RDF Roadmap
Progress Report 2021

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, and the NSB mandated the SPHN Data Coordination Center (DCC) to develop a roadmap for the implementation of the necessary infrastructure components. This document outlines the work packages (WP) of the RDF Roadmap and summarizes activities, developments and achievements of the involved actors in the year 2021.

WP 1: Further development of the SPHN RDF schema

The SPHN RDF schema provides an interoperable framework for the transport and storage of health data for SPHN-funded projects, making it one of the key deliverables of the SPHN strategy for meeting the FAIR criteria. The schema needs to be continuously optimized to exploit the full potential of RDF. It also facilitates the integration of and connection to existing external resources within the SPHN framework. The SPHN RDF schema, based on the SPHN dataset, transforms elements of the SPHN dataset into a formal structure. More information is available here.

Goal of this work package is to expand, improve and professionalize the SPHN RDF schema.

Aim 1: Define conventions for transforming the SPHN dataset representation in the SPHN RDF schema (e.g. concepts, properties, meaning binding, value sets)

Aim 2: Include new concepts from SPHN Driver Projects and other SPHN funded projects

Aim 3: Identify and address issues and gaps in the current SPHN RDF schema (e.g. versioning, connectivity to external terminologies)

Deliverable D1: New release of the SPHN RDF schema (this is a yearly deliverable)

Deliverable	Responsibility	Timeline	Status	Details
D1	DCC	May 2021	Done	The SPHN RDF schema 2021 can be downloaded from DCC GitLab or directly explored online on DCC WebProtégé. Major improvements addressed: • New Oncology and ICU concepts • Following good practices and RDF grammar rules (conceptual errors) • Implementation of controlled vocabularies as defined in WP2 • Instantiation of SPHN value sets in RDF and references to the external terminologies (ATC, LOINC, SNOMED CT etc.)

				• Introduction of hierarchies in classes and properties
D1	DCC	Sep 2021	Done	The SPHN RDF schema 2021.2 release with bug fixes is accessible on GitLab and in WebProtégé. All URIs are human and machine readable, available on https://www.biomedit.ch/rdf/sphn-ontology/sphn/2021/2.

WP 2: Use of controlled vocabulary

The use of controlled vocabulary greatly improves data interoperability and facilitates data understanding and reuse. In this work package, the focus is put on the selection, agreement and use of international vocabularies for all already defined concepts in the SPHN Dataset and Schema. For further releases, this is directly included in the concept definition for the SPHN Dataset.

Aim: Use international vocabulary (e.g. SNOMED CT, LOINC) to semantically describe SPHN concepts

Deliverable D1: Mapping between the SPHN concepts and different controlled vocabulary is included in the SPHN Dataset and SPHN RDF Schema, continuously updated at each release.

Deliverable	Responsibility	Timeline	Status	Details
D1	DCC	May 2021	Done	The SPHN Dataset 2021 includes, where possible: • Meaning binding for concepts to SNOMED CT and/or LOINC • Value set binding to SNOMED CT

WP 3: Guidelines for Driver projects to expand the ontology

The SPHN ontology can be extended by projects to fit their individual needs. However, the extension needs to be carefully thought through to be consistent with the data interoperability strategy. The goal of this work package is to build templates, guidelines and comprehensive documentation on the extension of the SPHN RDF schema.

Aim: Develop common guidelines for driver projects to expand the SPHN ontology following best practices

Deliverable D1: Guidelines for expanding the SPHN ontology

Deliverable D2: A template ontology to be used by researchers to easily start their project-specific definition of the semantics (a new template will be provided for each new RDF Schema)

Deliverable	Responsibility	Timeline	Status	Details
D1	DCC	Jun 2021	Done	The guidelines include

				• Use of the Ontology Template (D2) • Design of new concepts • Expansion of existing concepts
D2	DCC	Sep 2021	Done	The Template Ontology helps projects to develop their project-specific RDF schema. This template imports the SPHN RDF schema and other required RDF-related and terminology-related libraries as well as pre-filled basic metadata annotations (User guide).

WP 4: Guidelines for data instantiation

The SPHN RDF schema provides the specification data providers have to follow to represent the data. The process of data instantiation must follow specific rules to be compliant with the RDF framework. The goal of this work package is to develop understandable guidelines and examples to facilitate the process of data instantiation.

Aim: Produce common guidelines for the instantiation of data in RDF across all data providers

Deliverable D1: National guidelines for data instantiation according to the SPHN RDF Schema

Deliverable	Responsibility	Timeline	Status	Details
D1	DCC	Sep 2021	Done	The User guide includes • Versioning of the data • IRI naming conventions • Instantiation of external resources • Several examples of data instantiation • Cardinality of the data implemented in the SHACL rules

WP 5: Development of quality control guidelines and scripts

Once data is generated from the data providers, an important step is to ensure that the exported data meets the requirements of the SPHN RDF specification (quality control). In this work package, the goal is to develop guidelines and tools to facilitate the process of data validation.

Aim: Delivery of high-quality data according to the SPHN RDF specifications

Deliverable D1: A set of SHACL rules to validate the data, based on the SPHN RDF schema (a new SHACL set will be provided for each new RDF Schema)

Deliverable D2: A script/tool to create SHACL rules from a project specific RDF Schema

Deliverable D2.2: Include more features into the tool to create SHACL (depending on the community needs)

Deliverable D3: A set of SPARQL queries to provide summary statistics on the delivered data

Deliverable D4: A script/tool for automatic checks, accessible to data providers and researchers

Deliverable D5: A set of guidelines/rules on how to extend the SHACL rules for project specific schemas and data.

Deliv-erable	Re-sponsi-bility	Time-line	Status	Details
D1	DCC	Sep 2021	Done	The developed SHACL rules set includes: • Cardinality constraints • Compliance to ontology • Compliance to used standards • Restricting on individuals • Data type constraints
D2	DCC	Dec 2021	Done	A tool, the so called SHACLer (beta version available on request) was developed, which generates SHACL rules from an ontology that facilitates the SHACL generation.
D3	DCC	Dec 2021	Done	The developed Statistics queries provide information about: • SPHN concepts used in the data • Project specific concepts used • No. of patients, hospitals, and patients/hospital • Average values and ranges e.g. for measurements or dates
D4	HUG	Dec 2021	Done (beta version	SHACLs (D1) and SPARQL (D3) can be bundled in the tool developed by HUG.
D5	DCC	Dec 2021	Done	A first version of the User guide on how to use the SHACLer (D2) was developed. A more detailed User guide on manual extension will follow in 2022.

WP 6: A central RDF generation pipeline at each data providing institution

The transformation of clinical resources into a format that fits the SPHN semantic framework requires to map internal data to international standards and transform the data into SPHN compliant RDF. With every new release of the SPHN RDF Schema, these pipelines need to be updated and extended. Until a common SPHN architecture is in place (coordinated by the SPHN Architecture WG), there is a need to invest in the pipelines at the various data providing institutions to systematically facilitate data export according to the standards in order to provide data in RDF to (i) the ongoing SPHN/PHRT (driver) projects and (ii) the upcoming National Data Streams (assuming that the SPHN architecture solution will not be ready until summer 2022).

These processes should be replaced (or better: centralized) by the implementation of the common SPHN architecture.

Aim 1: Ability to extract clinical data according to newest SPHN RDF specifications in a (semi) automated way

Aim 2: Committed contribution to the development of a harmonized SPHN Architecture (with a Graph Database at the core, linking routine data for research)

Aim 3: Generalize pipelines to be used by other hospitals (in the realm of the Architecture WG)

Deliverable D1 (each hospital): Process definition of data delivery in RDF (mapping to SPHN concepts, coding in required standards, RDF generation, quality control)

Deliverable D2 (each hospital): Central implementation of the above-mentioned process for high quality RDF data export from CDWs.

Deliverable D3 (each hospital): Technical documentation of the RDF Export pipeline

Deliverable D4 (each hospital): Process definition and timeline how to update the pipeline according to new SPHN RDF Schema releases also following the WP4 guideline.

Deliverable D5 (each hospital): Implementation of the above-mentioned process for RDF schema updates

These deliverables have been postponed to the HospFAIR program in 2022.

WP 7: Terminology service

External standard terminologies are being used in the SPHN semantic framework to facilitate data integration and understandability thanks to the use of existing dictionaries. The use of these terminologies needs to be facilitated for both data providers and data users with a common platform that delivers these terminologies in SPHN compliant RDF, properly versioned and directly usable.

Aim: Provide terminologies used in the RDF schema to data providers and researchers in RDF

Deliverable D1: CHOP, ATC, SNOMED CT, UCUM, ICD-10 and LOINC are available in RDF on request to the DCC. (Including older versions and new versions as soon as a new version is released)

Deliverable D2: Set a list of additional terminologies needed (this is a yearly occurring milestone)

Deliverable D3: Make additional terminologies available in RDF (this is a yearly occurring milestone)

Deliverable D4: First prototype of a terminology service for BioMedIT (automated generation, testing and loading via a CI/CD pipeline)

Deliverable D5: Download area for terminologies in the BioMedIT Portal

Deliv-erable	Respon-sibility	Time-line	Status	Details
D1	DCC	Mar 2021	Done	Python scripts were developed for all conversion. The following terminology files were created: • CHOP (historical versions since 2016) • ICD-10 GM (historical versions since 2014) • SNOMED CT (historical versions since 2021-01-31) • LOINC (historical versions since version 2.69) • ATC (2021 version) • UCUM
D2	DCC	Oct 2021	Done	Terminologies; ICD-O, Spezialitätenliste and EMDN
D3	DCC	Mar 2022	In pro-gress	EMDN (available on request) Spezialitätenliste and ICD-O (in progress)
D4	DCC	Oct 2021	Done	The DCC Terminology service consists of a gitLab Runner pipe-line and a MinIO server. The service is operational, accounts can be requested by the single responsible persons in the BioMedIT nodes and hospitals (for individual users see D5). A paper de-scribing the developed service is published https://www.mdpi.com/2076-3417/11/23/11311.
D5	DCC	Sep 2021	Done	The Terminology service in the BioMedIT Portal is developed, including user management for compliance with license agree-ment e.g. for SNOMED CT. All terminologies of D1 are available for download.

WP 8: Documentation and visualization

With all developments made in the SPHN semantic interoperability framework, it is crucial to have a detailed documentation that gathers all knowledge and guides the projects in their development.

Aim: Document content and developments of the SPHN RDF schema for record, information, and guidance purposes

Deliverable D1: Full documentation of the SPHN RDF schema using the readthedocs.io platform

Deliverable D2: Representation and visualization of the SPHN RDF schema (a new visualization will be provided for each new RDF Schema)

Deliverable D3: First version of a user guide for the extension and use of the SPHN RDF schema (linked to WP3, WP4 and WP5 and WP11)

Deliverable	Responsibility	Timeline	Status	Details
D1	DCC	Jun 2021	Done	A read the docs including: • SPHN framework Introduction • SPHN Dataset • SPHN RDF Schema • Quality Assurance Framework
D2	DCC	Oct 2021	Done	A web-based visualization is published under https://biomedit.ch/rdf/sphn-ontology/sphn
D3	DCC	Jun 2021	Done	A User guide including the following section was developed: • Access, read and use of SPHN content • Generating a SPHN project-specific ontology • Data generation following the project ontology • Data quality assurance for validation • Exploratory data analysis with triple stores • Querying data with SPARQL • How to use Python and R with RDF data

WP 9: Tools and services for RDF

To facilitate the different step along the project lifecycle, different tools and services are needed. This WP encompasses the selection of tools, central or local (project space) deployment, or license agreements for commercial software.

Aim: Provide researchers with a software stack to handle ontologies and be able to analyze RDF data

Deliverable D1: Outline an RDF ecosystem/tool stack for BioMedIT

Deliverable D2: Accessibility of recommended tools for handling and analyzing RDF data within the BioMedIT infrastructure

Deliver-able	Respon-sibility	Timeline	Status	Details
D1	DCC	Jun 2021	Done	Next to the self-developed tools (e.g. SHACLer and SPARQler), three main tools were identified: • GraphDB as a triple store, • Protégé as a desktop ontology editor and • WebProtégé, the web-based version of Protégé, for a collaborative ontology editing.
D2	DCC	Dec 2021	Done	• GraphDB can be used on all three BioMedIT nodes (licenses are provided centrally) • A DCC WebProtégé instance is hosted centrally and can be used by all SPHN users (31.12.2021 51 users).

WP 10: Training

To facilitate the use of semantic web standards and corresponding tools (WP9) a comprehensive training portfolio shall be developed and made freely accessible.

Aim: Provide researchers and data providers with a series of training on RDF and related standards

Deliverable D1: A set of modular training session on the SPHN RDF tool stack (see WP9) and W3C standards namely RDF, SPARQL and SHACL

Deliv-erable	Respon-sibility	Timeline	Status	Name and Link of the Training
D1	DCC	2021	Done	The DCC developed 7 FAIR trainings session on semantics and RDF. Training and Training material are made FAIR: • Videos are published on the SPHN website and Youtube • Training materials are published on the SPHN website and the DCC git • A user guide for each topic is available on the ReadtheDocs • All videos and training materials are under the CC-BY-SA license
		May 2021	Done	Intro to RDF and SPARQL
		Sep 2021	Done	How to design new concepts (Dataset and Protege)
		Sep 2021	Done	Introduction to existing standards used in SPHN
		Oct 2021	Done	Data exploration and visualization in GraphDB
		Oct 2021	Done	Querying data with SPARQL
		Oct 2021	Done	Use of RDF data in Python and R
		Nov 2021	Done	Validate graph data with SHACL

WP 11: Research support on BioMedIT

A user guide with comprehensive examples will train the researcher on the use of semantic web standards (e.g. RDF, SPARQL and SHACL). General training session will be organized as part of WP10. To provide a hands-on support, a research support position is created in each BioMedIT node.

Aim 1: Ability of researchers to use the data on BioMedIT either directly or to transform their data into a target format or application of interest

Aim 2: Capacity building at the BioMedIT nodes to provide research projects with support regarding the SPHN RDF tool stack (see WP9) and WC3 standards namely RDF, SPARQL and SHACL

Deliverable D1: Provide a set of SPARQL examples on how to extract data and transform them into another format (e.g. .csv)

Deliverable D2: A tool supporting researcher to generate SPARQL queries in an automated way.

Deliverable D3: A user guide on exploratory data analysis with triple stores.

Deliverable D4: A qualified RDF support person is available in each node to help researchers.

Deliver-able	Respon-sibility	Timeline	Status	Details
D1	DCC	Sep 2021	Done (beta version)	A set of helper queries for each class for better under-standing the content of a dataset is currently under devel-opment
D2	DCC	Dec 2021	Done (beta version)	A tool, called SPARQLer to extracts example SPARQL queries for each concept from an SPHN compliant input Ontology is available as beta version
D3	DCC	Oct 2021	Done	A first version of the user guide is available. Covering: • More SPARQL examples (inference) • Use of RDF data in R • Use of RDF data in Python
D4	Bio-MedIT	Oct 2021	In progress	SIS and SIB already allocated / hired. sciCORE person will a start Q1 2022. Education of these persons ongoing.

WP 12: Outreach

Outreach in form of presentation and written material is important to raise visibility and educate the community on the advancements of the SPHN Semantic Framework.

Aim: Spread the knowledge and understanding of the SPHN RDF schema and its usage in clinical settings

Deliverable D1: Participation in appropriate meetings and organization of workshops

Deliverable D2: Scientific publications

Deliverable D3: Other publication e.g. Website and Fact sheets

Deliver-able	Respon-sibility	Timeline	Details
D1	DCC	Continuous	The DCC gave/will give the following talks: • 02.06.2021 SPHN webinar • 11.09.2021 Full day BC2 Tutorial - Semantic representation of clinical data in RDF • 13.09.2021 Expert talk BC2 conference • 21.09.2021 Trivadis Brown bag • 03.10.2021 BioDataWorld congress • 14.10.2021 GA4GH Clin/Pheno WS Roundtable
D2	DCC	Continuous	Scientific publications: • Gaudet-Blavignac C, et al., JMIR Med Inform 2021;9(6):e27 • Österle S. et al., Preprints, 2021 • Krauss P, Tet al. *Applied Sciences*. 2021; 11(23):11311
D3	DCC	Continuous	Fact sheet: • SPHN Semantic Strategy • RDF SPHN Website: SPHN Semantic Strategy News: • SPHN Semantic Framework 2021 release • Terminology Service of the SPHN Data Coordination Center (DCC) • The SPHN Ecosystem Towards FAIR Data

For questions contact Dr. sc. ETH Sabine Österle sabine.oesterle@sib.swiss