New concept proposal

Data File

Author	**Kristin Gnodke** **Jan Armida**	**Date last updated**	**01/12/2022**
Project	-	**Contact person**	Kristin Gnodke
Status	Accepted	**Consulted expert**	Semantic WG

1 Rationale

In addition to structured data such as diagnosis and procedure, there is a need to share files with raw data for research purposes such as machine learning. Raw data files can have different formats and contain different types of information. A raw data file can be an image such as a CT scan, a document such as a pathology report or a binary file containing sequence alignment data. A file can be described by its name, format and location within a file system (locally or on a web site's folder structure) as well as other metadata, such as the creation date or encryption status. In order to enable the reference to single or multiple raw data file we propose *Data File* as a concept in the SPHN dataset.

2 Comparison to other standards/data models

2.1 GA4GH Phenopackets

The GA4GH Phenopackets have a direct implementation of the concept File to link the structured phenotypic data it contains to external files which can be used to inform analysis. The element requires a valid URI to identify the file (allowing both internal and external references) and a series of identifiers to associate the file to the individual ID and the biosample ID. Extra information about the file can be added using the *file_attributes* field which, however, does not make use of any controller vocabulary of fixed value-set. Each resource is free to add as many attributes as necessary to share information about the contents of a file.

Field	Type	Multiplicity	Description
uri	string	1..1	A valid URI e.g. file://data/file1.vcf.gz or https://opensnp.org/data/60.23andme-exome-vcf.231?1341012444. REQUIRED.
individual_to_file_identifiers	a map of string key: value	0..1	The mapping between the Individual.id or Biosample.id to any identifier in the file. RECOMMENDED.
file_attributes	a map of string key: value	0..1	A map of attributes pertaining to the file or its contents.

2.2 UMLS

In the UMLS Metathesaurus there are two concepts related to file and of interest.

- *Sequencing Data File* with the concept unique identifier C5401191 with the definition: “An electronic file containing nucleic acid sequencing data.”;
- File (record) with the concept unique identifier C0016094 with the definition: “A set of related records (either written or electronic) kept together.”.

2.3 SNOMED CT

SNOMED CT is certainly not the ontology to find a concept coming from information technology rather than from healthcare. However, the concept File exists in SNOMED CT with another meaning as we want to describe in this document: 12953007 |File, device (physical object)|, e.g. a bone file.

2.4 HL7 FHIR

In HL7 FHIR, there are three resources dealing with files.

- DocumentReference: A reference to a document of any kind for any purpose. Provides metadata about the document so that the document can be discovered and managed. The scope of a document is any serialized object with a mime-type, so includes formal patient centric documents (CDA), clinical notes, scanned paper, and non-patient specific documents like policy text.
- DiagnosticReport: The findings and interpretation of diagnostic tests performed on patients, groups of patients, devices, and locations, and/or specimens derived from these. The report includes clinical context such as requesting and provider information, and some mix of atomic results, images, textual and coded interpretations, and formatted representation of diagnostic reports.
- ImagingStudy: Representation of the content produced in a DICOM imaging study. A study comprises a set of series, each of which includes a set of Service-Object Pair Instances (SOP Instances - images or other data) acquired or produced in a common context. A series is of only one modality (e.g. X-ray, CT, MR, ultrasound), but a study may have multiple series of different modalities.

2.5 Other sources

Wikipedia describes a computer file as follows “A computer file is a computer resource for recording data in a computer storage device.” (https://en.wikipedia.org/wiki/Computer_file).

Segen's Medical Dictionary. © 2012 Farlex, Inc. All rights reserved: term: file, description: “A basic unit of storage on a computer; a collection of data which can be stored, accessed and transferred as a single unit. A file may contain text, calculations, graphics or software routines.”

3 Concept information

Concept name	Description	Type	Standard	Value set	Meaning binding
Data File	electronic resource of information, which can be stored, accessed and transferred as a single unit				
file name	name given to the data file	string			
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) |
format	format of the data file	Code	EDAM	descendant of: 1915 format	

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 Data File is to offer a construct capable of linking any concept to a file containing the raw data. Each file is uniquely identified by its resource ID which is a combination of its internal ID and data provider institute. In order to improve readability, a filename needs to be specified for each file. This attribute is normally composed by the base name of the file followed by its extension, although syntax and format for a valid filename might vary across operating systems.

When needed, one or more Uniform Resource Identifiers (URI) can be associated with the file to further specify the file location. Valid URI must follow a precise syntax consisting of various hierarchically organized components. In general, the following rule applies for absolute URI: <scheme>:<scheme-specific-part>. As an example, Uniform Resource Locators (URL) used to locate a resource on a computer network are a specific type of URI. A URL will typically follow this syntax: http://www.example.com/file.extension.
Although file extension and file format are often referred as interchangeable terms, these do not forcibly match in reality. For this reason, for each instance of *Data File* a format needs to be specified by means of the extensive format branch of the EDAM ontology. Ultimately, *Data File* provides a minimal set of meta-information and can be easily extended to accommodate the need of more complex file types (e.g., file containing genomic coordinates) that might benefit from additional attributes (e.g., genome assembly).

6 Examples

Example 1 – FTP URL:

file name: SRR1653111.fastq
uri: ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR165/001/SRR1653111/SRR1653111_1.fastq.gz
format:
identifier: format:1930
name: FASTQ
coding system and version: EDAM-1.25

Example 2 – No URI:

file name: PD-1-MELANOMA-00012.vcf
format:
identifier: format:3016
name: VCF
coding system and version: EDAM-1.25

Example 3 – Internal directory

file name: HF1316_report.pdf
uri : file://data/dataprovider/reports/HF1316_report.pdf
format:
identifier: format:3508
name: PDF
coding system and version: EDAM-1.25