New concept proposal
Semantic Mapping

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

1 Rationale

Various stakeholders have expressed interest in collecting and sharing the provenance of data. Information about data provenance i.e., metadata that explains how the data were generated, by whom, and where they came from, gives an indication about data type and quality and allows the researcher to assess whether the data are suitable for his or her type of data science (fit for purpose). The proposed new concept Semantic Mapping represents information about the transformation of data elements to a (ideally standardised) code and therefore refers to the source data as well as the mapping method.
From a knowledge-centric perspective, a process is an event which operates on input(s) and can yield output(s) (Figure 1A). In this sense, the concept Semantic Mapping represents the process which operates on source data (input) like non-standard codes or strings and transforms it into a standardised code (output) which may for example be linked to a Result (Figure 1B).

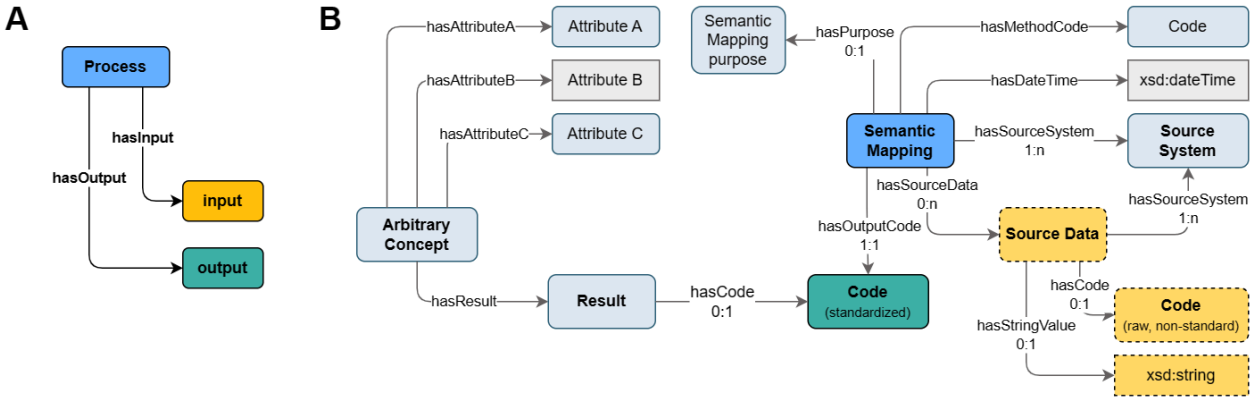


***Figure 1**: Semantic Mapping in the context of knowledge-centric (process-oriented) concept design*

2 Comparison to other standards/data models

2.1 OMOP

The OMOP Common Data Model contains variables referring to source data represented as a code or string:

- [Event]_SOURCE _CONCEPT_ID, concepts that represent the code used in the source, e.g. diagnosis code in the source system (it is a code)
- [Event]_SOURCE_VALUE, original representation of an Event recorded in the source data, e.g. notion of a code in the source system, e.g. m for male (it can be a code and it can also be a short free text phrase)

2.2 ECO

The Evidence and Conclusion Ontology (ECO) describes types of scientific evidence within the biological research domain that arise from laboratory experiments, computational methods, literature curation, or other means (https://www.ebi.ac.uk/ols/ontologies/eco).

2.3 NCIT

The NCI Thesaurus defines the term “Data Mapping” (C142485) as “The process of connecting an item or symbol to a code or concept.”. The CDISC-GLOSS definition is “In the context of representing or exchanging data, connecting an item or symbol to a code or concept.”

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	Semantic Mapping	process of transforming data elements to a code	Semantic Mapping	process of transforming data elements to a code					
composedOf	source data	source data associated to the concept	input data	source data used for semantic mapping	Source Data				0:n
composedOf	output code	output associated to the concept	mapping output	output of the semantic mapping	Code				1:1
composedOf	datetime	datetime of the concept	mapping datetime	datetime of the semantic mapping	temporal				0:1
composedOf	method code	coded information specifying the method of the concept	method code	method used for semantic coding, such as manual or automatic	Code	ECO	descendant of: ECO:0000217		0:1
composedOf	purpose	objective of the concept	purpose	objective of the semantic mapping	qualitative		Billing; Patient Care; Quality Control; Research		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
Semantic Mapping		1:1		1:n

Leave field empty if not applicable.

4 Impact on the SPHN Dataset

The datetime connected to the new *Semantic Mapping* concept replaces the coding datetime connected to Diagnosis and Medical Procedure concepts (see example in Figure 2). Thus, the composedOf 'coding datetime' of these concepts should be deleted.

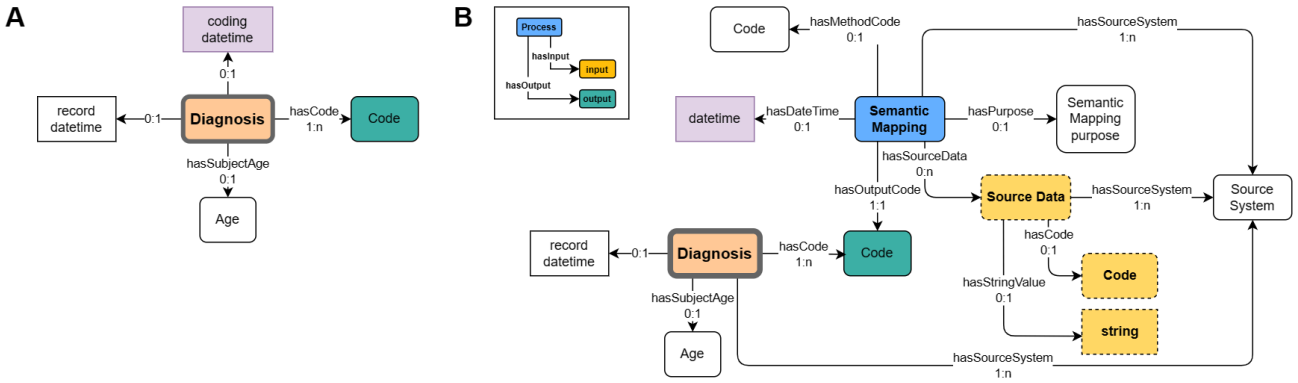


Figure 2: Representation of 'coding datetime' in the 'Diagnosis'-concept

(A) Explicit 'coding datetime' (purple) of Diagnosis (release 2023.2) (B) Implicit 'coding datetime' of 'Diagnosis' using the 'datetime'-composedOf of 'Semantic Mapping' (release 2024.1). The dashed borders of 'Source Data' (yellow) emphasise the optional character of 'Source Data' in the context of 'Semantic Mapping' (cardinality 0:n). Inset: Process-oriented design: Processes feature input and output.
Note: 'coding datetime' and 'record datetime' (A) as well as 'method code' and 'output code' (B) are shown as separate nodes for better accessibility.

5 Discussion

The concept 'Semantic Mapping' is designed and intended to cover both mapping and coding events. Its description "process of transforming data elements to a code" covers both of these cases. For coding events, however, the input data may not always be fully available or it may not be possible to represent all input components taken into account properly in the schema. The **cardinality of the composedOf 'source data'** (type 'Source Data') has therefore been **set to 0:n** to cover such cases and still be able to represent output codes (output code without formal input).

A single code can originate from multiple (local) codes, hence a cardinality of 1:n for the composedOf source data is suggested. This would apply, for example, if multiple local analysis codes map to a single LOINC. It is expected that the source data is either coded or present as string. Therefore, the cardinalities for the composedOf source data code and source data value should be 0:1.

Notably, **multiple instance of 'source data'** in a 'Semantic Mapping' **imply an AND-logic**, e.g., the SNOMED CT-codes 88897007 | Malignant tumor, fusiform cell type AND 9801004 | Spindle cell sarcoma are all required for mapping to the ICD-O-3 output code of 8032/3 Spindle cell carcinoma, NOS) (Figure 3A).

Different independent codes mapping to one identical code need to be represented by multiple instantiations of 'Semantic Mapping', e.g., the terms "female", "weiblich", "feminine", "f" etc., are all mapping to the SNOMED CT-code 248152002 |Female (finding) (Figure 3B).

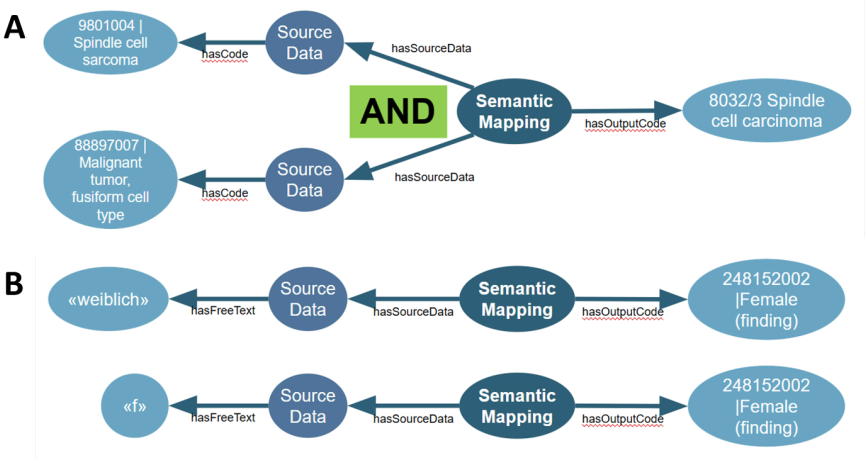


Figure 3: *Meaning of multiple instances of 'Source data' (A) and of 'Semantic Mapping' (B). Discussed in Semantic Working Group on 30.08.2023.*

The composedOf datetime is somewhat ambiguous, however, this may even be beneficial because the date and time it refers to may vary between data providers. For example, in one hospital, it may refer to the time a semantic mapping was applied, in another one it may refer to the point in time when a mapping table was created. In both cases, however, it shall be possible to trace back which mapping was applied, e.g., for quality control, provided that internal operating procedures and tracking are in place.

The composedOf 'source system' of Semantic Mapping covers meta-data to meta-data, e.g. what is the source system of the SemanticMappingPerformer.

The new concept will increase the FAIRness of data by providing solid provenance information. It is possible to extend the concept by additional properties, e.g. information on validation status, coder (person who carried out the mapping, e.g., data engineer, physician, study nurse, bioinformatician), confidence in the mapping, or further details on the mapping method (manual, automatic, AI, NLP etc.) should the need for this information surface at some point.

The composedOf 'purpose' of Semantic Mapping will provide the opportunity to see what was the objective or reason for semantic mapping. The table below provides example situations and the corresponding value. The value set is aligned between the concepts Semantic Mapping and Source System.

Example situation	Corresponding value for purpose
data has been coded in ICD 10 or CHOP for billing purposes	Billing
data has been coded in LOINC in the laboratory	Patient Care
data has been coded in SNOMED CT for a registry	Quality Control
data has been coded in SNOMED CT for a research project	Research

6 Example

6.1 Example 1:

Mapping of a local KISIM code (identifier: 1; name “weiblich” (=”female”) to a SNOMED CT code (identifier: 248152002; name: “Female (Finding)”)

Semantic Mapping

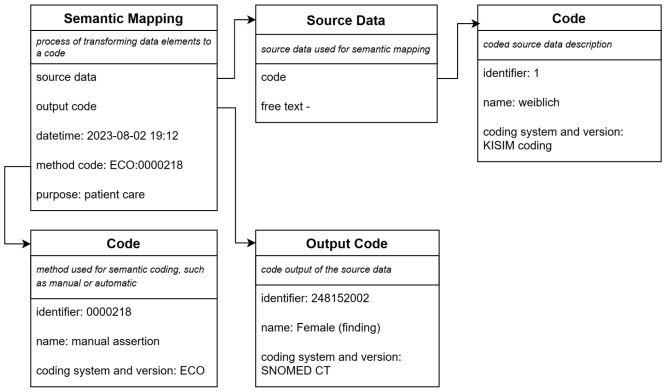
source data
- code
  - identifier: 1
  - name: weiblich
  - coding system and version: KISIM
- free text -

output code
- identifier: 248152002
- name: Female (Finding)
- coding system and version: SNOMED CT

datetime 2020-03-15T14:23:00+01:00

method code
- identifier: 0000203
- name: automatic assertion
- coding system and version: ECO

purpose: Patient Care



6.2 Example 2:

Mapping of a local KISIM term (free text: “Hausstaubmilbe” (=”dust mite”) to a SNOMED CT code (identifier: 260147004; name: “House dust mite (organism)”)

Semantic Mapping

source data
- code
  - identifier: -
  - name: -
  - coding system and version: -
- free text Hausstaubmilbe

output code
- identifier: 260147004
- name: House dust mite (organism)
- coding system and version: SNOMED CT

datetime 2020-03-22T15:01:00+01:00

method code
- identifier: 0000218
- name: manual assertion
- coding system and version: ECO

purpose: Patient Care