New concept proposal

Source Data

Author	**Kristin Gnodtke, Harald Witte**	**Date last updated**	**11.12.2023**
Project	General interest	**Contact**	DCC
Dataset release	2024.1	**Consulted expert**	Katie Kalt, Amanda Ramirez Ramos

1 Rationale

Adherence to the FAIR principles (Wilkinson *et al.*, 2016, DOI: 10.1038/sdata.2016.18) maximizes the value of research data by making it findable, accessible, interoperable, and reusable. This includes providing data origin and source data information to researchers as metadata along with the actual dataset. Source data information, e.g., on data transformation or mapping events, makes data FAIR, reliable, and suitable for reuse by third parties.
Importantly, **the proposed concept "Source Data" refers to source data in the context of codes**, e.g., locally used hospital analysis codes or gender codes (as opposed to standardized codes like LOINCs or SNOMED CT-codes as the outcome of a semantic mapping). **It does not entail source data in the sense of raw data**, for example unprocessed images or a data set underlying a machine learning model.

2 Comparison to other standards/data models

2.1 NCIT

NCIT uses the term "Source Document" (C142692) to describe "The original information in a variety of media, which may substantiate information contained in an electronic record of the study.". The longer, alternative description describes "Source Document" as "Original documents, data, and records (e.g., hospital records, clinical and office charts, laboratory notes, memoranda, subjects' diaries or evaluation checklists, pharmacy dispensing records, recorded data from automated instruments, copies or transcriptions certified after verification as being accurate copies, microfiches, photographic negatives, microfilm or magnetic media, x-rays, subject files, and records kept at the pharmacy, at the laboratories, and at medicotechnical departments involved in the clinical trial)"

2.2 OMOP (Observational Medical Outcomes Partnership)

OMOP features various terms related to data provenance spread across several tables, e.g., OBSERVATION_PERIOD, VISIT_DETAIL, VISIT_OCCURRENCE, or PERSON (https://ohdsi.github.io/CommonDataModel/cdm54.html#Clinical_Data_Tables). (Syntax: xxx_source_xxx attribute, e.g., gender_source_value "[...] provider's gender as it appears in the source data.")

2.3 DICOM Controlled Terminology

The DICOM Controlled Terminology knows various terms revolving around Source Data.
The term "Source raw data" (DICOM: 128226) is used without further information, probably assuming it is self-explanatory.
Source Document (DICOM: 121335): Document whose content has been wholly or partially transformed to create the referencing document.
Source Image (DICOM: 121324): Image used as the source for a derived or compressed image.
Source report (DICOM: 128225): Report used as the source for derivation.
Source measurement (DICOM: 121335): Measurement used as the source for derivation.

3 Concept information

Concept or concept compositions or inherited	General concept name	General description	Contextualized concept name	Contextualized description	Type	Standard	Value set or subset	Meaning binding	Cardinality for composedOf
Concept	Source Data	unprocessed data	Source Data	raw data that has not been processed for meaningful use					
composedOf	code	coded information specifying the concept	code	coded information specifying the source data description	Code				0:1
composedOf	string value	textual representation	textual source data description	source data description expressed using free text	string				0:1

Concept cardinalities:

General concept name	cardinality for concept to Administrative Case	cardinality for concept to Data Provider	cardinality for concept to Subject Pseudo Identifier	cardinality for concept to Source System
Source Data				1:n

Leave field empty if not applicable.

4 Impact on the SPHN Dataset

Optional (if existing concepts need to be adapted because of this new concept, state here the currently released version of the existing concept and the proposed adapted version)

5 Discussion

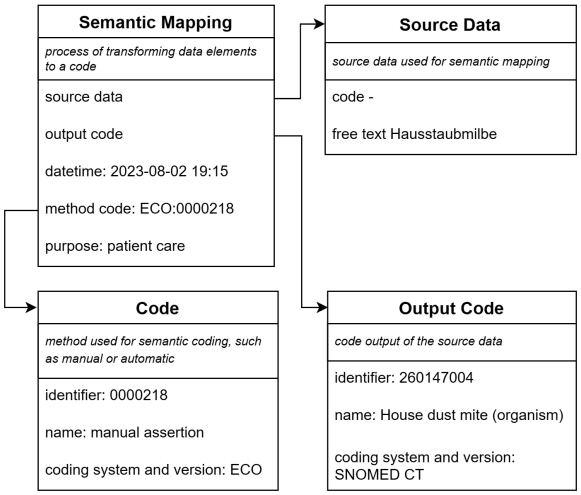
A direct link between the Source Data concept and Source System had initially been discussed but ruled out to avoid duplicating information as the Source System is linked to most data concepts in the SPHN Dataset on the concept level.
After an evaluation using comprehensive data from CDWHs, however, the link of Source Data to Source System has been introduced with a cardinality of 1:n. The lack of a direct link may cause problems in the context of other concepts, e.g., Semantic Mapping. When data originates from different source systems (e.g., KISIM and PDMS), codes of the Source Data are different but the output code of a semantic mapping is identical. This renders the original code used in the source data non-traceable.
The cardinality 1:n for the link between single concepts and the Source System concept reflects that data elements represented by one concept can come from different source systems. For example, multiple source systems of Diagnosis imply that DiagnosisCode may be from source system A and DiagnosisSubjectAge from source system B.

6 Example

6.1 Example 1

Source data

code:
identifier: -
name: -
coding system and version: -
string value: Hausstaubmilbe



6.2 Example 2

Source data

code:

identifier: 1

name: weiblich

coding system and version: KISIM

string value: -