New Concept proposal

Time Series Data File

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

Time series data can be very hard to generate and manage in RDF. Also most projects will process time series data in a ML pipeline for which they need to extract the data from the graph into a csv or parquet file anyway.

As a consequence we have decided to allow time series to be represented in an alternative format than RDF, as an external file. This could be either CSV or Parquet format. Additionally the data needs to be captured by sensors.

Data recorded manually e.g. at the wards must remain by default in the RDF format but could be added into a time series data file if needed. This has to be defined by the project requesting the concepts. Note that within a project there can not be a mix of 'time series' data provided as RDF and provided as a time series data file.

The concepts that comply with the new Measurement -> Result pattern and that can be represented in this alternative file format are the following:

- Blood Pressure
- Body Temperature
- Cardiac Output
- Heart Rate
- Oxygen Saturation
- Respiratory Rate

This pattern would allow the project to re-integrate the time series data in the alternative format into the knowledge graph using e.g. ontop mappings. As a consequence there should be separate time series files for each concept. A Time Series Data File should be created for each patient (Subject Pseudo Identifier).

The Time Series Data File concept takes into account a Data File change request proposal for release 2024.1: https://docs.google.com/document/d/1nVtV1Sm_j6JRv6P_3SZkq2SagVNeUl0V/edit.

For more information about the structure of the data file, see section 7 later in this document.

2 Comparison to other standards/data models

Please refer to the evaluation and specification file for further details:
https://docs.google.com/document/d/1SASCTsoGlrmo1toN6Gn8TctdhheMSdi2SuFX_GOY69A/edit#

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
Concept	Time Series Data File	electronic resource that contains all the results related to a measurement as time series	Time Series Data File	electronic resource that contains all the results related to a measurement as time series	Data File				
inherited	name	name associated to the concept	file name	name given to the time series data file	string				0:1
inherited	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
inherited	format code	coded information specifying the format of the concept	format	format of the data file	code	EDAM	descendant of: EDAM:format_1915		0:1
inherited	hash	hash associated to the concept	hash	hash associated to the file	Hash				0:1
inherited	creation datetime	datetime the concept was created	creation datetime	datetime the file was created	temporal				0:1

inherited	encoding	encoding of the concept	encoding	encoding of the file	qualitative		UTF-8; UTF-16; ASCII; ISO-8859-1		1:1
composedOf	entry count	count of entries in the concept	entry count	count of the time series entries	Quantity	UCUM	unit -> code restricted to: {#}		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
Time Series Data File	0:1	1:1	1:1	1:1

4 Pros and cons

4.1 Advantages

Allows additional metadata about a data file.

4.2 Disadvantages

none

5 Impact on SPHN Dataset

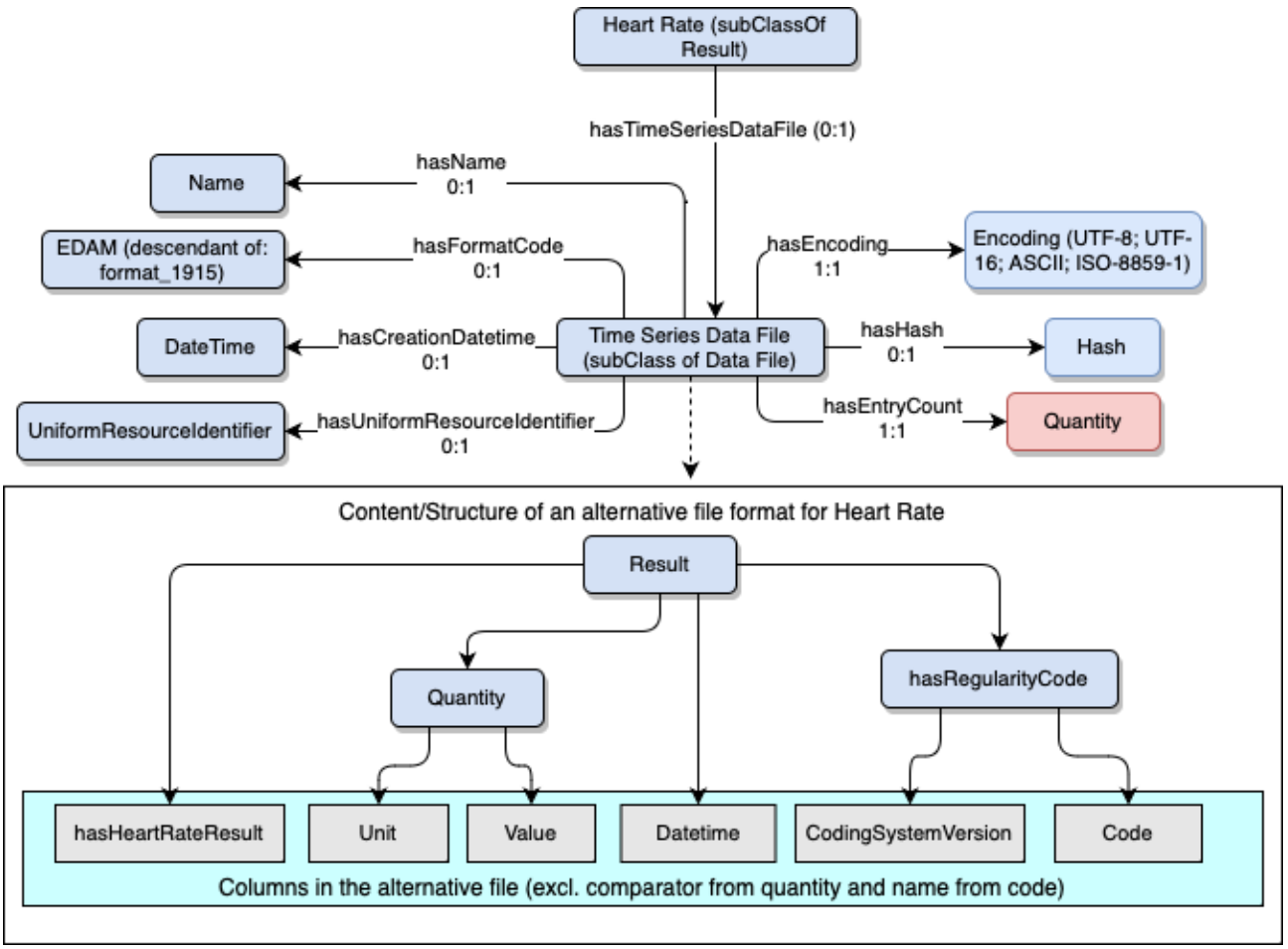
As there is no change in the existing definition there is no impact.

6 Discussion

The taxonomy EDAM does not list the format Parquet yet which is a possible file format for Time Series Data File. A request has been sent here: https://github.com/edamontology/edamontology/issues/740.

7 Content/Structure of the Time Series Data File

The content of the data file are the properties of the concept Result excluding the string value but adding a link to the related result instance. The coding system and version and the code are only present if there is also a code linked to the result like the regularity code in the Heart Rate.



Column naming according to SPHN Connector schema

"sphn_hasHeartRate",
"sphn_hasDateTime",
"sphn_hasQuantity__id", (default = “Quantity-hasValue-hasUnit__id”)
"sphn_hasQuantity__sphn_hasComparator__iri", (optional)
"sphn_hasQuantity__sphn_hasUnit__id",
"sphn_hasQuantity__sphn_hasUnit__sphn_hasCode__iri",
"sphn_hasQuantity__sphn_hasValue",
"sphn_hasRegularityCode__iri", (optional)
"sphn_hasRegularityCode__termid", (optional)

- Aligned to the SPHN Concept and would allow in the future for the SPHN Connector to create the alternative file.
- Clear naming of the content
- Easier to map back into the graph

8 Example

name: heart-rate_4193EB08-5A4E-47F5-95DD-516D3A348484.csv
uri: file:/data/CHE-108_904_325/heart-rate_4193EB08-5A4E-47F5-95DD-516D3A348484.csv
format code: EDAM:format_3752 |CSV|
creation datetime: 15.09.2023
hash:
    string value: 34044584BEE2
    algorithm: SHA-256
encoding: UTF-8
entry count: 36293