New concept proposal

Data Processing

Author	Eelke van der Horst, Femke Kopmels	**Date last updated**	31/10/2023
Project	General interest	**Contact**	DCC
Dataset release	2024.1	**Consulted expert**	-

1 Rationale

An essential part of scientific disciplines is processing the data produced by assays to retrieve an analysis result. Especially in data-intensive domains such as omics, data processing makes up a significant part of the experiment. Usually, individual processing and analysis steps are chained together into a (bioinformatics) pipeline. As part of data processing, data may be transformed from one format or structure to another, or may be subjected to computing to produce aggregates and other analysis results. To evaluate and reproduce these results, metadata on the data processing steps, such as the software/script that was used, is required.

2 Comparison to other standards/data models

2.1 OBI and EFO

OBI has a 'data transformation' class (OBI:0200000) that has the synonym 'data processing'. It is defined as "A planned process that produces output data from input data". EFO imports this class. This class is equivalent to the proposed concept, and its subclasses may be used as terms to indicate the type of data processing.

2.2 EDAM

Most items under EDAM's Operation (EDAM:operation_0004) are data processing steps, and may be used as code to indicate the particular type of processing.

2.3 SIO

The Semanticscience Integrated Ontology (SIO) has an 'information processing' class that is similar to data processing but broader, since it also includes 'data collection'.

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	Data Processing	a process that produces output data from input data	Data Processing	a process that produces output data from input data					
composedOf	code	coded information specifying the concept	code	code specifying the nature of data processing	Code	EDAM, OBI or other	for EDAM: descendant of EDAM:operation_0004 |Operation|; for OBI: descendant of: OBI:0200000 |data transformation|		1:1
composedOf	software	software associated to the concept	software	software used for data processing	Software				0:1
composedOf	input	input associated to the concept	input	input data file	Data File				0:n

composedOf	output	output associated to the concept	output	output data file	Data File				0:n
composedOf	start datetime	datetime at which the concept started	start datetime	datetime at which the data processing started	temporal				0:1
composedOf	quality control metric	quality control metric associated to the concept	quality control metric	quality control metric related to the output of the data processing	Quality Control Metric				0:n
composedOf	predecessor	a preceding process associated to the concept	predecessor	process preceding this data processing	Data Processing; Assay			RO:0002087 |immediately preceded by|	0:n
composedOf	standard operating procedure	standard operating procedure associated to the concept	standard operating procedure	standard operating procedure that was followed for this data processing	Standard Operating Procedure				0:1

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**
Data Processing	0:n	1:1	0:n	1:1

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

Data Processing can be used for any data processing step for which the used software should be indicated, such as BCL to FASTQ conversion. The *Data Processing* concept can be used to indicate sub-steps of a broader process when there is a need to provide metadata for individual steps.

Usually, a data processing step has at least one input file. However, there are cases where intermediate files between steps are not known or important. Therefore, the minimum cardinality is 0.

6 Example

Genome annotation (a data processing step in a genomics pipeline)

code: EDAM:operation_0362 |Genome annotation|
processing datetime: 2023-06-26
input:
name: ExampleFile1
data provider institute: EXAMPLE01
output:
name: ExampleFile2
data provider institute: EXAMPLE01
predecessor:
code: EDAM:operation_3182 |Genome alignment|