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 2 Workshop

NLP and Ontology

Stefan Schulz, Sareh Aghaei, Remzi Celebi

Funded by the European Union

15 February 2024

The meaning of “Ontology”

- System of identifiers that
  - Denote the meaning of **concepts (types of entities, classes)** and **relations (properties)**
  - Provides **formal or textual definitions** and **axioms and constraints**
- Example:
  - The SNOMED CT code 1303400001 denotes the class “Fracture of phalanx of right thumb”. It is defined as the class of all fractures of bone at some Thumb with the laterality Right
  - The relation hasBodySite in SPHN denotes the relations between a procedure and a body site
- Kinds of ontologies:
  - Foundational ontologies: provide domain-independent categories and relations, e.g. Quality, Disposition, Process, inheres_in, continuant_part_of. Example: BFO
  - Domain upper-level: describes what is important in a domain, e.g. SPHN…
  - Domain ontologies, e.g. SNOMED CT
  - Information models such as FHIR can also be seen as ontologies of information
- Ontologies do not describe all kinds of knowledge; they are limited to what is universally true for all members of a class

Important ontological class distinctions

- Process-like vs. object-like (e.g., operation vs. tissue)
- Dependent vs. independent (e.g., blood pressure vs. blood)
- Domain entities vs. information entities (e.g. cancer and “suspected cancer”

Additionally, important distinctions in health care:

- Conditions (problems) – everything a patient can have: being bald, having a seizure, having renal failure, having a broken nose, being short-sighted,..
- Procedures – everything that is done on a patient by a health professional: taking blood pressure, making a chest x-ray, transplanting a heart
- Observables – everything that can be measured and takes a value to make sense (body weight, glucose in blood, number of pregnancies)
- Specimens – everything taken from an organism for examination (tissue specimens, blood, urine, organs, organ parts after operations of autopsies)
- Supportive categories: body parts, organisms, devices, substances, drug products, roles, qualities, dispositions

Aligning SNOMED CT upper-level concepts with SPHN

AIDAVA

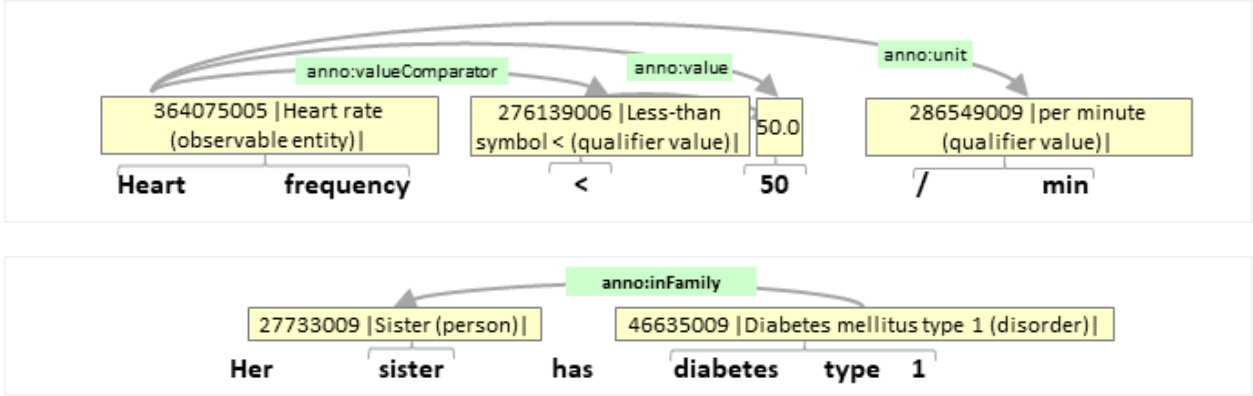
Relevant upper-level concepts

SPHN	SNOMED CT	Issues
sphn:ProblemCondition sphn:Diagnosis, sphn:PhysiologicState, sphn:NursingDiagnosis	Clinical Finding, Disorder, Event	No relations for body site. No mechanism to express factuality (confirmed / suspected / negated)
sphn:Allergy	SNOMED 418038007 | Propensity to adverse reactions to substance (finding)	Slot for factuality (verificationStatusCode) would be equally necessary for ProblemCondition
sphn:MeasurementMethod Sphn:LabTest	<< 386053000 |Evaluation procedure (procedure)|	
sphn:Measurement sphn:LabResult sphn:Quantity	<< 363787002 |Observable entity (observable entity)|	If 363787002 |Observable entity (observable entity)| exists, no need to specify sphn:MeasurementMethod. Otherwise: subclass of 386053000 |Evaluation procedure (procedure)| and instantiate unspecified observable. sphn:Quantity is not instantiated
sphn:Procedure	Procedure	No link to substance
sphn:Drug	373873005 |Pharmaceutical / biologic product (product)|	Difference between product and substances unclear
sphn:TumorSpecimen	258435002 |Tumor tissue specimen (specimen)|	Only Tumor Specimen
(...)		

Observations

- Many SPHN classes can fully be expressed by more general ones, e.g. sphn:BloodPressure by sphn:Measurement
- Important relations missing (e.g. verificationStatus only for Allergy but not for ProblemCondition
- Cardinality = 1 for codes precludes use of postcordinated codes
- Targets sometimes retired SNOMED concepts (419199007 \|Allergy to substance (finding)\|) , concepts to be retired (363743006 \|Navigational concept (navigational concept)\|) or wrong concepts (439401001 \|Diagnosis (observable entity)\|,)

Aligning annotation predicates with SPHN



- Standard annotation semantics:
  - A text span denotes a particular entity in the domain
  - The annotation of a span identifies the concept it instantiates
  - The predicate annotation identifies the relationship that holds between two domain entities
- Annotation relations (binary predicates)
  - Correspond to OWL object properties or data properties
  - Used in annotation and therefore in NLP output
  - Are grounded in SNOMED CT and/or FHIR relational patterns
- From annotations to knowledge graph
  - Direct correspondences: above example
  - Indirect correspondences, e.g.
    Diabetes_e6w4r anno:inFamily ***Sister_iw8l*** <=>
    Diabetes_e6w4r inverse(246090004 |Associated finding (attribute)|) o 408732007 |Subject relationship context (attribute)| ***Sister_iw8l***
- SPHN relations can mostly be mapped to anno: relations
- Some anno: relations do not correspond to SPHN relations. SNOMED relations must be used instead

Aligning anno: properties with SPHN

Frequent predicates in text mining

anno: predicates	Range::Domain according to SNOMED, cf. Annotation Guideline	correspondence to SPHN
after	Finding::Finding, Event::Event, Procedure::Procedure	No predicate
beginTime	Finding::Finding, Event::Event, Procedure::Procedure	sphn:hasStartDateTime
dueTo	Finding::Finding, Event::Event, Procedure::Procedure	sphn:hasReasonToStopCode However, the description of this predicate and domain and range do not fulfil the requirement
endTime	Finding::Finding, Event::Event, Procedure::Procedure	sphn:hasEndTime
inFamily	Finding or Event::Social concept	not in sphn
ingredient	Product::Substance	sphn:hasActiveIngredient sphn:hasInactiveIngredient With the range and domain of sphn:Drug and sphn:Substance, respectively.
laterality	BodyStructure or Finding or Procedure or Specimen::Side	sphn:hasLaterality no chain of relations from sphn:ProblemCondition to sphn:Laterality
sameAs	*::*	To be managed at the instance-level (merge of two instance).
siteDirect	Finding or Procedure or Specimen or Event::BodyStructure	sphn:hasBodySite – Domain incomplete, e.g. sphn:ProblemCondition missing
value	*::decimal ObservableEntity::Qualifier Value	
valueComparator	*Mathematical sign	
unit	Product or Substance::Unit of measure(qualifier value)	sphn:hasUnit With the range and domain of sphn:ReferenceRange and sphn:Quantity, respectively.
usingSubstance	Procedure::Substance	not in sphn
verificationStatus	Finding or Event::Qualifier Value	in sphn limited to Allergy

(Preliminary) conclusions

- Syntactically, SNOMED CT, AIDAVA anno predicates and SPHN are fully compatible: use of OWL
- Many SPHN classes are not necessary for AIDAVA: less diversification – focus on selected classes recommended. Example: Blood pressure, body height, BMI, etc. are observables (for which SNOMED CT concepts exist): no need for diversification at SPHM class level
- Part of relevant annotation relations can be mapped to SPHN, others cannot: SNOMED relations need to be used in the knowledge graph
- Important relations such as sphn:HasBodySite and sphn:hasVerificationStatusCode are not available where they are expected, particularly sphn:ProblemConditiom
- No need for SPHN value sets and individuals – if needed can be expressed by SNOMED concepts (most SPHN individuals are ontologically no individuals)

Next steps

- Better defining the scope of SPHN content that is used
- Identifying important gaps in SPHN and communicating to SPHN curators
- Analysing correspondences between AIDAVA text annotation scheme (which characterizes future NLP output) and SPHN output
- Creating one common OWL model
- Exploring compatibility with upper-level ontologies
- (...)