AI powered Data Curation & Publishing Virtual Assistant

Optimize interoperability & quality of health data to increase data sharing and reuse across Clinical Registries & Personal Data Intermediaries

---

WP 5 Workshop

Annotation Process and NLP Workflow

Markus Kreuzthaler, Sareh Aghaei, Kris Collins, Todor Primov, Stefan Schulz

Funded by the European Union

13 May 2023 10:45 – 12:30 Vienna

Importance of semantic annotation in AIDAVA

- Large parts of EHR content only available as free-text
- Manual markup of concepts and relations of and between text elements (words, word sequences) in clinical narratives
- Purpose:
  - Training / domain-fine-tuning NLP approaches for information extraction for AIDAVA
  - Gold standard for evaluating these approaches
- Annotation vocabulary
  - Concept and relation names/codes used for annotation, consistent with AIDAVA reference ontology
- Pre-annotation
  - Use of existing text analytics pipelines to facilitate the annotation process
- Annotation guideline
  - Set of instructions for training annotators
  - Consistent and comparable annotation results

Task 4.3 Current state: guidelines, tools & training

Call: HORIZON-HLTH-2021-TOOL-06
Topic: HORIZON-HLTH-2021-TOOL-06-03
Funding Scheme: HORIZON Research and Innovation Actions (RIA)

Grant Agreement no: 101057062

AI powered Data Curation & Publishing Virtual Assistant

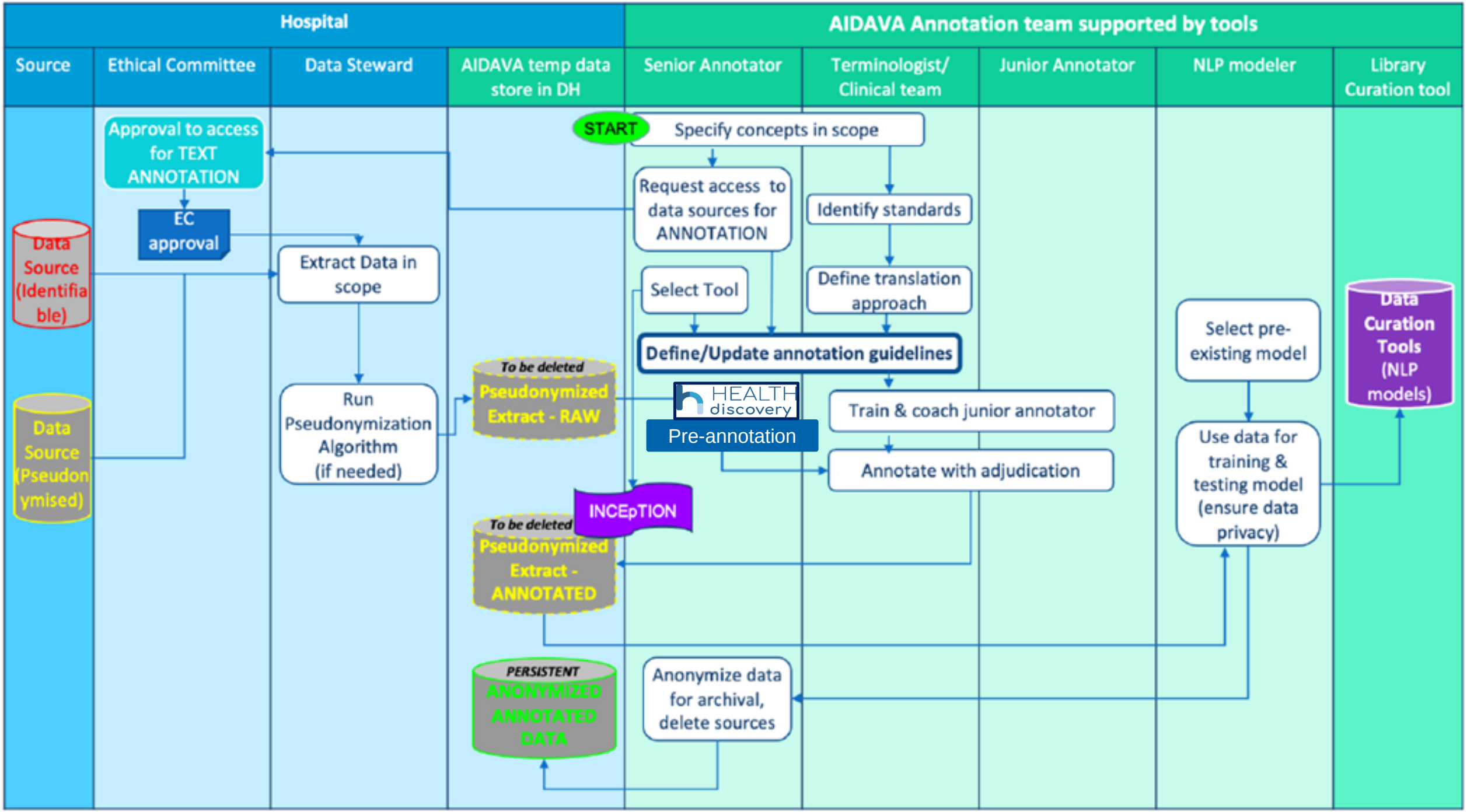
Deliverable No. 4.3
Update to Annotation guidelines, tools & training

Preparation phase until December 2022

- Access to **some** clinical narratives
- Overview of the annotation process
- **Selection of the annotation tool**
- **First draft** of annotation guidelines

Pilot phase until May 2023

- Revise annotation guidelines
- Access to **use case oriented** clinical narratives
- Feedback from annotators
- First steps for pre-annotation tooling

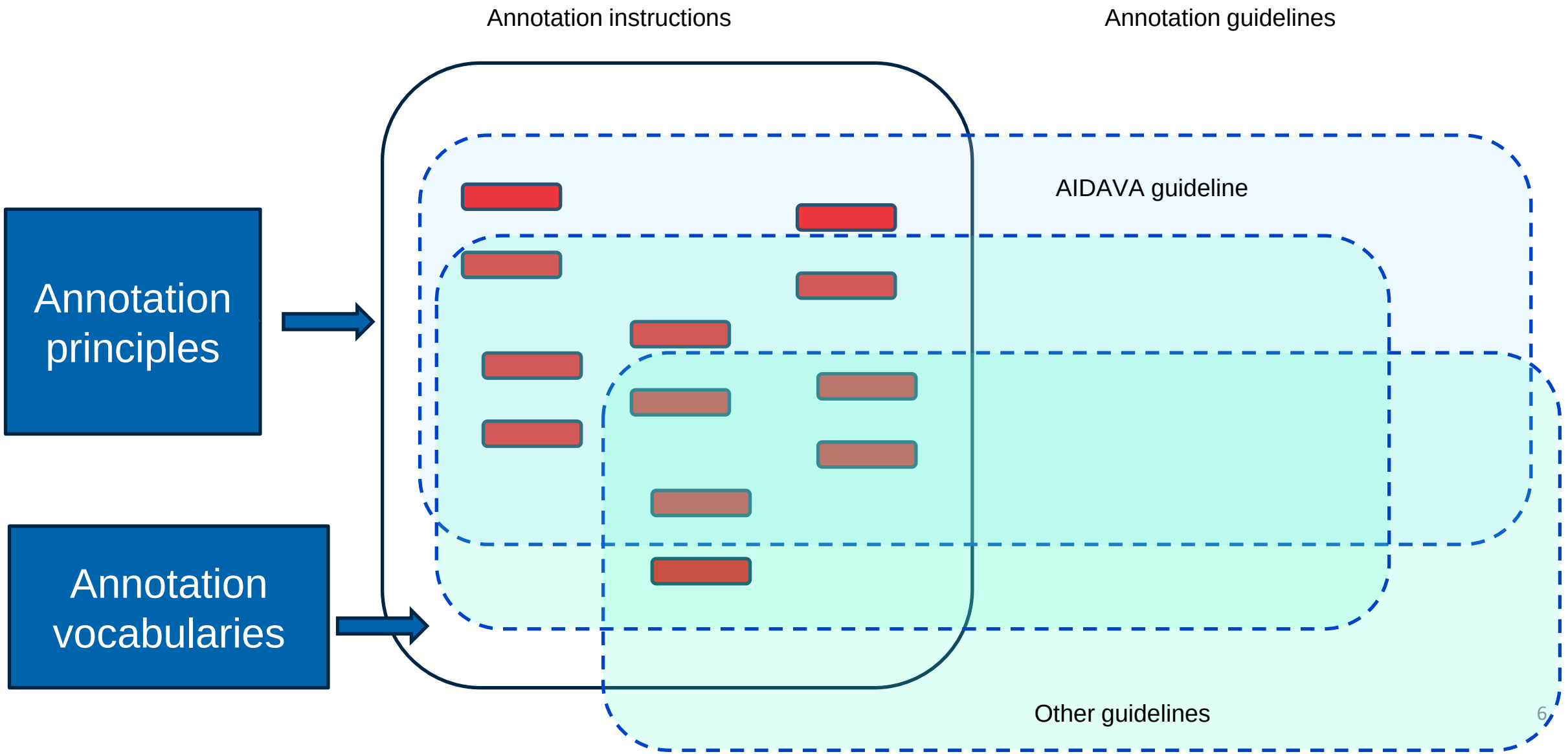

Hospital
AIDAVA Annotation team supported by tools
Source
Ethical Committee
Data Steward
AIDAVA temp data store in DH
Senior Annotator
Terminologist/ Clinical team
Junior Annotator
NLP modeler
Library Curation tool
Approval to access for TEXT ANNOTATION
EC approval
Data Source (Identifiable)
Data Source (Pseudonymised)
Extract Data in scope
Run Pseudonymization Algorithm (if needed)
START
Specify concepts in scope
Request access to data sources for ANNOTATION
Identify standards
Select Tool
Define translation approach
Define/Update annotation guidelines
To be deleted
Pseudonymized Extract - RAW
HEALTH discovery
Pre-annotation
Train & coach junior annotator
Annotate with adjudication
INCEpTION
To be deleted
Pseudonymized Extract - ANNOTATED
PERSISTENT
ANONYMIZED ANNOTATED DATA
Anonymize data for archival, delete sources
Select pre-existing model
Use data for training & testing model (ensure data privacy)
Data Curation Tools (NLP models)

From annotation principles to exemplification

- Semantic annotation
  - Following annotation principles
- Assignment of codes and relations to
  - Documents
  - Database entries

- Codes
  - SNOMED CT subsets
  - LOINC? – mapping at a later stage
- User friendly relation “predicates”
  - Harmonized with SNOMED and FHIR

Annotation: Principles, instructions, guidelines

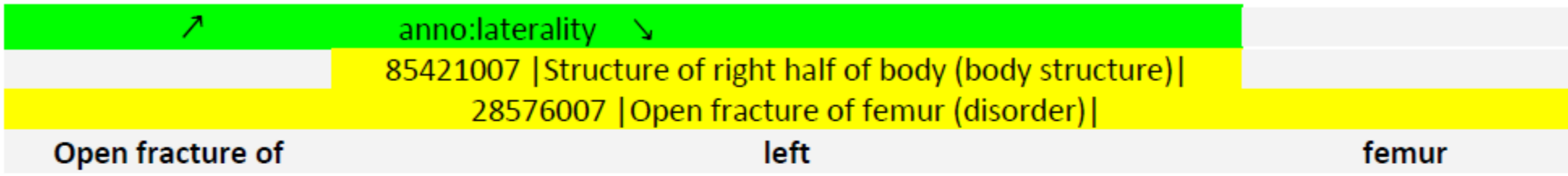


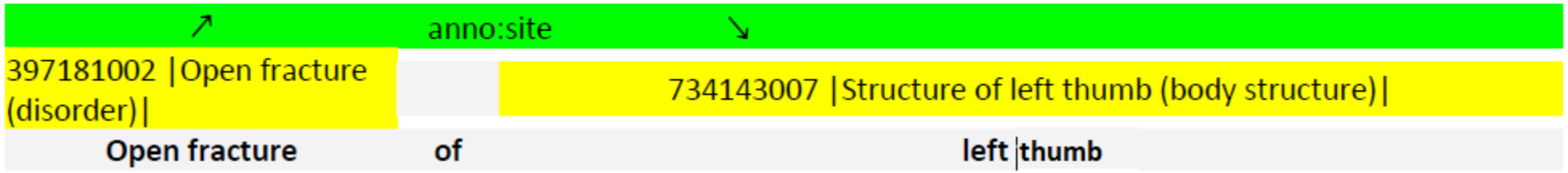
Guideline – Example 1

- Annotation principle:

  Annotate the span that corresponds to the meaning represented by a concept in the annotation vocabulary

  -> instruction: give preference to SNOMED CT precoordinated content





Guideline – Example 2

- Annotation principle:

  Annotate , what is in the text, not what you interpret

  -> instruction: don’t interpret temporal sequence as causality unless explicitly stated

↗	anno:after	↘
25064002 |Headache (finding)|		110030002 |Concussion injury of brain (disorder)|
Headache	**following**	**brain concussion**

↗	anno:dueTo	↘
162049009 |Left flank pain (finding)|		45816000 |Pyelonephritis (disorder)|
Left flank pain	**in**	**pyelonephritis**

Guideline – User-friendly predicates

- During annotation: choice of predicates dependent on domain and range constrained to be learned by annotators
- Automatic transformation to standards-conforming expressions after annotation

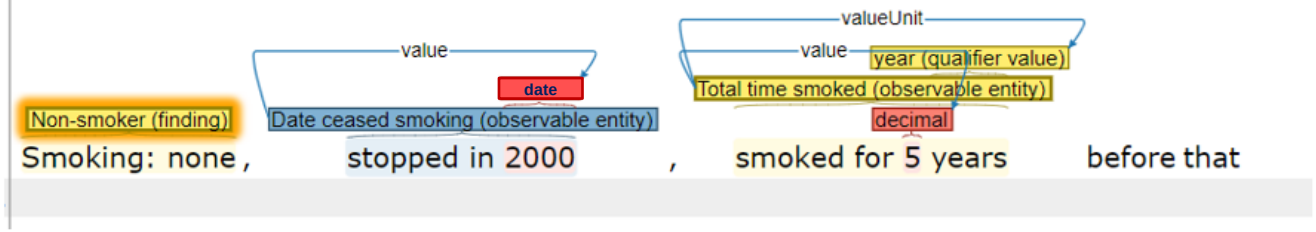
Example binary predicates for relation annotations with their translation into SNOMED CT and FHIR syntax. "INV" = inverse, "||" = concatenation. The relation paths marked with [a] refer to the concatenation of SNOMED CT relations, those marked with [b] to roughly equivalent FHIR elements

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'	Qualifier value' (cf. Tab. 1)

Task 4.3 Manual Annotations
Exemplification “tickets”

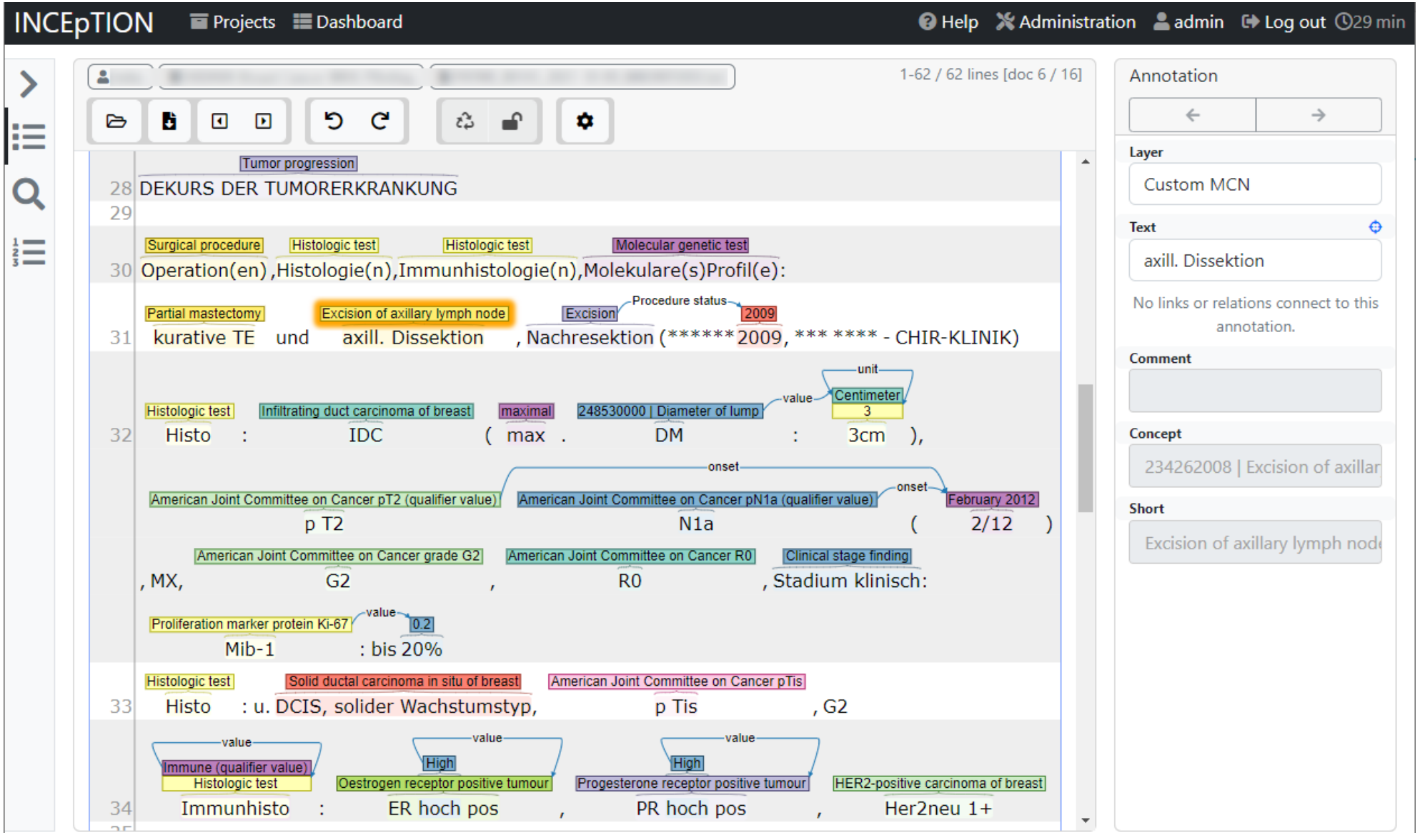
- Growing repository of annotation examples (“tickets”)

Table 10: Example of smoking behaviour

Input	Smoking: none, stopped in 2000, smoked for 5 years before that
INCEpTION screenshot	
Adjudication Description	Two predicted patterns (stop at a specific date, and time duration) in smoking data appear in the input text. Therefore, instructions X and Y are followed to annotate ‘stopped in 2000’ and ‘smoked for 5 years’ as ‘Date ceased smoking’ and ‘Total time smoked’, respectively. Also, the numbers 2000 and 5 were coded as decimals. Moreover, the time-unit needs to be identified (as instructed in Section 6.2), so ‘years’ correspond to ‘year (qualifier value)’. As shown in Table 5, the predicate ‘value’ has observable entity and decimal as its domain and range, respectively. Thus, we use the predicate ‘value’ between the identified observable entities and their corresponding decimals. The same explanation applies to ‘valueUnit’ between the observable value and the qualifier value (i.e., year).

Task 4.3 Manual Annotations

From exemplification tickets to full document annotations



Task 4.3 Manual Annotations
Scope and granularity

Determined by **reference ontology**

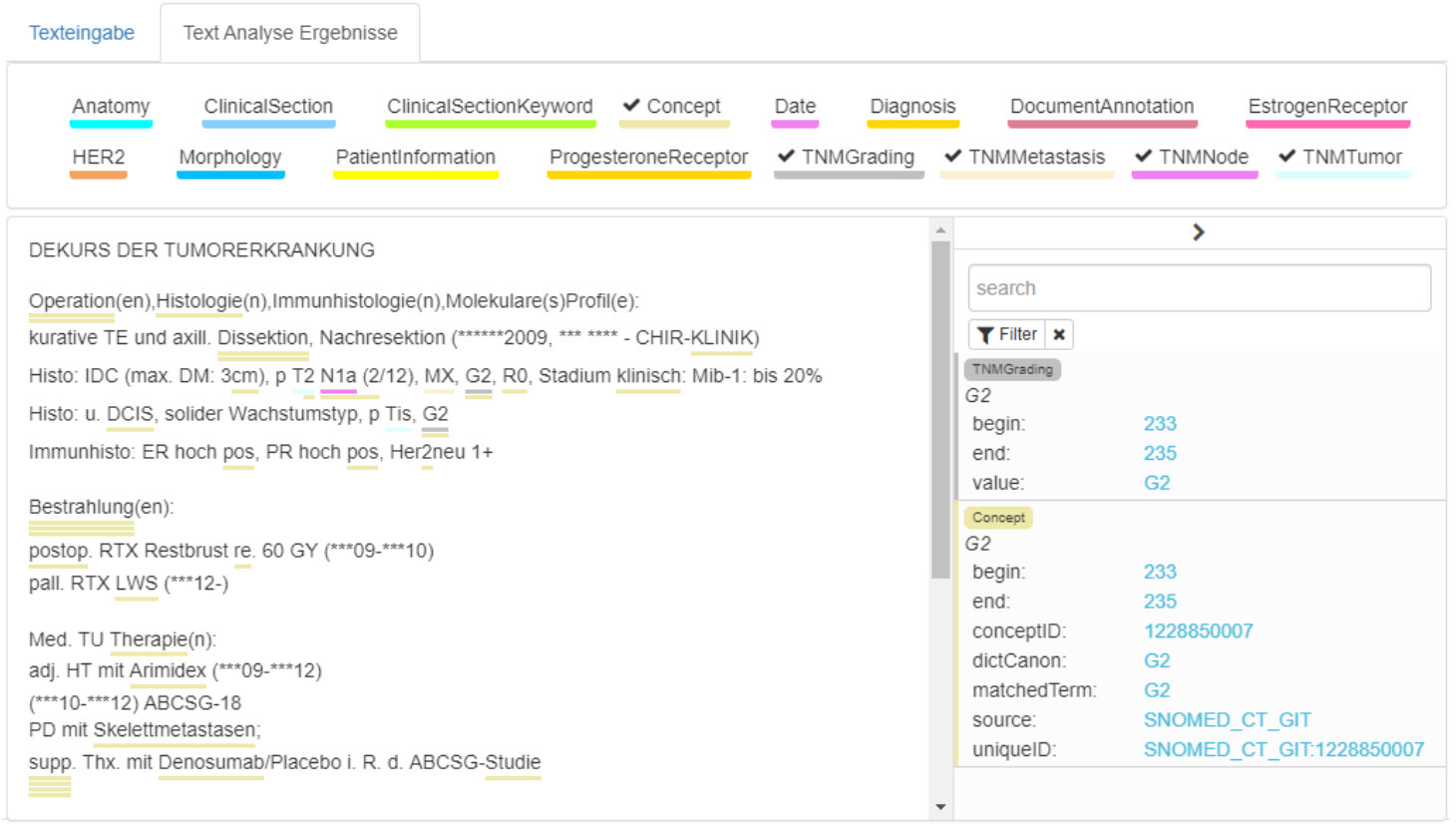
- Subsets of SNOMED CT with the maximum coverage for the prioritized items of the use cases

When and how to access the reference ontology?

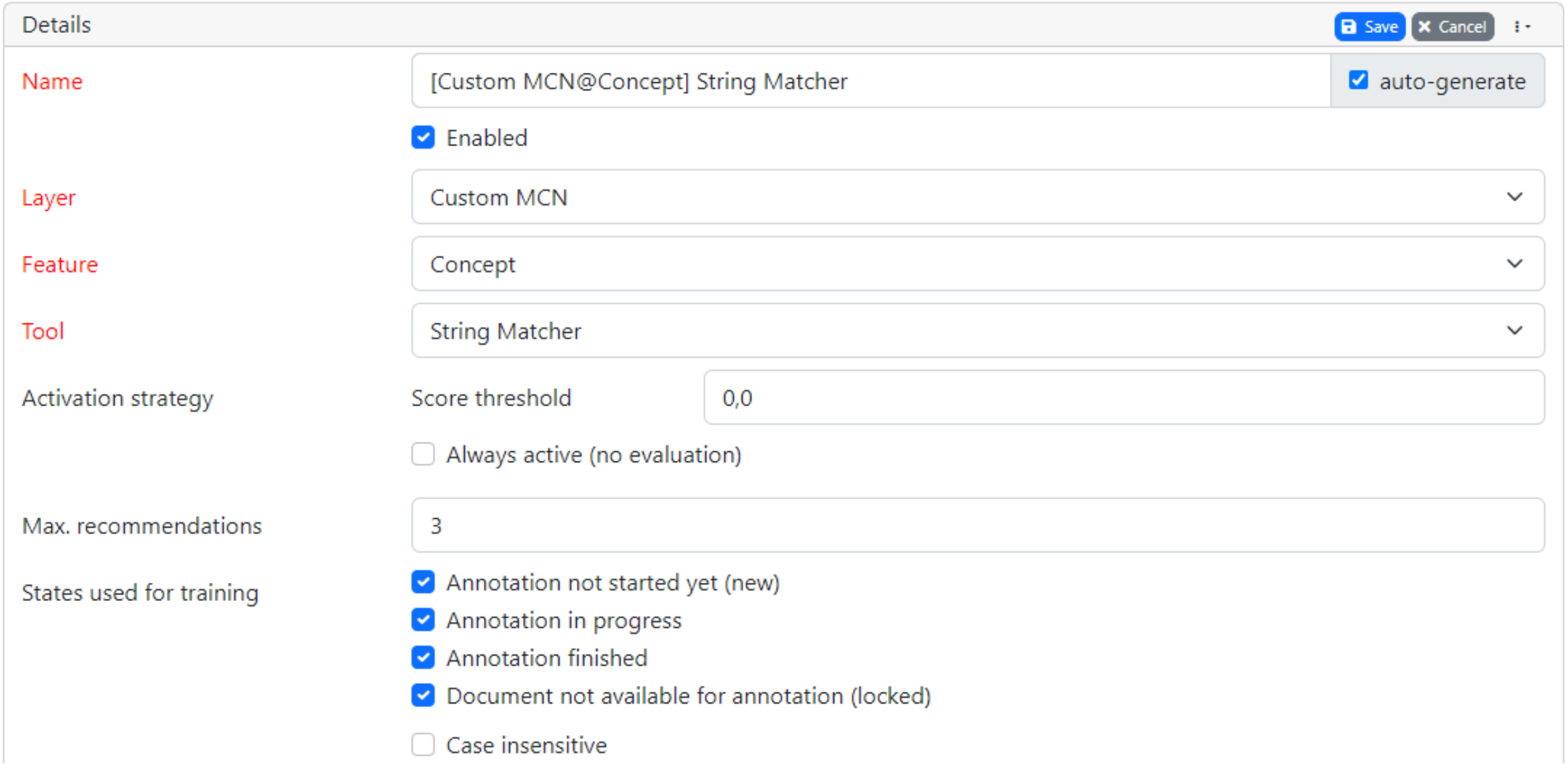
- Annotators use just the SNOMED CT browser, abstraction to reference ontology content done automatically?
- Annotators have only access to reference ontology?
- Reference ontology imported into annotation tool?

Task 4.3 Manual Annotations
Pre-annotation

Using existing NLP pipeline: **Averbis Health Discovery**

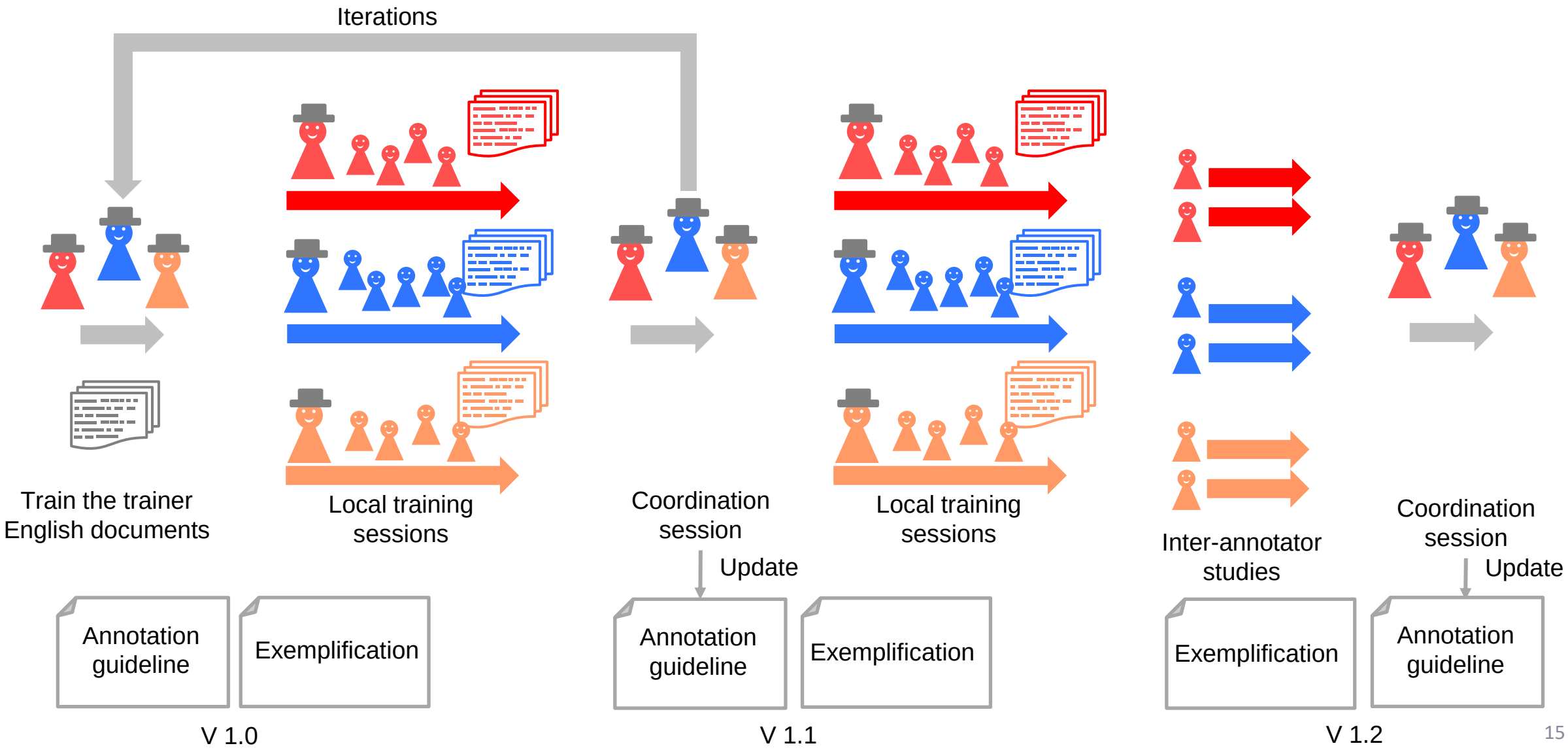


INCEpTION recommender functionality

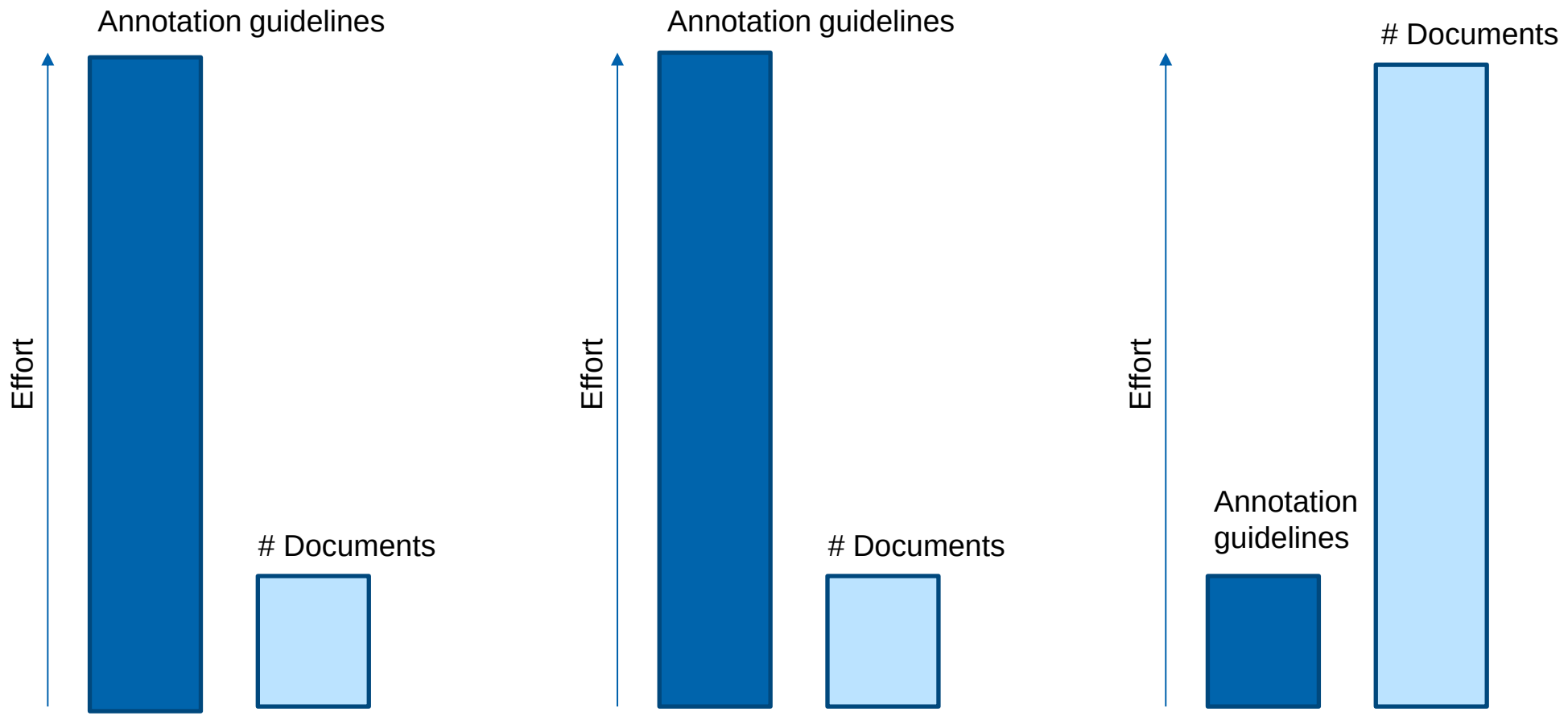


Task 4.3 Manual Annotations

Update process in the production phase



Annotation work load



Preparation Phase
End 2022

Piloting phase
End April 2023

Productive phase
Start June 2023

Vision

- **Canonical form** of semantic representation of clinical narrative
- **Maximally consistent** inter-annotator agreement (IAA)
- **Annotation graph** as the primary knowledge graph (KG)

Synergies

Principles of ontology-based annotation of clinical narratives*

Stefan Schulz[1,2,*], Warren Del-Pinto[3], Lifeng Han[3], Markus Kreuzthaler[1], Sareh Aghaei Dinani[1] and 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*

Task 4.3 Manual Annotations

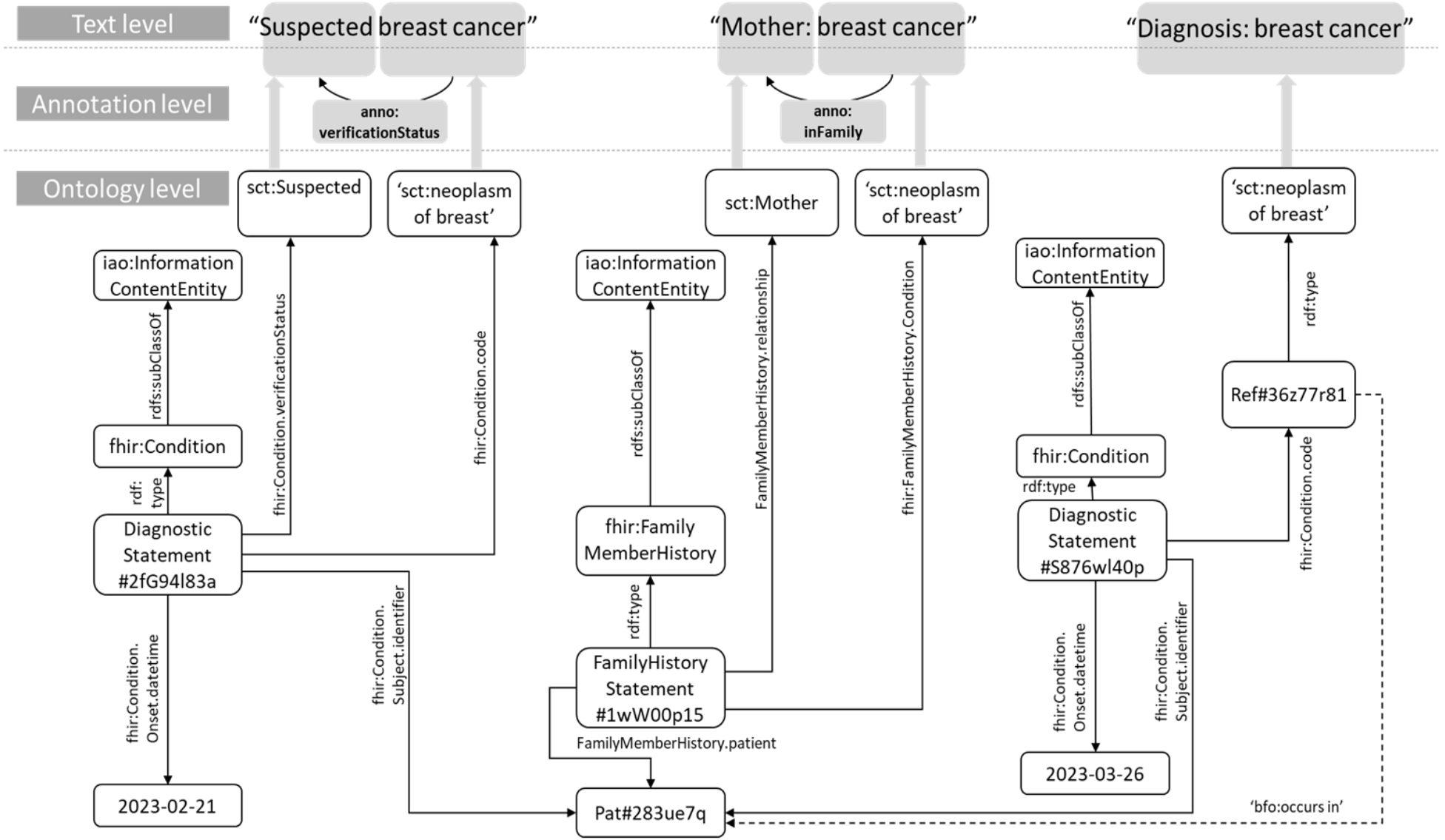
Scientific output - Synergies

- Technical University of Munich (Germany)
  - Research Consortium DIFUTURE
  - Methods Platform GeMTex
- University of Erlangen - Nürnberg (Germany)
- German Research Centre for Artificial Intelligence
- University of Manchester (UK)
- University of Murcia (Spain)
- University of Chiang Mai (Thailand)

Alignment Annotation – Reference ontology

- Discussion

From annotation to ontology-based knowledge graph



Example coreference

t_1

126926005 |Neoplasm of breast (disorder)|

rdf:type

NoB34u73axn4us

owl:sameAs

t_2

254837009 |Malignant neoplasm of breast (disorder)|

rdf:type

MoBkj88935elp

owl:sameAs

t_3

278054005 |Lobular carcinoma of breast (disorder)|

rdf:type

LCBrrp009g65t

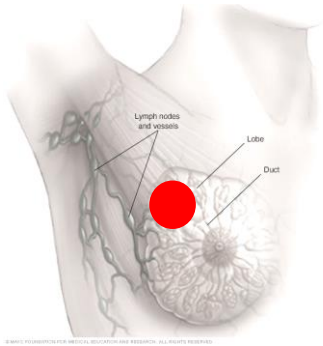
owl:sameAs

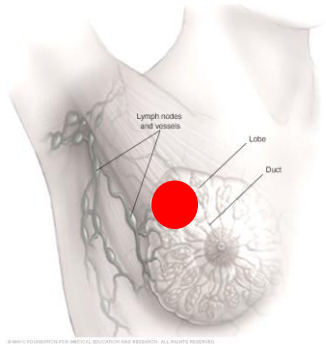
t_4

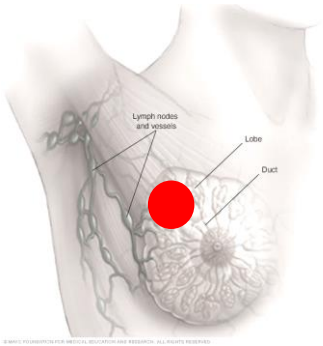
713609000 |Invasive carcinoma of breast (disorder)| AND 278054005 |Lobular carcinoma of breast (disorder)|

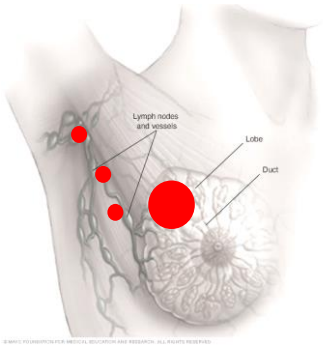
rdf:type

ILCB4tz5pplkll









Lobular carcinoma of breast

Invasive Lobular carcinoma of breast

Thank you