Mandate: RDF working group

27 July 2022, version 4

1 Background

The Swiss Personalized Health Network (SPHN) aims to develop, implement, and validate a coordinated infrastructure to make health and related data FAIR (Findable, Accessible, Interoperable and Reusable) for research in Switzerland. From the SPHN initiative, various efforts are currently underway to harmonize and define data standards to ensure the interoperability of health-related data, with a focus on data obtained from the five Swiss University Hospitals (UH).

On recommendation of the Hospital IT "Data Exchange Format" Task Force, the SPHN National Steering Board (NSB) endorsed the use of Resource Description Framework (RDF), a Semantic Web Standard (W3C). RDF has been chosen as one of the formats of choice for data transport and storage in SPHN projects in autumn 2020. The SPHN RDF schema, developed by the SPHN Data Coordination Center (DCC) in collaboration with representatives of the UH, is in constant evolution and regular releases are done to improve the schema content representation but also enrich it with more metadata as expressed in the SPHN dataset. It represents the "template" to which data submitted for research purposes must conform. Therefore, hospitals are building RDF data extraction pipelines to transform data coming from their clinical data warehouses into an RDF format, compliant with the SPHN RDF schema.

Until today, the DCC and the SPHN partners started the development of several infrastructure elements (see also RDF Roadmap Progress Report 2021[1]), among them:

- The SPHN RDF schema;
- Quality control tools with the SHACLer and the Quality check tool;
- Terminology service;
- A user guide on how to generate and analyze RDF data;
- RDF extraction pipeline at each university hospital;
- SPHN trainings on RDF data.

[1] https://sphn.ch/wp-content/uploads/2022/02/Progress-Report-RDF-Roadmap-2021.pdf

Based on this development experience, the following challenges and gaps have been identified and need to be tackled:

- Currently, the RDF data generation in the hospitals is very time consuming;
- Many steps are done manually which can lead to implementation errors;
- The necessary compute capacity for data generation and validation is often missing in the hospitals;
- Changes and adaptations to SPHN or project-specific RDF schemas require major, often manual, adaptations of the hospital-specific pipelines;
- Quality control checks (SHACLs) cannot be executed at some sites.

2 Vision and Mission

The vision is to build an easy-to-use streamlined processes to deliver high-quality SPHN compliant RDF data in each hospital. This includes:

- further development of the SPHN semantic framework and related tools;
- ensuring the compliance of data deliveries with the SPHN semantic framework and the implementation of relevant controlled vocabulary and ontologies,
- warranting that the data generation mechanisms are efficiently implemented in the hospital Information and Communication Technology (ICT) infrastructure in alignment with the SPHN Information Technologies (IT) architecture goals,
- allowing adjustments and improvements of the interoperable data generation workflow in a timely manner (e.g. with regards to the new SPHN RDF releases),
- ensuring quality control (SPHN compliance) of the delivered data packages;
- providing education and documentation on the use of the SPHN semantic framework and related tools.

3 Composition of the Working Group

The HIT-STAG nominated the following representatives of the UH into the WG:

- CHUV: Yves Jaggi, Alexandre Wetzel
- HUG: Pierre Dethare, Adel Bensahla
- Insel: Guido von Matt
- USB: Rita Achermann
- USZ: Katie Kalt, Barbara Jesacher

The WG is chaired by a member of the Personalized Health Informatics (PHI) Group. The RDF working group will work closely with the BioMedIT RDF Support WG, the SPHN IT Architecture WG, the SPHN Semantic WG and the Data Standards and Data Quality WG.

4 Work Packages

4.1 WP1 Continuous development and improvement of the SPHN RDF schema

Responsible person	**WG chair**
Aim	The goal of WP1 is to continue the development of the SPHN RDF schema as needed with new semantics and improve it as required.
Task WG members	• Provide feedback when new developments are presented. • Suggest and present new implementation ideas for particular topics. • Review of the SPHN RDF schema and related tools and services
Timeline	Continuous until end of SPHN funding period

4.2 WP2 Professionalization of the Hospital RDF pipeline development

Responsible person	**Hospital representatives**
Aim	The goal of WP2 is to further improve the hospital internal pipelines for the RDF generation to meet the following criteria (defined in the HospFAIR agreement): • sufficient compute power for the data validation and the overall data delivery pipeline (comparable to a low latency large size database; exact requirements will be determined according to the individual set-ups) • flexibility of the pipeline to react to new RDF schema releases and additional requirements by the projects • automation of repetitive steps • triple store deployment with the capacity to load at least all data of one data delivery including incremental loads UH provides the final validation reports and pipeline runtimes of each data delivery to DCC.
Task WG members	• Implementation and documentation of the RDF pipelines as stated in the aim.
Timeline	HospFAIR phase 1 until 31.3.2023

4.3 WP3 HospFAIR data deliveries

Responsible person	**Hospital representatives**
Aim	Provide RDF data in the realm of the HospFAIR project for quality control.
Task WG members	• Extract the required data; • Use the RDF pipeline to generate RDF data; • Validate the RDF data with the SPHN SHACLs; • Send data to the HospFAIR BioMedIT project space.
Timeline	HospFAIR phase 1 until 31.3.2023, timeline for each data delivery will be defined in collaboration with all HospFAIR WGs.

4.4 WP4 Quality analysis of the HospFAIR data deliveries

Responsible person	**Hospital representatives**
Aim	Improve the quality (compliance to specification) of RDF data produced in the realm of SPHN.
Task WG members	• Validate the quality control (against defined structural quality parameters) of RDF data; • Support the DASAQ in analyzing RDF data (SPARQL or SHACL); • Provide feedback to the provider of the analyzed dataset; • This task will be supported by the RDF Support WG and the DCC.
Timeline	HospFAIR phase 1 until 31.3.2023, timeline will be defined in collaboration with all HospFAIR WGs.

5 Reference and further reading

[1] S. Österle, V. Touré, and K. Crameri, “The SPHN Ecosystem Towards FAIR Data,” *preprint.org*, 2021, doi: 10.20944/preprints202109.0505.v1.

[2] P. Krauss, V. Touré, K. Gnodtke, K. Crameri, and S. Österle, “DCC Terminology Service—An Automated CI/CD Pipeline for Converting Clinical and Biomedical Terminologies in Graph Format for the Swiss Personalized Health Network,” *Appl. Sci.*, vol. 11, no. 23, p. 11311, Nov. 2021, doi: 10.3390/app112311311.

[3] SPHN semantic framework available on git https://git.dcc.sib.swiss/sphn-semantic-framework/sphn-ontology

[4] SPHN semantic framework documentation and user guide available on https://sphn-semantic-framework.readthedocs.io/en/latest/