New concept proposal

Quality Control Metric

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

Quality control metrics are used to express the quality of a product or process. They help identify defects, errors, or deviations from established standards. By capturing these metrics, users can ensure that the data meets quality criteria.

2 Comparison to other standards/data models

2.1 FAIR Genomes

FAIR Genomes *Sequencing* module has some QC metrics defined as elements, such as percentage Q30 and percentage Tr20. However, there is also an additional "Other Quality Metrics" *catch-all* attribute that allows for additional QC metrics. The value of this attribute is a string and it is defined as: "Other NGS quality control metrics, including but not limited to (i) sequencer metrics such as yield, error rate, density (K/mm2), cluster PF (%) and phas/prephas (%), (ii) alignment metrics such as QM insert size, GC content, QM duplicated reads (%), QM error rate, uniformity/evenness of coverage and maternal cell contamination, and (iii) variant call metrics such as number of SNVs/CNVs/SVs called, number of missense/nonsense variants, common variants (%), unique variants (%), gender match and trio inheritance check. (EDAM:data_3914)". We propose a more structured and consistent approach by introducing a *Quality Control Metric* concept that is used for each QC metric value and that reuses the *Quantity* concept for its values.

2.2 EDAM

The EDAM ontology defines a 'Quality Control Report' as "A human-readable collection of information about how a scientific experiment or analysis was carried out that results in a specific set of data or results used for further analysis or to test a specific hypothesis.". The Quality' Control Report' is a type of 'Report' which is a type of 'Data'.

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	Quality Control Metric	report of the quality control review that was made of factors involved in a procedure	Quality Control Metric	report of the quality control review that was made of factors involved in a procedure				EDAM: data_3914 |quality control report|	
composedOf	code	coded information specifying the concept	qualitative result code	coded information specifying the quality control metric	Code	SNOMED CT, GENEPIO or other			0:1
composedOf	quantity	an amount or a number of the concept	quantitative result	value and unit of the quantitative result of the quality control metric	Quantity				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
Quality Control Metric	-	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

Some of the proposed concepts require a *Quality Control Metric*. Note that although this information can be mandatory, it is not necessarily guaranteed that data quality will meet or exceed a specific standard; it still needs to be evaluated by the data user, and not all users hold on to the same data quality standards. One or multiple *Quality Control Metric*s can be assigned to a single concept. As there is no (relatively) complete list available for quality control metrics in existing ontologies/vocabularies, we propose no specific value set for the 'code' composedOf.

6 Example

code: GENEPIO:0000089 |phred quality score|
quantity:
value: 78.33
unit: %