Change request
Data File, Hash

Author	**Katie Kalt**	**Date of request updated**	**01.11.2023**
Project	IICU	**Contact person**	DCC
Dataset release	2024.1	**Consulted expert**	IICU Time Series Team

1 Change request input / rationale

Every data file has a creation date time and this should be part of the properties of the data file. This allows the validation if the data file referenced in the concept is the same one as physically present.

Change 1: add creation datetime as composedOf
Every data file has a creation date time and this should be part of the properties of the data file. This allows the validation if the data file referenced in the concept is the same one as physically present.

Change 2. Change cardinalities
The cardinality for Subject Pseudo Identifier is now changed from mandatory (1:1) to relaxed (0:1) to allow for scenarios where the Data File can exist without a corresponding Subject Pseudo Identifier.
Another change is the introduction of a Hash concept which will be associated with the Data File.

Change 3: Add hash code as composedOf
Hash codes associated with data files are important for verifying data integrity, ensuring reproducibility, and enhancing security by acting as digital fingerprints that detect changes or corruption. This is particularly important in the case of large-scale biological data, such as omics, where the integrity of the file directly affects the validity of all downstream applications.

2 Comparison to other standards/data models

n/a

3 Change content

3.1 Currently released concept

Concept or concept compositions or inherited	General concept name	General description	Contextualized concept name	Contextualized description	Type	Standard	Value set or subset	Meaning binding
Concept	Data File	electronic resource of information, which can be stored, accessed and transferred as a single unit	Data File	electronic resource of information, which can be stored, accessed and transferred as a single unit				
composedOf	name	name associated to the concept	file name	name given to the data file	string			
composedOf	uniform resource identifier	unique identifier of the concept that allows the system to identify all the information needed to access the resource	Uniform resource identifier	unique identifier that allows the system to identify all the information needed to access the resource	string			SNOMED CT: 1119461003 |Uniform resource identifier (foundation metadata concept)|
composedOf	format code	code, name, coding system and version describing the format of the concept	format	format of the data file	code	EDAM		

3.2 Proposed new concept (Data File)

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	Data File	electronic resource of information, which can be stored, accessed and transferred as a single unit	Data File	electronic resource of information, which can be stored, accessed and transferred as a single unit					
composedOf	name	name associated to the concept	file name	name given to the data file	string				0:1
composedOf	uniform resource identifier	unique identifier of the concept that allows the system to identify all the information needed to access the resource	uniform resource identifier	unique identifier that allows the system to identify all the information needed to access the resource	string				0:1
composedOf	format code	coded information specifying the format of the concept	format	format of the data file	code	EDAM	descendant of: EDAM:format_1915 |Format|		0:1
composedOf	creation datetime	datetime the concept was created	creation datetime	datetime the file was created	temporal				0:1
composedOf	hash	hash associated to the concept	hash	hash associated to the file	Hash				0:1
composedOf	encoding	encoding of the concept	encoding	encoding of the file	qualitative		UTF-8; UTF-16; ASCII; ISO-8859-1		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 File	-	1:1	0:n	1:1

3.3 Proposed new concept (Hash)

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	Hash	irreversible unique number computed on an information entity used to check its validity and integrity	Hash	irreversible unique number computed on an information entity used to check its validity and integrity					
composedOf	string value	textual representation	hash code value	string representation of the hash value	string				1:1
composedOf	algorithm	algorithm applied to the concept	algorithm	algorithm used to calculate the hash code	qualitative		SHA-256; MD5; SHA-512; SHA-1		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
Hash	-	-	-	-

4 Pros and cons

4.1 Advantages

Allows cross checking the referenced file with the physical file.

4.2 Disadvantages

none

5 Impact on SPHN Dataset

As there is no change in the existing definition there is no impact.

6 Discussion

Parquet is a possible format that would be used in SPHN projects. A request is sent to EDAM for integrating it into their terminology: https://github.com/edamontology/edamontology/issues/740.

Note: SHA-1 is listed as a possible hashing algorithm, but it is important to note that security concerns have been raised on the use of this algorithm. Therefore, it would be recommended that providers use one of the other algorithms' listed.

7 Example

name: SRR1653111.fastq

uri: ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR165/001/SRR1653111/SRR1653111_1.fastq.gz

format code: EDAM:format_1930 |FASTQ|

creation datetime: 15.09.2023

encoding: ASCII

hash:

- string value: 7b1646fc6239e8eac8544233cc94ade7
- algorithm: MD5