Semantics@Roche – September 27th, 2017

Stefan Schulz
Medical University of Graz (Austria)

Reference terminologies vs. Interface terminologies

purl.org/steschu

Common problems with specialised domain terminologies

- I need to automatically annotate textual content
  - that uses an idiosyncratic sublanguage
  - which abounds of abbreviations and errors (syntax, spelling)
  - which uses highly ambiguous terms
  - the meaning of which depends on local contexts
  - is characterised by constantly new terms
    I need to find close-to-user terms for structured entry
- Current situation:
  - Numerous quasi-standard terminologies
  - None of them cover all my concepts
  - Many terms I use are not covered
    (although concepts are available)

Popularity of Terms in Pubmed

Preferred term (SNOMED CT)	Count	Synonym	Count
Primary malignant neoplasm of lung	0	Lung cancer Bronchial carcinoma	120682 3452
Cerebrovascular accident	3819	Stroke	191559
Block dissection of cervical lymph nodes	1	Neck dissection	7512
Electrocardiographic procedure	1	Electrocardiogram ECG	33670 55120
Backache	3489	Back pain	38132
Capillary blood specimens	32	Capillary blood samples	574

Two Aspects of Terminologies

- Normative
  - Codes + Labels (names) denote well-defined entities in a realm of discourse
  - "Explanatory“ labels, e.g. "*Primary malignant neoplasm of lung (disorder)*".
  - Scope notes and definitions explain meaning of concepts
  - Meaning formalised by logic descriptions → formal ontology
- Descriptive
  - Describes language as being used: lexicon "*Lung cancer*"

Reference terms – Interface terms

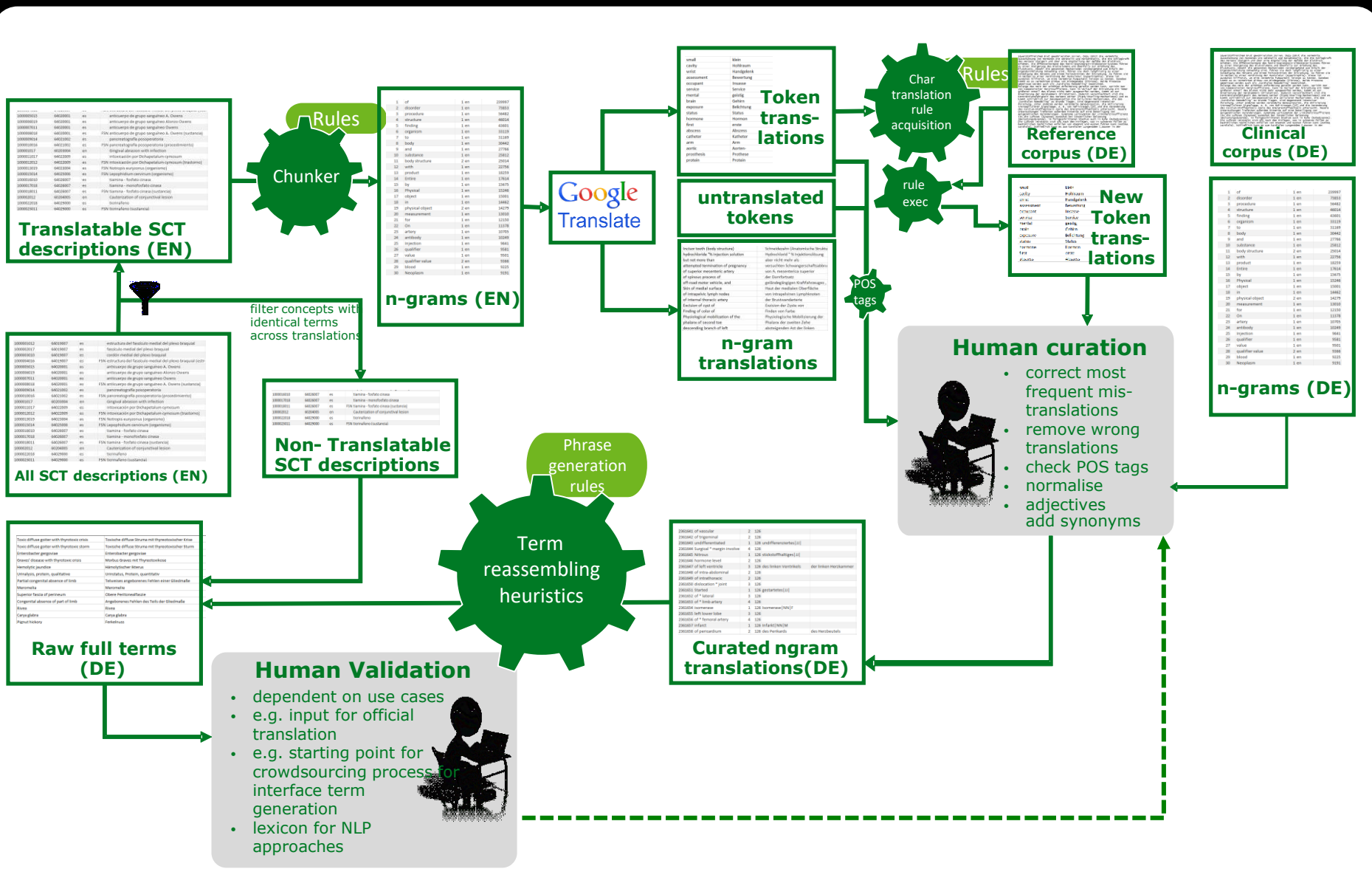
- Reference terms: Normative
  - Primarily language-independent: described by textual and / or formal (ontological) definitions
  - Labels univocally identify a concept, often not terms of choice in communication / documentation
- Interface terms: Descriptive
  - Collection of human language terms / expressions used in (informal) written and oral communication
  - Grounding by connection with reference terminologies
  - Inherent ambiguity of interface terms, often not perceived
  - Need to deal with short forms
- Many domain terminologies contain interface terms, but far from being exhaustive

Case study: German interface terminology for SNOMED CT

- Context: Information extraction from clinical narratives → CBmed IICCAB* (see poster)
- Limited resources, incremental approach
- Top-down:
  - Modularization of original terms (split into noun phrases):
  - Translating / finding synonyms of derived, highly repetitive phrases:
    - "Magnetic resonance imaging" in 627 SNOMED terms
    - "Second degree burn" in 166 SNOMED terms
  - Acquisition of translations and synonyms by decreasing frequency (machine translation, automated synonym acquisition)
  - Manual revision
- Bottom-up:
  - Addition of terms from n-gram frequency lists from reference corpora

*Schulz S. Innovative Nutzung von Informationen für klinische Versorgung und Biomarkerforschung. http://goo.gl/wHMedz

MUG-GIT: Erstellung einer deutschen Interface-terminologie für SNOMED CT (II)



Case study: German interface terminology for SNOMED CT

- Core vocabulary:
  - Constantly checked and enhanced by domain experts (medical students)
  - Priorisation by Use cases
- Guidelines
  - Avoidance of ambiguous entries: inclusion of composed terms (e.g. "delivery" → "drug delivery", "preterm delivery")
  - Acronyms only in context: not "CT" but "CT guided"
  - Inflection, compositions rules requires special markers (German) and rules for reassembly of terms
- Current state:
  - ca. 2 Million Interface-Terms
  - Automatically generated from core vocabulary with 92,500 n-grams, out of 85,400 English n-grams
  - Benchmark: parallel corpus extracted from Medline: term coverage 33.1% for German vs. 55.4% for English

Core n-gram vocabulary

vaginal	1	1478	vaginales|JJ	Scheiden-	
fluoroscopic guidance	2	1477	Durchleuchtungskontrolle|NN|F		
disc	1	1476	Scheibe|NN|F		
lower limb	2	1473	unteres|JJ Extremität|NN|F	Bein|NN|N	
brain	1	1468	Gehirn|NN|N	Hirn|NN|N	Encephalon|NN|N
preparation	1	1464	Zubereitung|NN|F	Aufbereitung|NN|F	Präparation|NN|F
method	1	1463	Verfahren|NN|N	Methode|NN|F	
of bone	2	1462	des Knochens	_Knochen_	
Red	1	1455	rotes|JJ		
Monitoring	1	1453	Überwachung|NN|F	Monitoring|NN|N	
Computed	1	1453	berechnetes|JJ	Computer-	
phalanx	1	1449	Phalanx|NN|F		
subsp.	1	1449			
anastomosis	1	1447	Anastomose|NN|F	Anastomosierung|NN|F	
vessel	1	1446	Blutgefäß|NN|N	Gefäß|NN|N	
Computed tomography	2	1443	Computertomographie|NN|F		
uterus	1	1436	Uterus|NN|M	Gebärmutter|NN|F	
difficulty	1	1432	Schwierigkeit|NN|F		
elbow	1	1429	Ellbogen|NN|M	Cubitus|NN|M	Ellbogengelenk|NN|N
high	1	1429	hohes|JJ		
food	1	1423	Lebensmittel|NN|N	Speise|NN|F	Nahrungsmittel|NN|N
Observation	1	1423	Beobachtung|NN|F		
using fluoroscopic	2	1422			
unable	1	1421	unfähiges|JJ		
Peripheral	1	1419	peripheres|JJ		
unable to	2	1418	unfähig zu		
Vascular	1	1417	vaskuläres|JJ	Gefäß-	
using fluoroscopic guidance	3	1416	mit Durchleuchtungskontrolle		
Benign neoplasm	2	1415	gutartiges|JJ Neubildung|NN|F	gutartiges|JJ Neoplasie|NN|F	benignes|JJ Neoplasie|NN|F

Automatically generated interface terminology

20170315_240011_002	126952004	Neoplasm of brain	Gehirnneubildung
20170315_240011_003	126952004	Neoplasm of brain	Neubildung des Hirns
20170315_240011_004	126952004	Neoplasm of brain	Hirnneubildung
20170315_240011_005	126952004	Neoplasm of brain	Neoplasie des Gehirns
20170315_240011_006	126952004	Neoplasm of brain	Gehirnneoplasie
20170315_240011_007	126952004	Neoplasm of brain	Neoplasie des Hirns
20170315_240011_008	126952004	Neoplasm of brain	Hirnneoplasie
20170315_240011_009	126952004	Neoplasm of brain	Neoplasma des Gehirns
20170315_240011_010	126952004	Neoplasm of brain	Gehirnneoplasma
20170315_240011_011	126952004	Neoplasm of brain	Neoplasma des Hirns
20170315_240011_012	126952004	Neoplasm of brain	Hirnneoplasma
20170315_241010_001	126953009	Neoplasm of cerebrum	Neubildung des Großhirns
20170315_241010_002	126953009	Neoplasm of cerebrum	Neoplasie des Großhirns
20170315_241010_003	126953009	Neoplasm of cerebrum	Neoplasma des Großhirns
20170315_242015_001	126954003	Neoplasm of frontal lobe	Neubildung des Frontallappens
20170315_242015_002	126954003	Neoplasm of frontal lobe	Neubildung des Lobus frontalis
20170315_242015_003	126954003	Neoplasm of frontal lobe	Neoplasie des Frontallappens
20170315_242015_004	126954003	Neoplasm of frontal lobe	Neoplasie des Lobus frontalis
20170315_242015_005	126954003	Neoplasm of frontal lobe	Neoplasma des Frontallappens
20170315_242015_006	126954003	Neoplasm of frontal lobe	Neoplasma des Lobus frontalis
20170315_243013_001	126955002	Neoplasm of temporal lobe	Neubildung des Temporallappens
20170315_243013_002	126955002	Neoplasm of temporal lobe	Neubildung des Lobus temporalis
20170315_243013_003	126955002	Neoplasm of temporal lobe	Neoplasie des Temporallappens
20170315_243013_004	126955002	Neoplasm of temporal lobe	Neoplasie des Lobus temporalis
20170315_243013_005	126955002	Neoplasm of temporal lobe	Neoplasma des Temporallappens

Crowdsourcing for terminology development

- Functionality: entry of new terms, commenting and validating existing terms
- Possible central data element :
  Interface Term – External Code
  "DM" - *81827009 |Diameter (qualifier value)*
- Possible Attributes:
  - Creator, creation type, date, (sub)domain, user group
    Max Muster, manual, 20170803, Dermatology Graz, Doctos
  - Example annotation, e.g.
    "ein 3 cm im DM haltender Tumor"
  - Validation/ commenting by other users
    John Doe, 20180912, ★★★★☆
    "Example incomprehensible – additional examples needed"

Short forms – Document preprocessing

- Ambiguous short forms not in dictionary
- Difficulty of maintenance
  - Abundance of concurring readings
  - High productivity
- Instead:
  - Main assumption: short forms and expansions occur in the same corpus
  - Automatically create N-gram model from specific reference corpus
  - Replace short forms by most plausible expansions
  - The same for other out of-lexicon words, e.g. misspellings
  - If assumption fails: try Web mining

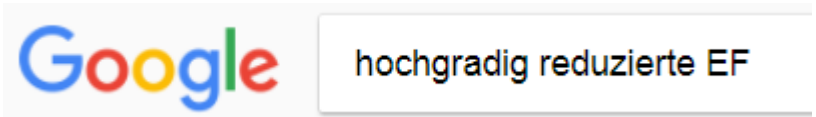
Example: Resolution of short forms

- "dilat. Kardiomyopathie, hochgr. red. EF"
- Word-n-gram model (30,000 discharge summaries)

```
1035      dilat. Kardiomyopathie
1442      dilatative Kardiomyopathie
```

```
   7      hochgr. red. EF
   4      hochgradig reduzierte EF
```

- Web mining



Ejektionsfraktion – Wikipedia
https://de.wikipedia.org/wiki/Ejektionsfraktion ▾ Translate this page
Die Ejektionsfraktion (**EF**) oder Auswurffraktion (auch Austreibungsfraktion) ist ein Maß für die ... 30 %, **hochgradig** eingeschränkt ... Eine **reduzierte** Ejektionsfraktion wird als objektivierbarer Parameter

Example: Resolution of short forms

- "Pat. mit rez. HWI und VUR"
- Wort-N-gram Modell

```
381  Pat. mit
120  Patient mit
  2  rez.
707  rezenten
468  rezidivierende
```

- No clear expansions of "HWI" and "VUR"
- Web mining 

Example: Resolution of short forms

- Interpretation:
  - Frequency
  - Acronym-definition patterns
  - Regular expressions created from short forms

Der vesikoureterale Reflux (VUR) - KidsDoc.at
www.kidsdoc.at/vesikoureteraler_reflux_vur.html ▾ Translate this page
Der vesikoureterale Reflux (VUR) ist ein Zurückfließen des Urins aus der Harnblase in den oberen Harntrakt (Harnleiter **und** Nierenbecken). Der Harn soll von ...

[PDF] HWI Ursachen und Folgen Leoben 2012
www.paediatrie.at/.../HWI_Ursachen_Folgen_Leoben_2012.pdf ▾ Translate this page
Nov 30, 2012 - Harnwegsinfekt (**HWI**) **und**. Harnwegsfehlbildungen ... Erbrechen **und** Durchfall seit heute ... Vesico--ureteraler Reflux (**VUR**). • Obstrukyve ...

[PDF] Diagnose und Behandlung von ... - Swiss Paediatrics
www.swiss-paediatrics.org/sites/default/files/10-13_0.pdf ▾ Translate this page
tik, Behandlung, Abklärung **und** Nachkontrol- ... nen, Säuglingen, Kindern **und** Jugendlichen bis. 16 Jahre. ..... wegsinfektion noch das Vorliegen eines **VUR**.

[PDF] Rezidivierender Harnwegsinfekt beim Kind - Krause und Pachernegg
www.kup.at/kup/pdf/7439.pdf ▾ Translate this page
by J Oswald
bestätigten **HWI** um das 6. Lebensjahr liegt bei 7 % für Mädchen **und** 2 % bei. Knaben. Die meisten ... der mit rezidivierenden HWIs ein **VUR** dokumentieren.

[PDF] Harnwegsinfektionen im frühen Kindesalter - CME-Portal - MGO ...
https://cme.mgo-fachverlage.de/uploads/exam/exam_94.pdf ▾ Translate this page
der ersten **HWI** deutlich jünger als die betrof- fenen Mädchen. Bei 36 % der Mädchen **und** bei. 24 % der Jungen fand sich ein vesikoureteraler. Reflux (**VUR**) ...

[PDF] Harnwegsinfektion - Klinik und Poliklinik für Kinder- und ...
kik.uniklinikum-leipzig.de/red_tools/dl_document.php?id=251 ▾ Translate this page
monas **und** Enterokokken je 2%. CAVE bei Jungs 30% Proteus ab 2. Lj. Prädisponierende Faktoren für Auftreten einer **HWI**: Vesikoureteraler Reflux (**VUR**): 30% ...

Recommendations

- Need for interface terms not satisfied by domain terminologies
- Acquisition of interface terms
  - Top-down (from reference terminologies)
  - Bottom up (from corpora)
- Use collaborative approach (crowdsourcing)
- Avoid ambiguous terms in lexicon
- Use n-gram models (alternatively deep learning ?) for out-of-lexicon terms, extracted from related corpora
  - Benefitting from typical local contexts in "similar" documents
    - Resolution of short forms
    - Correction of typos

Semantics@Roche – September 27th, 2017

Stefan Schulz

Medical University of Graz (Austria)

purl.org/steschu

Questions?

Contact:
stefan.schulz@medunigraz.at

Additional Slides

Example SNOMED CT

http://browser.ihtsdotools.org

ASSESS CT

H2020: Assessing SNOMED CT for Large Scale eHealth Deployments in the EU

"Is SNOMED CT well suited as an European reference terminology ?"

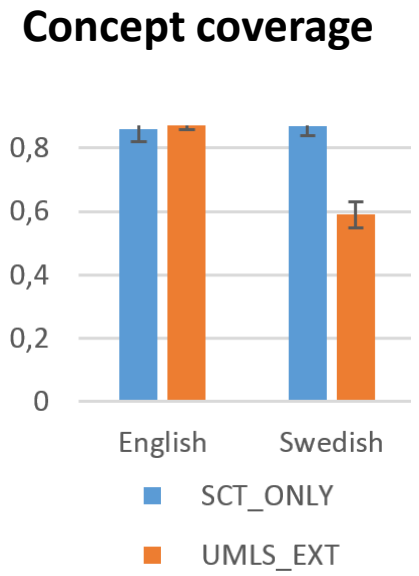
- Manual annotation of a corpus of clinical texts with SNOMED CT vs. UMLS-Extract
- Assessment:
  - concept coverage
  - term coverage

ASSESS CT

H2020: Assessing SNOMED CT for Large Scale eHealth Deployments in the EU

"Is SNOMED CT well suited as an European reference terminology ?"

- Manual annotation of a corpus of clinical texts with SNOMED CT vs. UMLS-Extract
- Assessment:
  - concept coverage
  - term coverage



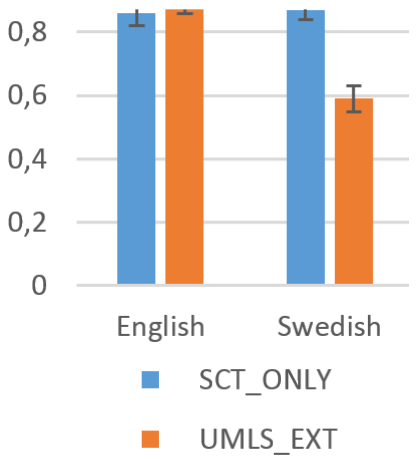
ASSESS CT

H2020: Assessing SNOMED CT for Large Scale eHealth Deployments in the EU

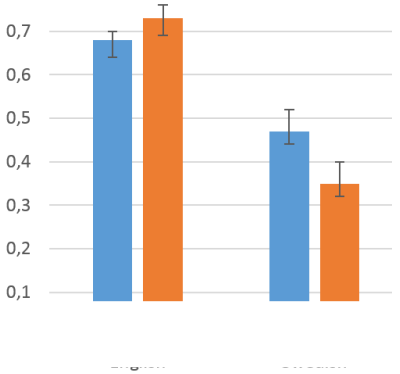
"Is SNOMED CT well suited as an European reference terminology ?"

- Manual annotation of a corpus of clinical texts with SNOMED CT vs. UMLS-Extract
- Assessment:
  - concept coverage
  - term coverage
- Differences SNOMED CT Swedish – English
  - Swedish:
    one term per concept
  - English:
    on average 2.3 terms
    per concept (Preferred terms , synonyms)

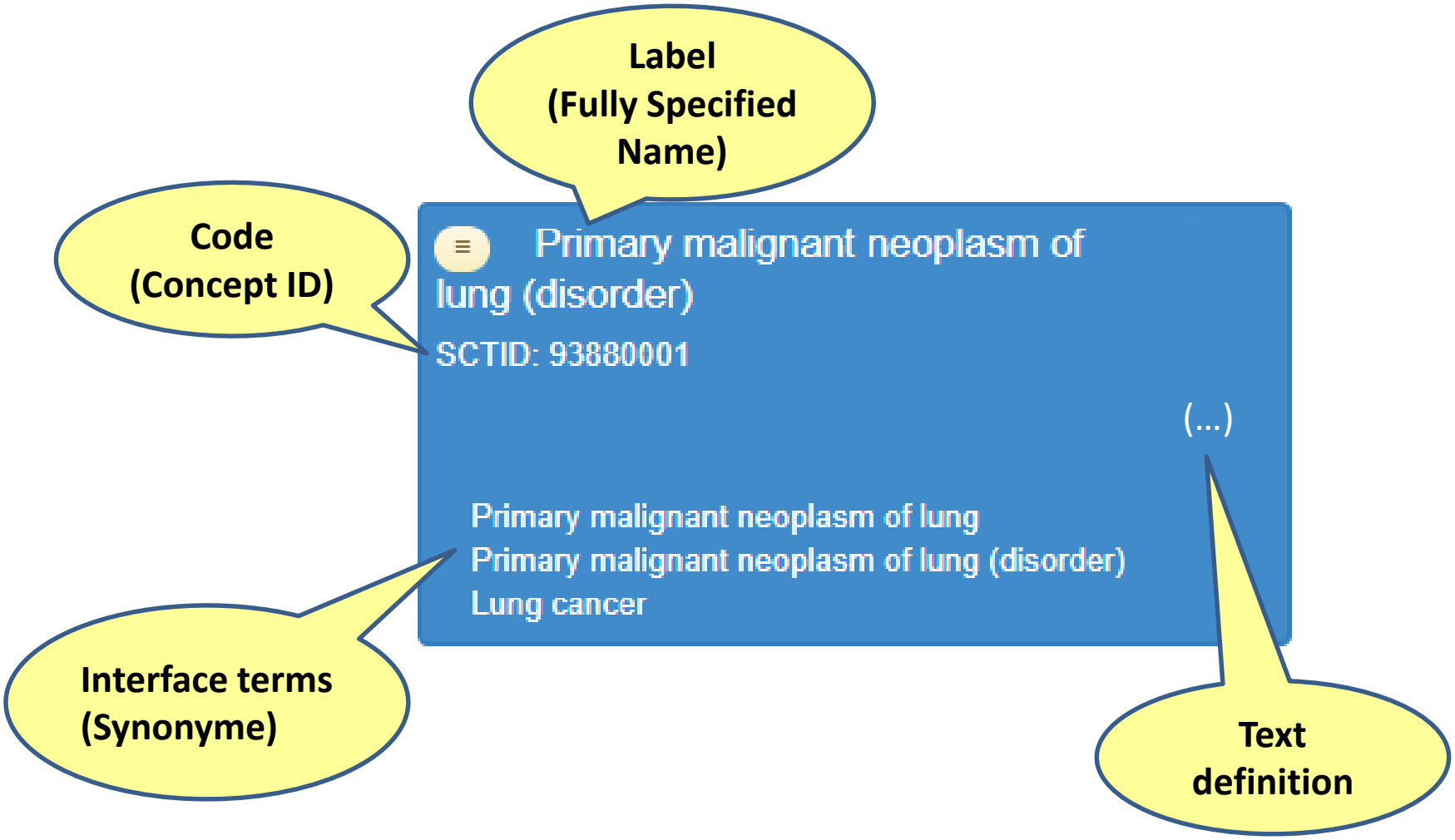
Concept coverage



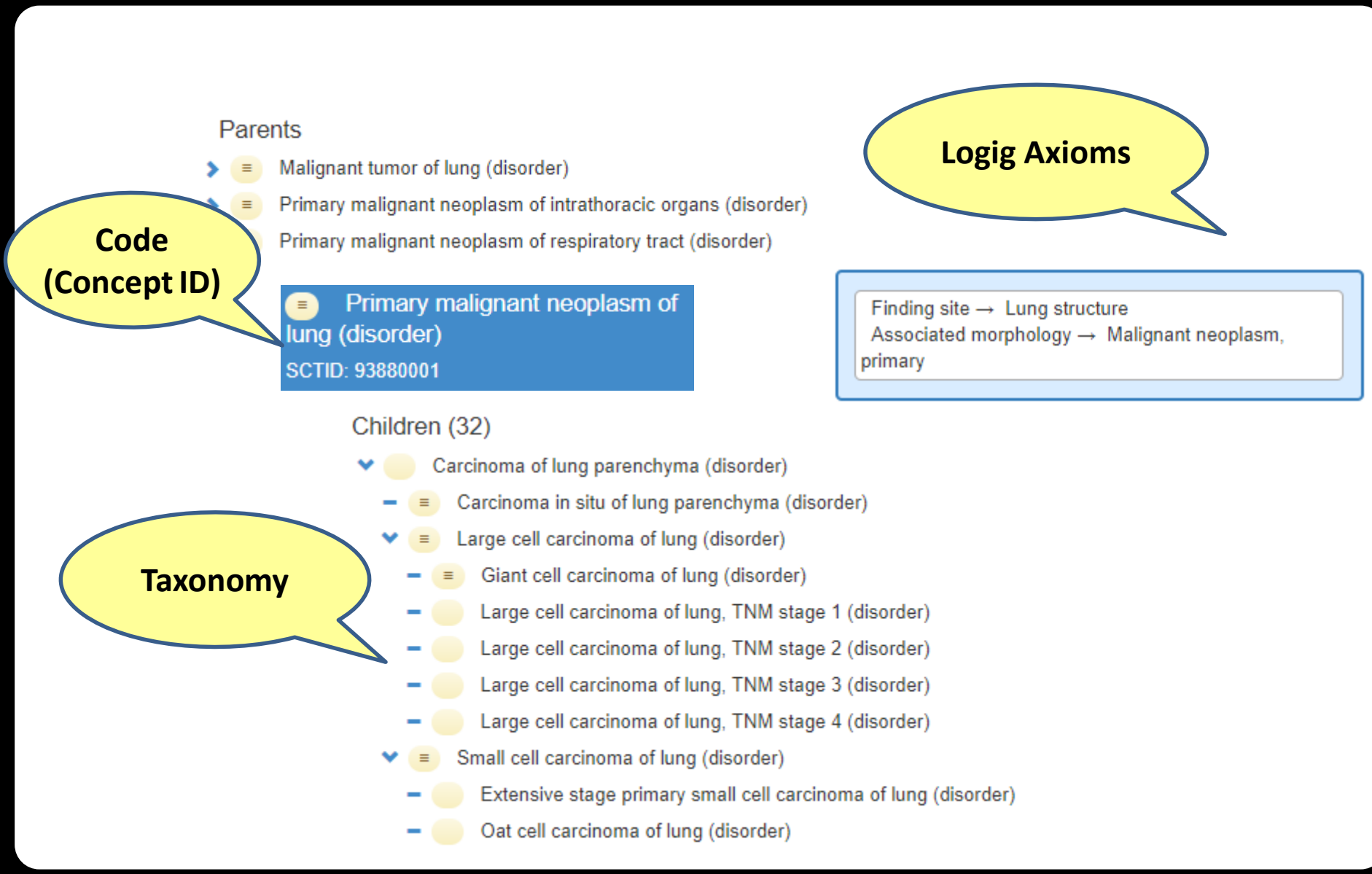
Term coverage



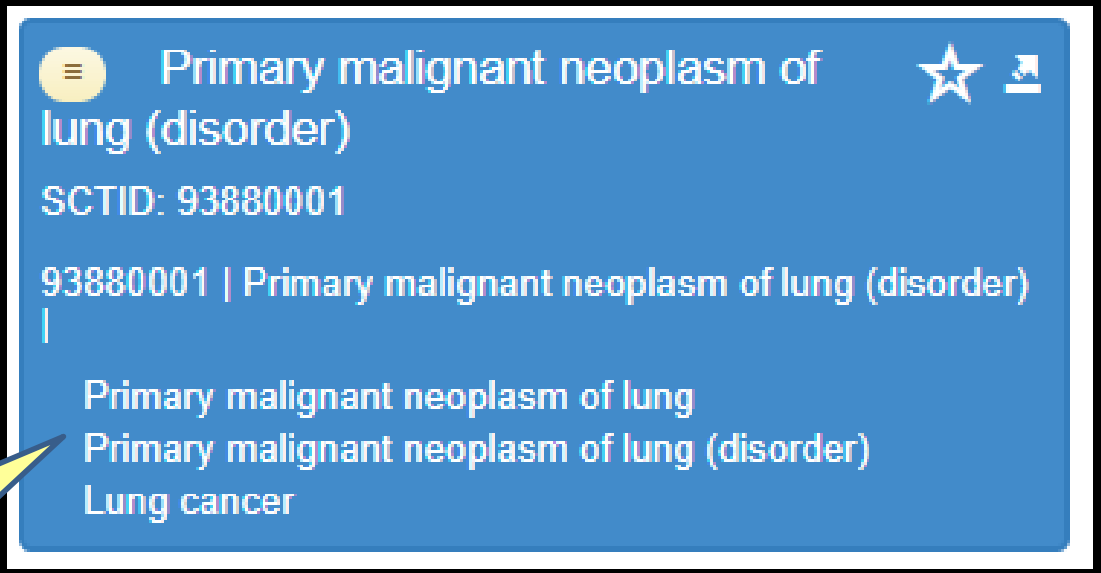
SNOMED CT als Terminology



SNOMED CT als formal Ontologie



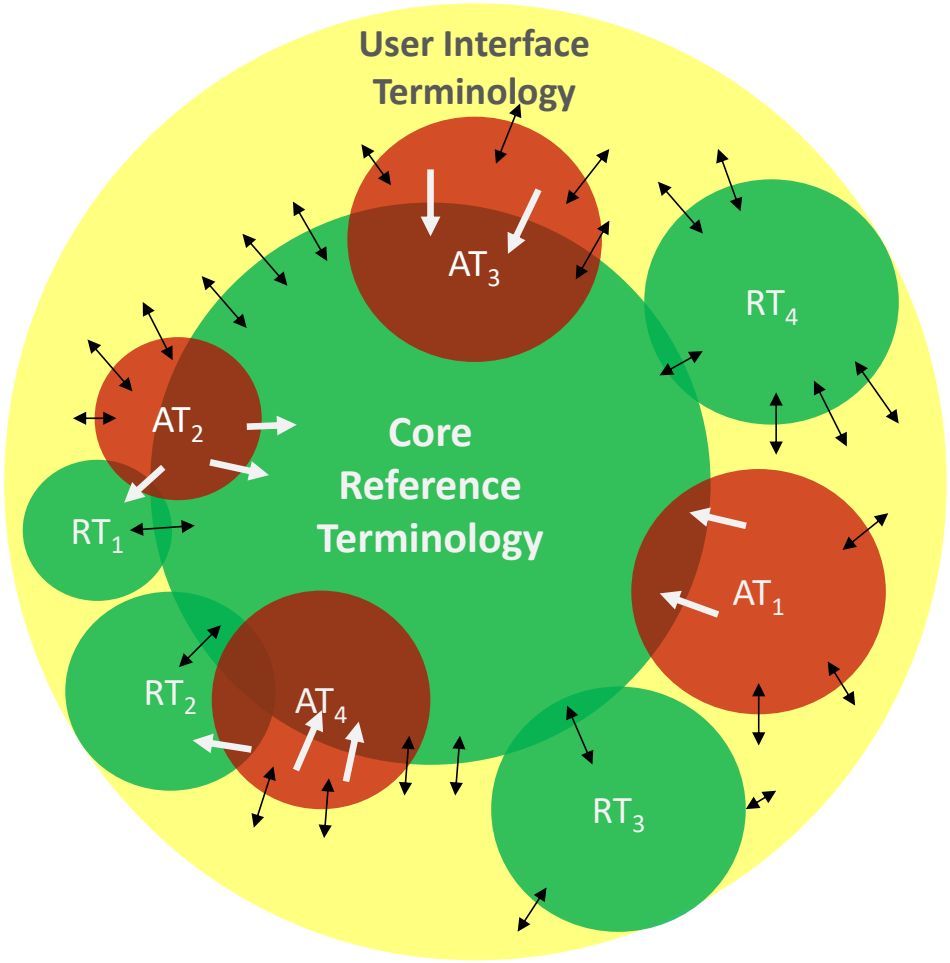
Interface-Terms → Interface-Terminology



http://browser.ihtsdotools.org

ASSESS CT

H2020: Assessing SNOMED CT for Large Scale eHealth Deployments in the EU



AT: aggregation terminology, RT: reference terminology

"SNOMED CT should be part of an ecosystem of terminologies, including international aggregation terminologies (e.g., the WHO Family of Classifications), and user interface terminologies, which address multilingualism in Europe and clinical communication with multidisciplinary professional language and lay language"

http://assess-ct.eu/fileadmin/assess_ct/final_brochure/assessct_final_brochure.pdf