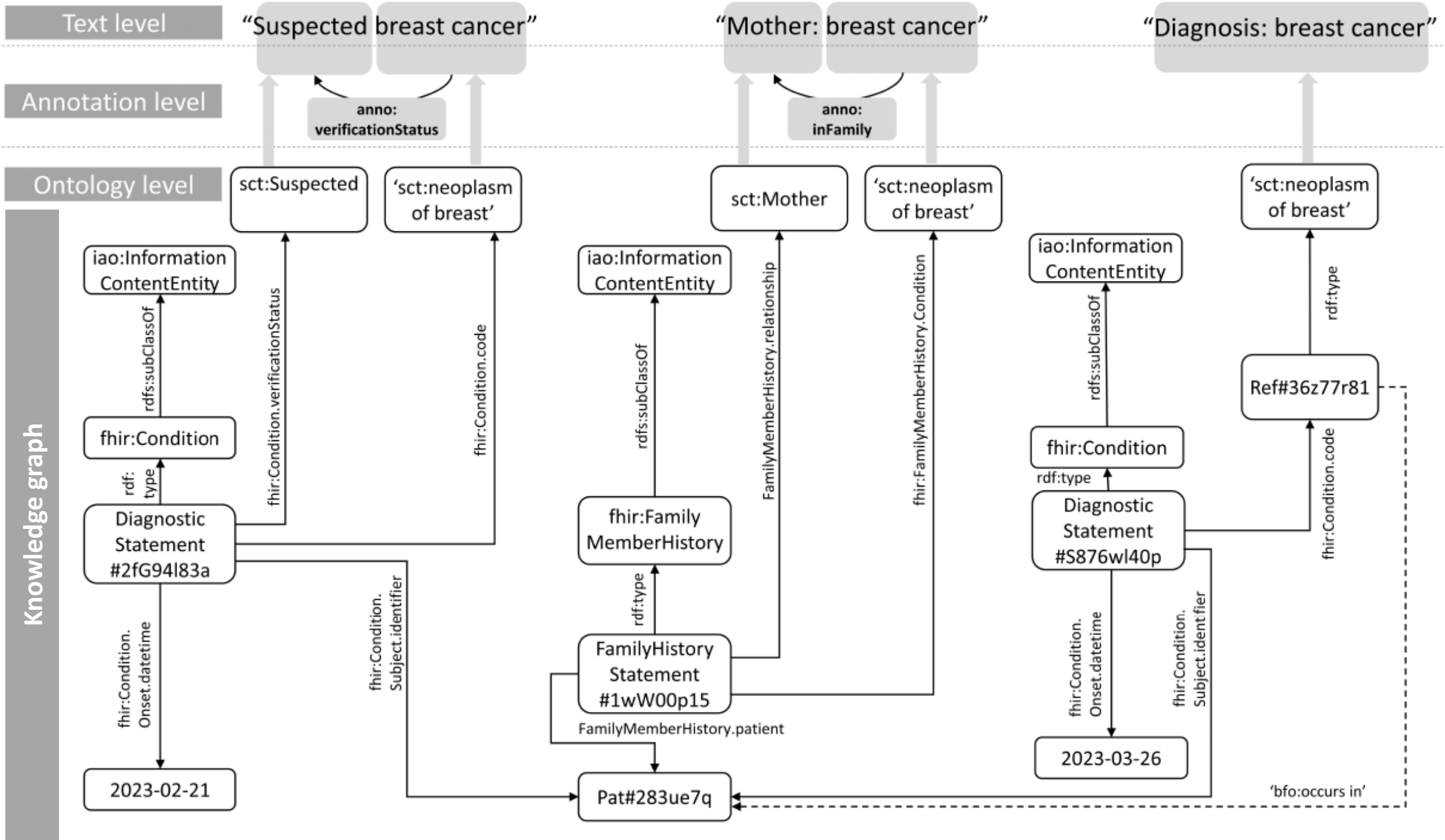
Text level

“Suspected breast cancer”

“Mother: breast cancer”

“Diagnosis: breast cancer”


Text level
Annotation level
Ontology level
“Suspected breast cancer”
“Mother: breast cancer”
“Diagnosis: breast cancer”
anno: verificationStatus
anno: inFamily
sct:Suspected
‘sct:neoplasm of breast’
sct:Mother
‘sct:neoplasm of breast’
‘sct:neoplasm of breast’


Text level
Annotation level
Ontology level
Knowledge graph
“Suspected breast cancer”
“Mother: breast cancer”
“Diagnosis: breast cancer”
anno: verificationStatus
anno: inFamily
sct:Suspected
‘sct:neoplasm of breast’
sct:Mother
‘sct:neoplasm of breast’
‘sct:neoplasm of breast’
iao:Information ContentEntity
iao:Information ContentEntity
iao:Information ContentEntity
rdfs:subClassOf
fhir:Condition.verificationStatus
fhir:Condition.code
rdfs:subClassOf
FamilyMemberHistory.relationship
fhir:FamilyMemberHistory.Condition
rdfs:subClassOf
rdf:type
fhir:Condition
rdf: type
fhir:Family MemberHistory
fhir:Condition
rdf:type
Ref#36z77r81
Diagnostic Statement #2fG94l83a
rdf:type
FamilyHistory Statement #1wW00p15
Diagnostic Statement #S876wl40p
fhir:Condition. Onset.datetime
fhir:Condition. Subject.identifier
FamilyMemberHistory.patient
fhir:Condition. Onset.datetime
fhir:Condition. Subject.identifier
fhir:Condition.code
2023-02-21
Pat#283ue7q
2023-03-26
‘bfo:occurs in’

Conclusion and outlook

- Annotated corpora are essential for training and benchmarking NLP tools, particularly in the current era of deep learning and large language models
- Semantic resources / ontologz-based standards are crucial:
  - Ontologies (description of entity types): Definitions / Axioms
  - Terminologies (description of natural language): labels, synonyms
  - Information Models (Instance-level templates, link to ontologies and values)
- Clinical free text annotation is a huge and challenging task. Facilitated by
  - Pre-annotations using existing NLP annotators
  - Simple, intuitive set of predicates that map to more complex graph structures in the background
- Adherence to detailed guideline principles
  - might take a long journey
  - Indispensable for high agreement between annotators → canonical clinical content representations

Contact: Stefan Schulz

stefan.schulz@medunigraz.at

http://purl.org/steschu

Comment on our annotation guideline:

https://bit.ly/3X4McGC

Acknowledgements:

- Grant 101057062 “AIDAVA” (funder: the European Commission, HORIZON-HLTH-2021,
- Grant “Assembling the Data Jigsaw: Powering Robust Research on the Causes, Determinants and Outcomes of MSK Disease” (funder: The Nuffield Foundation)
- Grant EP/V047949/1 “Integrating hospital outpatient letters into the healthcare data space” (funder: UKRI/EPSRC).

Annotation guideline for semantic annotations of clinical narratives based on SNOMED CT and FHIR

Stefan Schulz[1,2], Akhila Naz Kuppassery[1], Alexander Beger[1], Sareh Aghaei[1], Daniel Dür[1], Larissa Hammer[1], Kristian Kankainen[3], Markus Kreuzthaler[1], NN, NN…

[1]Medical University of Graz, Austria
[2]Averbis GmbH, Freiburg, Germany
[3]Tallin University of Technology, Estonia

Version 2023-06-14

Guests are welcome! Feel free to drop comments.
E-mail contact: stefan.schulz@medunigraz.at

In progress: consolidation of practical content for insertion into AIDAVA D4.2

Semantic equivalences

Clinical Summary > **Condition**

HL7 FHIR Release 4

Condition
- verificationStatus
- code: 390926006 |Suspected gallstones (situation)|
- bodySite

Condition
- verificationStatus: unconfirmed
- code: 235919008 |Gallbladder calculus (disorder)|
- bodySite

Condition
- verificationStatus: unconfirmed
- code: 313413008 |Calculus finding (finding)|
- bodySite: 3578005 |Structure of body of gallbladder (body structure)|

Condition
- verificationStatus: unconfirmed
- code: 56381008 |Calculus (morphologic abnormality)|
- bodySite: 3578005 |Structure of body of gallbladder (body structure)|

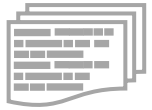
Condition
- verificationStatus
- code
- bodySite

```
EquivalentClasses(
    :41769001 |Disease suspected (situation)|
    ObjectIntersectionOf(
        :444433005 |Suspected clinical finding (situation)|
        ObjectSomeValuesFrom(
            :609096000 |Role group (attribute)|
            ObjectIntersectionOf(ObjectSomeValuesFrom(
                :246090004 |Associated finding (attribute)| :64572001 |Disease (disorder)|)
            ObjectSomeValuesFrom(:408729009 |Finding context (attribute)| :415684004 |Suspected (qualifier value)|)
            ObjectSomeValuesFrom( :408731000 |Temporal context (attribute)| :410512000 |Current or specified time (qualifier value)|)
            ObjectSomeValuesFrom( :408732007 |Subject relationship context (attribute)| :410604004 |Subject of record (person)| )))))
```

Open issue: Identity management

t_1 

Biopsy

t_2 

Imaging

t_3 

Operation

t_4 

126926005 |Neoplasm of breast (disorder)|

rdf:type

NoB34u73axn4us

owl:sameAs

254837009 |Malignant neoplasm of breast (disorder)|

rdf:type

MoBkj88935elp

owl:sameAs

278054005 |Lobular carcinoma of breast (disorder)|

rdf:type

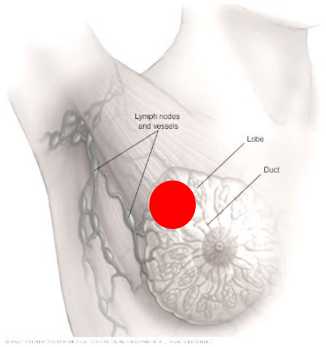
LCBrrp009g65t

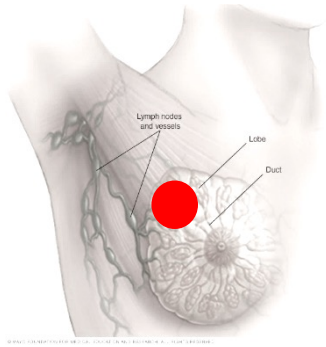
owl:sameAs

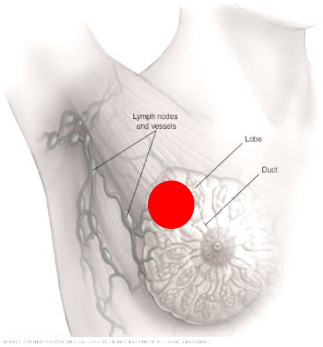
713609000 |Invasive carcinoma of breast (disorder)| AND 278054005 |Lobular carcinoma of breast (disorder)|

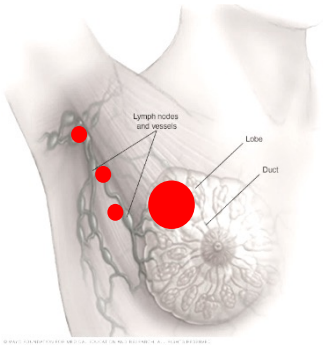
rdf:type

ILCB4tz5pplkll









Lobular carcinoma of breast

Invasive Lobular carcinoma of breast