New concept proposal

Software

Author	Jan Armida, Deepak Unni	**Date last updated**	31/10/2023
Project	General interest	**Contact**	DCC
Dataset release	2024.1	**Consulted expert**	The Hyve

1 Rationale

A piece of software is a computer program capable of executing a logical sequence of commands in a computer or in any machine and programmable electronic device. Often, the preparation and analysis of data requires the use of specific pieces of software, that by definition, greatly differ in terms of functionality and domains of application. To ensure reproducibility, pieces of software involved in the generation of the data needs to be properly described, including its version and, when possible, its source code.

In order to document any Software used regardless of the application (e.g., variant calling pipeline software, alignment software, image processing software) we propose a simple concept where the software used is identified by its name, version and URI. The URI, ideally, will be a unique link to a repository (e.g. an URL linking to a repository) where the software is deposited, version controlled and thoughtfully described. This solution applies to both custom and existing software.

2 Comparison to other standards/data models

2.1 IAO

In Information Artifact Ontology (IAO), there is a ‘Software’ concept (IAO:0000010) described as “Software is a plan specification composed of a series of instructions that can be interpreted by or directly executed by a processing unit”. This concept is used in various downstream ontologies like OBI, EFO, DUO and SWO.

2.2 SNOMED CT

In SNOMED CT, the concept of Software is conveyed in function of its application within the healthcare system. The concept 706687001 | Software (physical object) | lies under the Computer equipment parent concept and has three children that covers a large spectrum of software applications such as programs used to measure cardiac outputs or ECGs:

- Application program software (physical object)
- Operating system software (physical object)
- X-ray system software (physical object)

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 composed Of
concept	Software	set of procedures and instructions in a data processing system	Software	set of procedures and instructions in a data processing system				706689003 |Application	

								program software (physical object)|	
composedOf	name	name associated to the concept	name	formal name given to the software application	string				1:1
composedOf	description	description associated to the concept	description	human-readable description of the software application	string				0:1
composedOf	uniform resource locator	uniform resource locator (URL) associated to the concept	url	uniform resource locator (URL) associated with the software (e.g. link to a repository)	string				0:1
composedOf	version	version of the concept value or code	version	version of the software application used to generate the data	string				1:1

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
Software	-	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

The main purpose of the concept *Software* is to provide the minimal information needed to identify pieces of Software used to generate the data. One instance of the *Software* concept will therefore include the software name and a valid URL pointing to an external repository (ideally GitHub). The decision to leave the software description as a free-text field, instead of relying on controlled vocabularies, is dictated by the huge diversity of software applications that would need to be described. This solution will also allow more flexibility for pieces of software lacking a valid URL. Information about order of execution of software (e.g., classic pipeline analysis) should be defined by a separate concept.

6 Examples

In the following example, a simple implementation of the concepts is used to describe the HaplotypeCaller from the Genome Analysis Toolkit (GATK):

Software:
- name: HaplotypeCaller
- description: HaplotypeCaller is used to call SNPs and indels simultaneously via local de-novo assembly of haplotypes in an active region.
- url: https://gatk.broadinstitute.org/hc/en-us/articles/360037225632-HaplotypeCaller
- version: 4.1.4.1