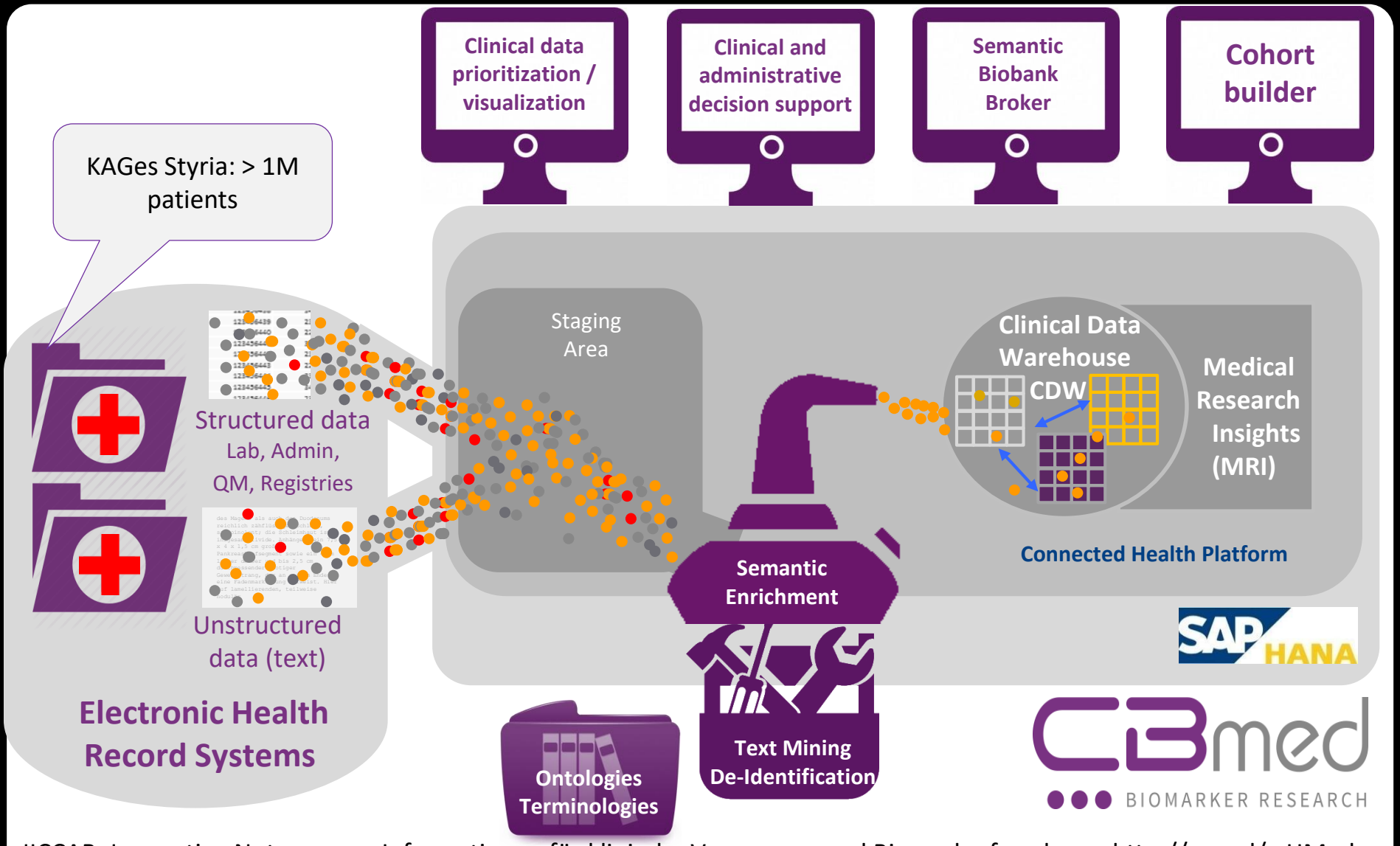
Stefan Schulz
Medical University of Graz (Austria)

http://purl.org/steschu

SNOMED CT Expo 2017
Bratislava | October 19-20

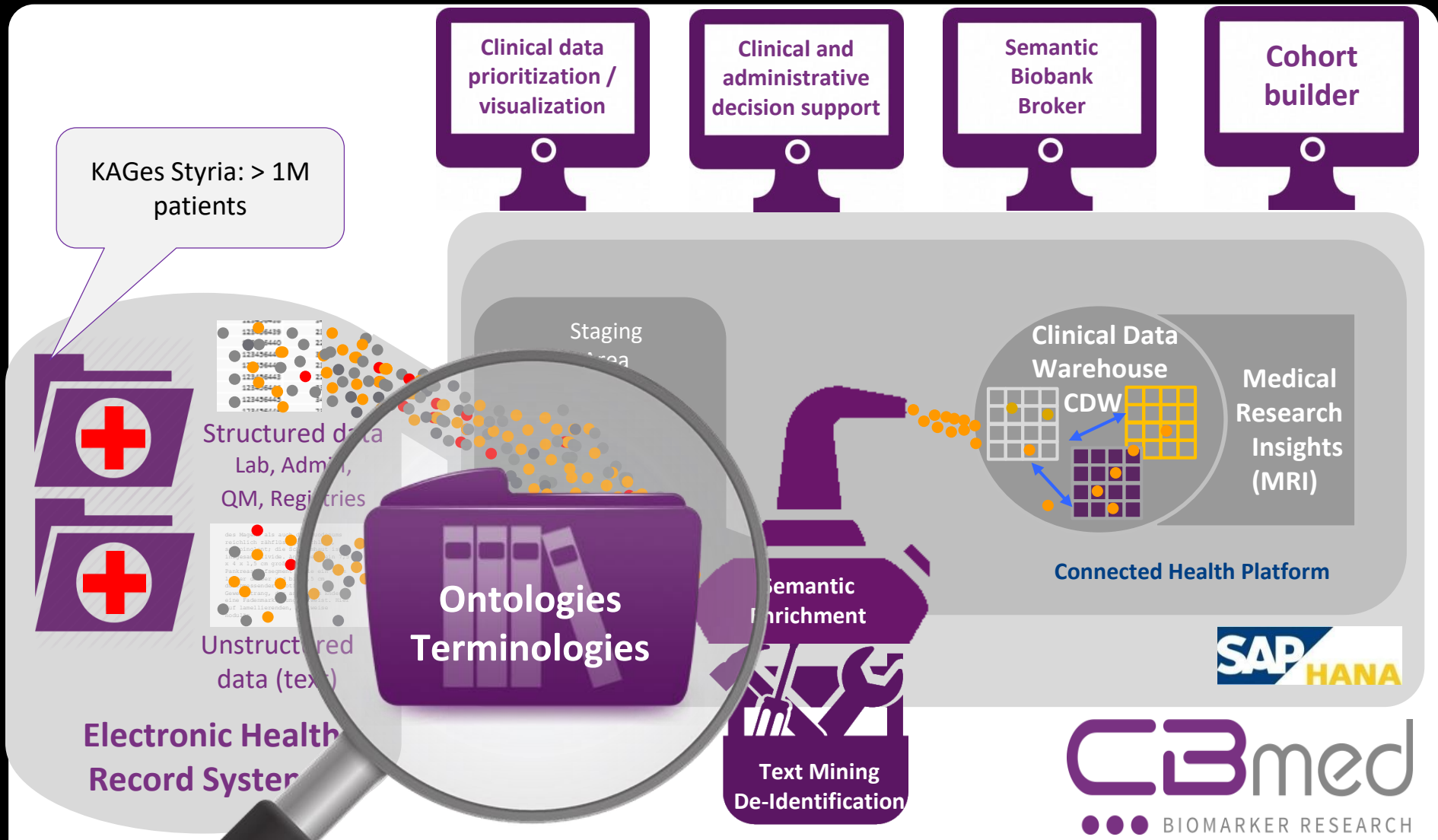
Building an experimental German user interface terminology linked to SNOMED CT

Context: Using SNOMED CT as annotation vocabulary for clinical narratives



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

Context: Using SNOMED CT as annotation vocabulary for clinical narratives



IICCAB: Innov… …zung von Informationen für klinische Versorgung und Biomarkerforschung. http://goo.gl/wHMedz

Using domain terminologies for NLP

- I need to automatically annotate textual content
  - which is in an idiosyncratic sublanguage
  - which abounds of abbreviations and errors
  - which uses highly ambiguous terms, only understandable in local contexts
  - which is characterised by constantly new terms
- 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 (SNOMED CT)	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

Popularity of terms in medical records

Preferred term (ICD, OPS – German)	Count	Synonym (German)	Count
Aortenklappenstenose	3749	Aortenstenose	3126
Hirninfarkt	7	Schlaganfall	65
Elektrokardiogramm	0	EKG	12208
Koronare Herzerkrankung	331	KHK	18455
Nicht-ST-Hebungsinfarkt	498	NSTEMI	3839
Magnetresonanztomographie	2	NMR	17

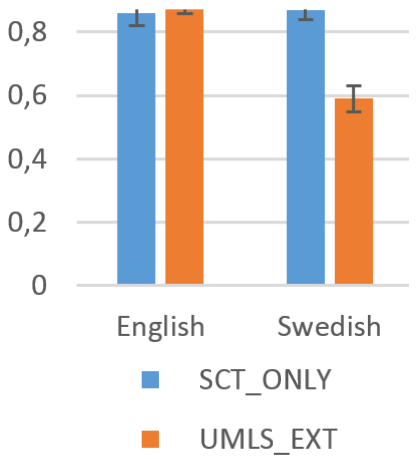
ASSESS CT

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

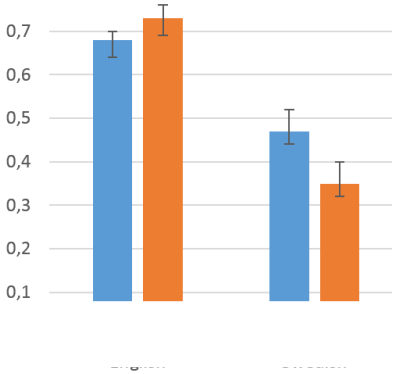
"Is SNOMED CT well suited as an European reference terminology ?"

- 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



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 definitions
  - (formal definitions → ontology)

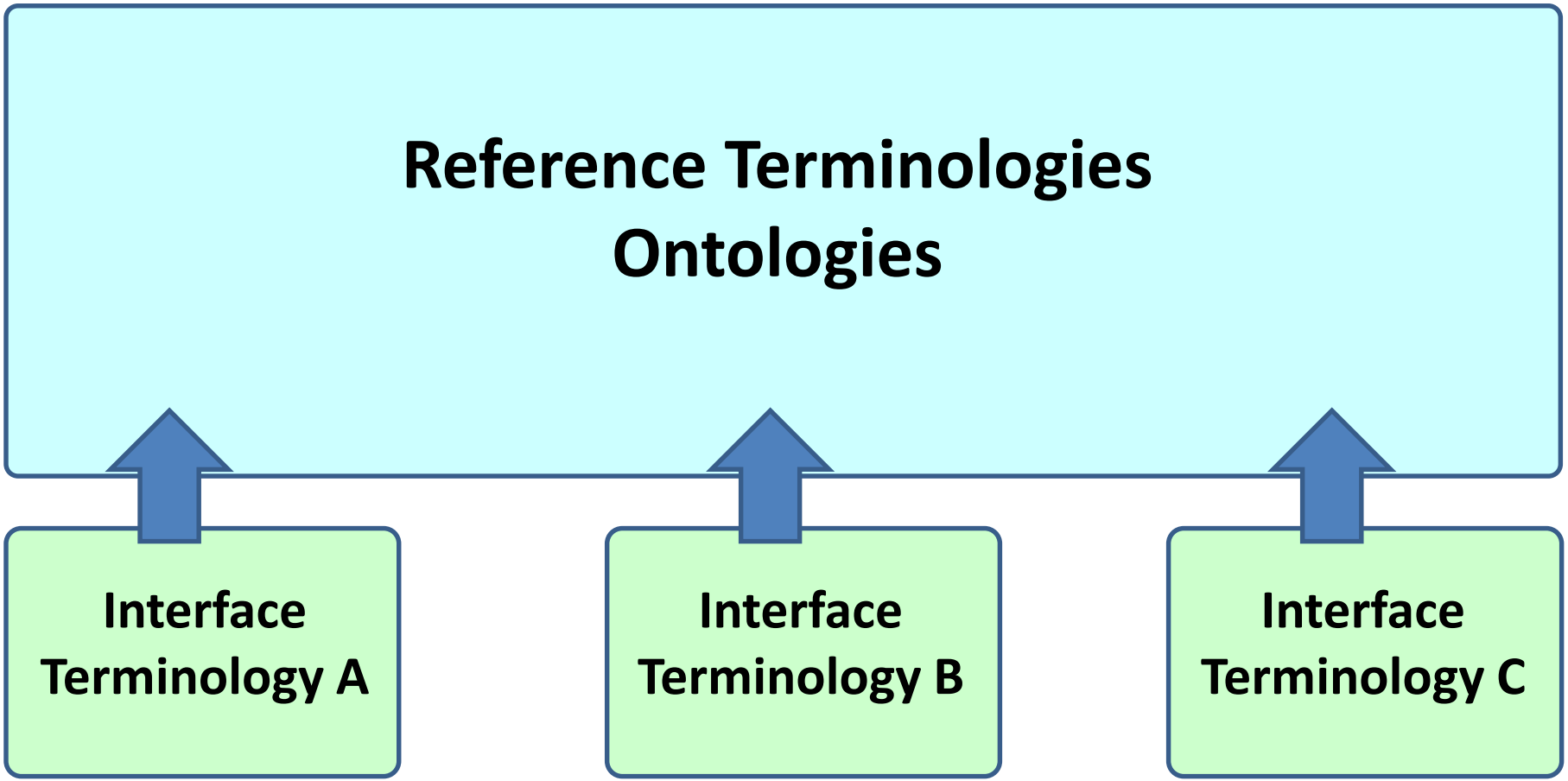
Reference Terms

- Descriptive
  - Describes human language expressions as being used "*Lung cancer*"

Interface Terms

- Many domain terminologies / ontologies contain interface terms, but far from being exhaustive

Separation



ASSESS CT

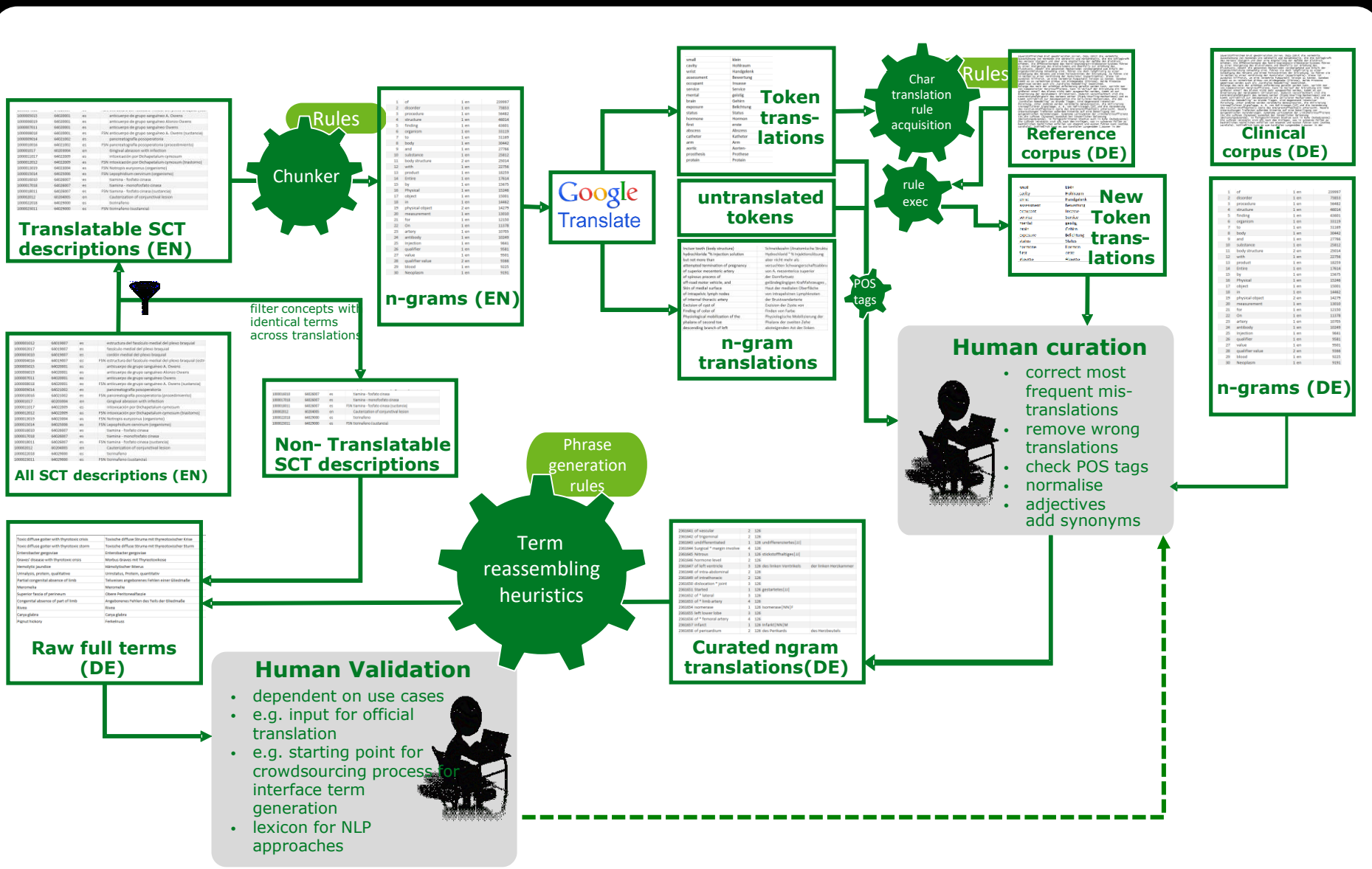
ASSESS CT Recommendations:
http://assess-ct.eu/fileadmin/assess_ct/final_brochure/assessct_final_brochure.pdf

Case study: German interface terminology for SNOMED CT

- Context: CBmed IICCAB* (see poster)
- Limited resources, incremental approach
- Top-down:
  - Modularization of original terms (split into ngrams):
    - "*Magnetic resonance imaging* *of neck* *with contrast*"
  - 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:
  - Collection of popular interface terms from n-gram frequency lists extracted for clinical corpora

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

Case study: German interface terminology for SNOMED CT



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: preference 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 German n-grams, out of 85,400 English n-grams
  - Benchmark: parallel corpus extracted from Medline: current 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

Scaling up…

- Taming size:
  Crowdsourcing for terminology development

- Taming ambiguity:
  Document preprocessing using language models

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, Doctor
  - 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

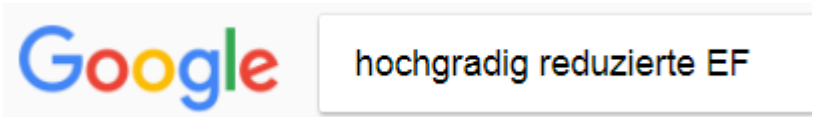
Example: Resolution of short forms

- "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
```

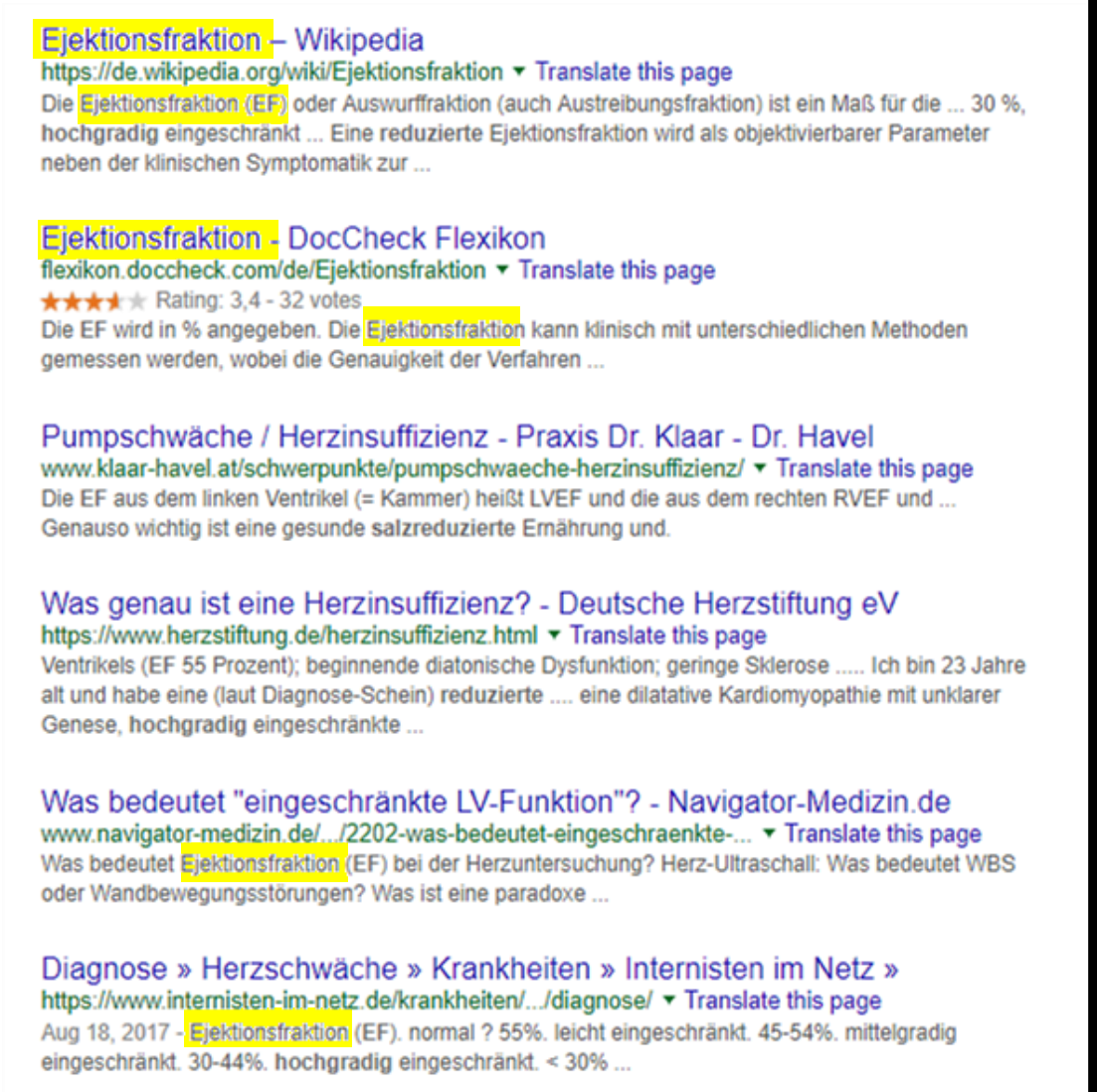
- Web mining



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

- Interpretation:
  - Salience
  - Acronym-definition patterns
  - Regular expressions created from short forms
- Accuracy of method currently studied



Recommendations

- Need for interface terms not satisfied by domain terminologies
- Dynamic acquisition of interface terms needed
  - 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