RDF Roadmap
Progress Report 2022

On the 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.
After a successful implementation of the RDF infrastructure in 2021, the DCC and the related working groups (BioMedIT RDF Research Support WG; RDF WG; Semantics WG; IT Architecture WG.) continued these developments in 2022. This document outlines the work packages (WP) of the RDF Roadmap and summarizes activities, developments, and achievements of the involved actors in the year 2022.

WP1: SPHN Dataset

The SPHN Dataset describes the semantic meaning of data shared within SPHN projects. The semantics are captured via concepts that follow ontological principles and are enriched with well recognized semantic standards. In 2022, the aim was to methodologically refine the SPHN Dataset according to best practices and to include further concepts and semantic standards that are/have been identified and used by SPHN driver projects and/or are foreseen to be needed by the National Data Streams (NDSs).

Deliverable 1: SPHN Dataset release (this is a yearly deliverable)

Deliverable	Responsibility	Timeline	Status	Details
D1.1	Semantic WG	June 2022	Done	The SPHN Dataset 2022.1 can be downloaded from DCC GitLab, the documentation of the change requests and new concepts can also be downloaded from DCC GitLab. Improvements and extensions to SPHN Dataset release 2022.1: - Broader use of semantic inheritance (concepts inherit composedOfs from their parent concept)

				- Extended reusability of elements by introducing general names and general descriptions for composedOfs - Enrichment with semantic standards (coded value sets where possible) - New concept: Quantity, Death Date, Diagnosis, Procedure, Lab Analyzer, Lab Test, Age, Reference Range, Laterality, Pharmaceutical Dose Form
D1.2	Semantic WG	March 2023	In progress	Concepts* for the 2023.1 release of the SPHN Dataset will include: - Gene, Protein, Transcript, Organism, Chromosome, Variant Descriptor, Genetic Variation, Single Nucleotide Variation, Genomic Position, Chromosomal Location, Reference, Variant Notation, Body Surface Area, Body Mass Index, Body Position, Cardiac Output, Cardiac Index, Access Device Presence, ECG Procedure, Electrocardiogram, Data File, Allergen, Health Care Provider Institute, Physiologic State **new concepts have been finalized in Dec 2022 and will be published together with the RDF schema and tools in Spring 2023*

WP 2: 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 representation using RDF. More information is available here.

The goal of this work package is to expand, improve and professionalize the SPHN RDF Schema.

Aim 1: Update the SPHN RDF Schema according to changes in the new SPHN Dataset.

Aim 2: Improve the naming of the unique identifiers (URI) of properties to make them better readable, less contextualized, and simplify the SPHN RDF Schema.
Aim 3: Include restrictions to add cardinality information for each property and encode all value constraints provided in the SPHN Dataset.
Aim 5: Improve and clarify the rule of inheritance for classes and properties.
Aim 6: Provide means to facilitate the migration from old to new versions of the SPHN RDF Schema.

Deliverable D1: New release of the SPHN RDF Schema (this is a yearly deliverable)
Deliverable D2: Migration path tool and file for comparing two versions of the SPHN RDF Schema

Deliverable	Responsibility	Timeline	Status	Details
D1.1	DCC	May 2022	Done	The SPHN RDF Schema 2022.1 can be downloaded from DCC GitLab or directly explored online on the BioMedIT page (generated thanks to the SPHN RDF Visualization tool, see WP 7). Major improvements addressed: - 20 new classes - Simplification of property identifiers - Cardinalities encoded as OWL Restrictions - Value restrictions encoded as OWL Restrictions - Inheritance rule: “all properties of a parent class are inherited by the child class. The children classes are not explicitly written as domain of that property”
D1.2	DCC	July 2022	Done	Bug fix release: 2022.2 version is accessible with a few fixes
D2	DCC	May 2022	Done	A migration path tool was developed, accessible in the DCC GitLab. The file comparing two SPHN RDF Schema versions (2021 and 2022) is accessible here.

WP 3: Dataset2rdf

Previously, we generated the SPHN RDF Schema using Protégé where concepts and properties were added to the SPHN RDF Schema manually. But this approach is not scalable, especially with

the increasing number of concepts that are requested to be added to the SPHN RDF Schema. To automate the process of generating the RDF Schema, this work package aims to build the Dataset2rdf tool which will be used to directly convert the SPHN Dataset to SPHN RDF Schema.
Aim 1: Build the Dataset2rdf tool to convert the SPHN Dataset to SPHN RDF Schema
Aim 2: Extend the Dataset2rdf tool to convert project-specific Dataset to project-specific RDF Schema
Aim 3: Prepare documentation and user guide

Deliverable D1: Create a prototype tool for converting the SPHN Dataset to RDF
Deliverable D2: Extend the tool to support cardinalities and complex OWL Restrictions
Deliverable D3: Finalize the tool for SPHN Dataset 2023.1 release

Deliverable	Responsibility	Timeline	Status	Details
D1	DCC	Aug 2022	Done	A prototype tool to transform SPHN Dataset to RDF was built: https://git.dcc.sib.swiss/sphn-semantic-framework/dataset2rdf
D2	DCC	Nov 2022	Done	The tool was extended to represent cardinalities on properties, and support valueset restrictions via owl:Restriction
D3	DCC	Jan 2023	Done	Finalize the Dataset2rdf tool for the pre-release of the SPHN Dataset 2023.1 release

Next steps (2023):
- Extend the tool for handling SPHN and project-specific schemas.
- Prepare documentation and user guide.

WP 4: Terminology Service

External standard terminologies are being used in the SPHN Semantic Interoperability 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. In 2021 the DCC developed the Terminology server (MINIO) and the Terminology download area (BioMedIT Portal).
Aim 1: Provide terminologies used in the RDF Schema to data providers and researchers in RDF

Deliverable D1: Maintain and provide new versions of CHOP, ATC, SNOMED CT, UCUM, ICD-10 and LOINC in RDF (Including older versions and new versions on a bi-yearly basis)

Deliverable D2: Include RDF versions of the Sequence Ontology (SO); the Genotype Ontology (GENO); and the Human Genome Organization (HGNC) into the Terminology service
Deliverable D3: Tooling to support the automatic download of terminologies from the Terminology server
Deliverable D4: Strategy for handling different versions and a prototype for versioning of terminologies
Deliverable D5: A user guide and script to support researchers to FAIRify their vocabulary

Deliverable	Responsibility	Timeline	Status	Details
D1.1	DCC	Jan 2022	Continuous	Yearly release of external terminologies - ICD-10 GM - ATC - CHOP
D1.2	DCC	Jan 2022 July 2022	Continuous	Biannual release of external terminologies - SNOMED CT - LOINC
D1.3	DCC	July 2022	Done	Updated the SNOMED CT pipeline with the ELK reasoner to generate an RDF version of SNOMED CT that explicitly includes all 'subClassOf' information that were missing in the schema. The 'snomed' prefix is also added into this new version of the RDF file.
D1.4	DCC	Dec 2022	Done	Updated the ICD-10 GM pipeline to include French labels and to remove special characters like *,! from the code. These updated versions will be released with the new terminology bundle used in the 2023 SPHN RDF Schema release in Spring 2023.
D2	DCC	Dec 2022	In progress	The DCC Terminology service was updated to include HGNC, SO and GENO (GitLab development branches). The terminologies will be released with the new terminology bundle used in the 2023 SPHN RDF Schema release in Spring 2023.
D3	RDF Support WG	April 2022	Done	Terminology Server Downloader script for automatic download of terminology bundle from the Terminology server https://git.dcc.sib.swiss/sphn-semantic-framework/terminology-server-downloader/
D4	DCC	Dec 2022	Done	A strategy for versioning of concepts from different versions of a terminology and providing the versioned concepts in RDF for downstream use. The

				proposal and demo implementation can be found here. A full implementation is planned for the new terminology bundle release in Spring 2023.
D5	DCC	Sep 2022	Done	A user guide to describe FAIRification of vocabulary is available here. "The External terminologies to RDF.ipynb" notebook provides a step-by-step example of transforming an exemplary vocabulary into RDF, and can be downloaded on the DCC GitLab.

Next steps (2023):

- Implementation of the versioning strategy for ICD-10-GM, CHOP and ATC.
- Release of the 2023 terminology bundle including the updates of D1.4; D2 and D4.

WP 5: SPHN Connector

The SPHN Connector is an automated solution to generate and execute de-identification, conversion, and validation pipelines on SPHN compliant data (SPHN core and project-specific). It can be installed on the premises of a data provider and can be seen as an interface (in Tables or structured documents). These interfaces hide the complexity of the steps to supply data to SPHN projects. It can be seen as a connector for data from a provider to SPHN and can replace the individual solutions of the data providers developed with a unified, consistent, and performant solution. New data providers get the benefits by installing the SPHN Connector to speed up their onboarding, by not needing to develop a RDF solution on their own, but using the same proven solution. For the SPHN Connector to be successful, different additional tooling is developed: A Load Testing Framework, User Testing Framework, and the SPHN Connector Integration Framework.

Deliverable D1: Define requirements, architecture, and align with University Hospitals
Deliverable D2: Setup the project, funding, staffing, and commitment
Deliverable D3: Development of SPHN Connector, Load Testing Framework, and Integration Framework

Deliverable	Responsibility	Timeline	Status	Details
D1	ITAC	May 2022	Done	Define requirements, architecture and align with hospitals
D2	ITAC	May 2022	Done	Setup project, funding, staffing, and commitment
D3	ITAC	Dec	Done	Successful execution of the SPHN Connector

	2022	Projects
		- SPHN Connector: https://git.dcc.sib.swiss/hospfair/sphn-connector
		  - 4 of 5 hospitals installed the release version 1.0.0
		  - KISPI successfully installed a development version
		- Load Testing Framework : https://git.dcc.sib.swiss/hospfair/loadtesting-sphnconnector
		  - Helped to reduce the runtime dramatically (in version 1.0.0 of the SPHN Connector the conversion takes 1/17th of the time than in 0.5.0)
		- Integration Framework: https://git.dcc.sib.swiss/hospfair/sphn-connector-integration
		  - Automatic setup/update script which may be used by different hospitals
		  - Blueprint solution for integration data using the tabular interface (used at least by USZ)
		- User Testing: https://git.dcc.sib.swiss/hospfair/sphn-connector-user-testing
		  - Automatic testing of outbound facing API

WP 6: 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 1: Provide tools to help users deliver 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 tool to create SHACL rules from a project-specific RDF Schema
Deliverable D3: A set of SPARQL queries to provide summary statistics on the delivered data
Deliverable D4: A tool for automatic checks, accessible to data providers and researchers

Deliver-able	Respon-sibility	Time-line	Status	Details
D1	DCC	May 2022	Done	For data quality and control, the SPHN provides a set of SHACL rules to validate the compliance of the RDF data produced. The following resource is available: - SPHN SHACLs on GitLab
D2	DCC	July 2022	Done	The SHACLer is a Python tool that extracts SHACL rules from an SPHN-compliant input ontology for facilitating data validation. The following resources are available: - SHACLer on GitLab - Read the Docs
D3	DCC	May 2022	Done	For data quality assurance, the SPHN provides a set of SPARQL queries to generate basic summary statistics of the RDF data produced. The following resource is available: - SPHN Statistical SPARQLs on GitLab *To provide more comprehensive statistics the SPARQLer Tool (WP9 SPARQLer) was developed.*
D4	HUG and Research Support WG	July 2022	Done	The Java-based RDF Quality Check tool facilitates the validation process of SPHN RDF data at the data provider level. Based on the SHACL shapes generated by the SHACLer and statistical queries in SPARQL, it generates a human-friendly report with information about data conformance to the schema and some basic statistics. The following resources are available: - QC Tool on GitLab - Read the Docs

WP 7: GraphDB Data Loader

When data or external terminologies come to the BioMedIT nodes, data needs to be loaded into the triple store (GraphDB at BioMedIT). In this work package, the goal is to develop a script to automate this process.

Aim 1: Automatize the steps of data loading into the GraphDB as a named graph to reduce CLI interaction for users.

Deliverable D1: A script for a GraphDB data loader usable in the three BioMedIT nodes

Deliverable	Responsibility	Timeline	Status	Details
D1	RDF Support WG	April 2022	Done	The GraphDB data loader, a bash script, can be used to trigger the import of triples from "server files" using the GraphDB API. The loader allows for quick loading of project data into specified named graphs. The following resources are available: - Data Loader script on GitLab

WP 8: SPHN RDF Visualization Tool

With all developments made in the SPHN Semantic Interoperability Framework, it is crucial to have detailed documentation that gathers all knowledge and guides the projects in their development. In this work package, the goal is to provide a tangible representation and visualization of the SPHN RDF Schema.

Aim 1: Provide a tool, which generates an easily readable representation/visualization of the SPHN and project-specific RDF Schema in HTML format.

Deliverable D1: A tool to create a html document that visualizes the SPHN RDF Schema
Deliverable D2: Add support for new features (cardinalities, restrictions, etc.)
Deliverable D3: Make the tool available for the visualization of project-specific ontologies

Deliverable	Responsibility	Timeline	Status	Details
D1	DCC	March 2022	Done	The SPHN Schema Visualization Tool generates a human-readable HTML document describing the project's RDF Schema directly from the schema. It is based on pyLODE and covers detailed information about the classes, objects, datatype and annotation

				properties, and the named individuals. The following resources are available: - Schema Visualization Tool on GitLab - Example visualization of the SPHN RDF Schema - User Guide
D2	DCC	July 2022	Done	Added support for, - Cardinalities on properties - owl:Restriction on classes - skos:scopeNote on classes
D3	DCC	Dec 2022	Done	Support for project-specific RDF Schema was developed. A short user guide is provided here. The first example published is the SwissBioRef RDF Schema.

WP 9: SPARQLer

Manually writing and adapting SPARQL queries for Quality Control purposes is a time consuming process, therefore this work package aims to develop a tool, which generates the queries automatically from the SPHN RDF Schema.

Aim 1: Provide a tool, which generates SPARQL queries for statistical analysis that can be executed against a SPARQL endpoint by data managers and hospitals to retrieve the content of the RDF data in a tabular format.

Deliverable D1: A tool, which generates SPARQL queries for concept-flattening (list of resources defined for a concept together with the direct property values)
Deliverable D2: Add SPARQL queries with simple statistics (additional features).
Deliverable D3: Make the tool available for project-specific ontologies.

Deliverable	Responsibility	Timeline	Status	Details
D1	DCC	Dec 2022	Done	The SPHN SPARQL Generator (SPARQLer) is a Python tool that accepts as input the SPHN RDF Schema (Turtle format) and generates a series of SPARQL queries in a standard RDF/OWL format based on the concepts present in the schema. The SPARQL generation functionality has been integrated into the SPHN Framework Schema

				Visualization Tool. Thus, the formerly known SPARQLer has now been retired in favor of the new implementation.
D2	DCC	Dec 2022	Done	Additional feature - Counting instances per concept and predicates - Minimum and maximum values per predicate - List and count of all used codes for the hasCode property
D3	DCC	Dec 2022	Done	Support for project-specific RDF Schemas

Next steps (2023):

- Provide documentation and a description of the queries in the User Guide.

WP 10: Technical Documentation and User Guide

The SPHN Interoperability Framework is a complete infrastructure that needs to be documented to provide the necessary information to users for understanding the different tools and services, and therefore facilitate their use.

The goal of this WP is to provide, in a single space, a thorough and complete documentation of the different elements of the framework.

Aim 1: Document the content and developments of the SPHN RDF Schema and related tools and services for record, information, and guidance purposes.

Deliverable D1: Provide background documentation about the different semantic web technologies used in SPHN.
Deliverable D2: Provide detailed information and documentation of the SPHN Interoperability Framework components.
Deliverable D3: Provide detailed user guide documentation on how to accomplish certain tasks in the context of SPHN, targeting both data providers and data users.
Deliverable D4: Generate BioMedIT node internal documents to facilitate the understanding process to handle RDF-related tasks.

Deliverable	Responsibility	Timeline	Status	Details
D1	Research Support	March 2022	Done	The ReadtheDocs documentation was extended for the following background sections

	WG			- Semantic Web - RDF - SPARQL - SHACL
D2	DCC	Oct 2022	Done	The ReadtheDocs documentation provides full documentation of the SPHN Interoperability Framework components including the tools developed in 2022 - SPHN Dataset including updates and new features in the 2022.1 release (e.g. inheritance) - SPHN RDF Schema including updates and new features in the 2022.1 release (e.g. cardinalities, OWL restrictions...) - SHACLer - SPARQLer - Quality Assurance Framework - Terminology Service
D3	DCC	Oct 2022	Done	The ReadtheDocs was extended with the following sections: - SHACL constraints in the SHACLer - Concept Flattening in the SPARQLer - Download external terminologies from the Terminology Service - External Terminologies in RDF for projects
D4	Research Support WG	March 2022	Done	Node onboarding guideline document is accessible here. They contain information about: - The use of the Terminology Service from the BioMedIT perspective

WP 11: Mock Data Generator

There is a growing need for large amounts of mock data for testing various tools and implementations developed in the context of SPHN. The mock data can be used,

- to test drive graph tools with large amounts of realistic data
- for projects to develop analysis pipelines based on realistic data without restrictions imposed by sensitive data

Aim 1: To develop a tool to generate mock data from the SPHN RDF Schema and/or project-specific schemas.
Deliverable D1: Demonstrate technical feasibility and build a prototype

Deliverable D2: Tool for generation of mock instance data from the SPHN RDF Schema

Deliver-able	Respon-sibility	Time-line	Status	Details
D1	Research Support WG	Nov 2022	Done	
D2	Research Support WG	Dec 2022	Done	- Parse the complete JSON schema of the SPHN ontology and create generic mock data - Enable schema transformations and implement custom Faker providers for generation of more realistic mock data - Expose execution options and enable loading of configuration details from a JSON file - https://git.dcc.sib.swiss/hospfair/mockdatageneration

Next steps (2023):

- Improvement of the tool (e.g. add new sampling methods, integration with the SPHN Connector).
- Extension of the generator for project-specific instance data from project-specific RDF Schema
- Prepare documentation on how to generate mock data using the tool

WP 12: Outreach and Training

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

Aim 1: Spread the knowledge and understanding of the SPHN RDF Schema and its usage in clinical settings

Deliverable D1: SPHN Trainings
Deliverable D2: Participation in appropriate meetings and organization of workshops
Deliverable D3: Scientific publications
Deliverable D4: Other publications e.g. Website and Fact sheets

Deliverable	Responsibility	Timeline	Details
D1.1	DCC	Aug 2022	Training: FAIR implementation in health data
D1.2	DCC	Sep 2022	Training: How to design a Concept
D1.3	DCC	Continuous	All training material is published on the TeSS catalogue and the SIB training collection to make it FAIR and to reach a wider audience.
D2	DCC	Continuous	The DIT gave a talk at the following conferences/events - Databits, Basel, CH - Workshop on safe access to sensitive data, Force, CH - SIB Resource Day, Bern, CH - BioDataWorld, Basel, CH - LOINC conference, Annecy, F - SNOMED CT Expo, Lisbon, PT - Semantics@Roche, Basel, CH - Ontotext Forum, Basel, CH - SIB RDF Day, Lausanne, CH - Life science cluster, HKBB, Basel, CH - The Data Warehousing Institute (TDWI), online, Basel, CH - SWAT4HCLS, Leiden, NL - MILA Seminar, Greifswald, DE - Life science cluster Basel, CH - Medical informatics initiative, data interoperability working group, DE The DIT presented a poster at the following conferences/events - Semantics@Roche, Basel, CH - SIB days, Biel, CH - Second Joint Personalized Health Day, Bern, CH The DIT was involved in teaching in the following courses - CAS Modern concepts in clinical research, Real World Data, ETH, CH - CAS Ethics and Legal in Clinical Trials, EPCM, UniBasel, CH

D3	DCC	Con- tinu- ous	Scientific publications: Touré *et al*. FAIRification of health-related data using semantic web technologies in the Swiss Personalized Health Network (under review)

D4	DCC	Con-
		tinu-
		ous
			Fact sheet:
			- 2022 Semantic Strategy
			Website:
			- Update of the sphn.ch website
			- New Biomedit.ch website
			  - SPHN Semantic Web Stack
			  - Listing of the "Scientific Tools and Services"
			News:
			- SPHN Quality Control Tools
			- SPHN Dataset & RDF schema new release