 doi:10.20944/preprints202109.0505.v1

The SPHN Ecosystem Towards FAIR Data

Sabine Österle[1][0000-0003-3248-7899], Vasundra Touré[1][0000-0003-4639-4431] and Katrin Crameri[1][0000-0003-3656-3457]

[1]Personalized Health Informatics Group,
SIB Swiss Institute of Bioinformatics,
Basel, Switzerland
dcc@sib.swiss

Abstract. Health-related data originating from diverse sources are commonly stored in manifold databases and formats, making it difficult to find, access and gather data for research purposes. In addition, so-called secondary use scenarios for health data are usually hindered by local data codes, missing dictionaries and the lack of metadata and context descriptions. Following the FAIR principles (Findable, Accessible, Interoperable and Reusable), we developed a decentralized infrastructure to overcome these hurdles and enable collaborative research by making the meaning of health-related data understandable to both, humans and machines. This infrastructure is currently being implemented in the realm of the Swiss Personalized Health Network (SPHN), a research infrastructure initiative for enabling the use and exchange of health-related data for research in Switzerland. The SPHN ecosystem for FAIR data consists of the SPHN Dataset (semantic definitions), the SPHN RDF Schema (linkage and transport of the semantics in a machine-readable format), a project RDF template, extensive guidelines and conventions on how to generate SPHN RDF schema, a Terminology Service (converter of clinical terminologies in RDF), and a Quality Assurance Framework (automated data validation with SHACLs and SPARQLs). The SPHN ecosystem has been built in a way that it can easily be adapted and extended by any SPHN project to fit individual needs. By providing such a national ecosystem, SPHN supports researchers in generating, processing and sharing FAIR data.

Keywords: Semantics, standards, clinical research infrastructure, terminology, graph data, data-driven medicine

1 Starting position

To optimize the use of health-related data for Personalized Health Research (PHR), both transdisciplinary scientific research and a broad range of infrastructural efforts are required: Established structures and procedures are needed that enable rapid and wide-ranging, controlled access to fit-for-purpose, interoperable and standardized health(care) data, which are able to interact and be linked to state-of-the-art IT infrastructures, research platforms and biobanks, while meeting data protection, privacy and information security requirements [1]. In view of the federal structures in Switzerland – with 26 cantons responsible for the provision of healthcare services – but also due to

 doi:10.20944/preprints202109.0505.v1

data protection and data security arguments, the Swiss Personalized Health Network (SPHN) has opted for a decentralized approach in which data remain at their source and are shared and combined solely in a project-specific manner [2]. The various data sources that can potentially be linked for PHR range from health care facilities providing routine clinical data, to laboratory facilities providing bioanalytical and -omic data, to patient-oriented clinical research registries and cohort studies, to citizen-controlled health data. The heterogeneity of the data types and data sources poses a variety of challenges, especially with regard to formats and standards, but also concerning the level of granularity when it comes to the description of the data. Local data codes, missing dictionaries and the lack of metadata and contextual descriptions make it tremendously difficult to link data derived from different sources and bring them together for PHR purposes.

This federated approach with the above described characteristics calls for a comprehensive framework for data providers to generate and deliver data in an SPHN-compliant and FAIR (Findable, Accessible, Interoperable and Reusable) way [3]. Local production or collection and preparation of data must ensure that the data is understandable to both humans and machines. Moreover, since individual providers only have insight into their own databases, harmonized and detailed guidelines as well as a solid data quality framework are necessary to ensure that the data arriving at the researcher is compatible and interoperable with that of other providers. The SPHN ecosystem towards FAIR data addresses the various requirements described above and is intended to support both data providers and researchers as a service infrastructure in their PHR endeavors.

2 Design and Components of the Ecosystem

The SPHN Data Coordination Center (DCC) managed by the Personalized Health Informatics Group of the SIB Swiss Institute of Bioinformatics is responsible for the design and implementation of the Ecosystem for FAIR data in SPHN. The core component of the ecosystem (see Figure 1 and Table 1) is the SPHN Resource Description Framework (RDF) schema, which incorporates the semantic definitions of the SPHN Dataset [4] in a machine processable way, following the W3C standard [5]. This RDF schema integrates national and international standards such as the International Statistical Classification of Diseases and Related Health Problems, 10th revision, German modification [6] (ICD-10-GM), the Swiss classification for procedures [7] ("Schweizerische Operationsklassifikation" CHOP), the Anatomical Therapeutic Chemical Classification System [8] (ATC), the Systematized Nomenclature of Medicine – Clinical Terms [9] (SNOMED CT), the Logical Observation Identifiers Names and Codes [10] (LOINC) and, the Unified Code for Units of Measure [11] (UCUM). All these standards can be used to express the data – be it through value set binding and/or to precisely code data. In addition, SNOMED CT and LOINC are used as controlled vocabularies to provide a meaning binding to some concepts defined in the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 September 2021 doi:10.20944/preprints202109.0505.v1

SPHN Dataset. External terminologies are provided via the terminology service in an SPHN-compliant RDF format. The terminology service is a built-in tool that converts different versions of the standard terminologies and classifications into RDF and provides them to the data providers as well as data users in a secure and controlled environment.

To allow a project to extend the SPHN RDF schema with its own concepts, the DCC provides an RDF template that contains the basic metadata and imports needed for building an SPHN-compliant RDF schema. A user guide describes how classes, properties, ranges, domains, etc. can be extended in an SPHN compliant way. The project RDF schema and the external terminology files are used in the ETL (Extract, Transform, Load) process, where data is extracted from the source, e.g. the clinical data warehouse of a hospital, coded in the SPHN recommended standards and transformed into RDF. The generated data are validated against a set of global SPHN Shapes Constraint Language (SHACL) rules, or against a project specific SHACL set created by using the SHACLer tool, both provided by the DCC. Valid RDF data is then sent from different data providers to the place of analysis, where it can be combined, link, and analyzed. Researchers are free to either use the RDF directly as input for their R or Python scripts, transform it into a data model of choice or convert it into a flat file. To facilitate the conversion, SPHN provides a set of SPARQL (SPARQL Protocol and RDF Query Language) queries to extract each SPHN concept with their metadata in a flat format. To allow automatic generation of such SPARQLs, the SPARQLer tool is provided, which generates these SPARQL queries from any SPHN-compliant RDF schema.

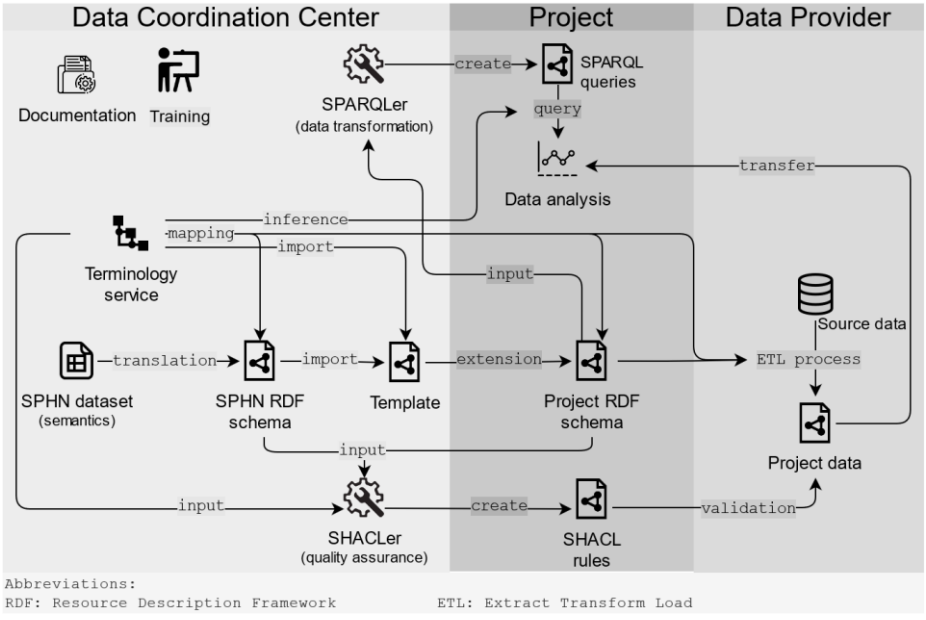

Fig. 1. SPHN ecosystem for data interoperability.

 doi:10.20944/preprints202109.0505.v1

Table 1. Summarizes the main components of the SPHN ecosystem.

Component	Description
SPHN Dataset	The SPHN dataset contains the semantic description of the SPHN concepts (definitions, properties, and recommended standards).
SPHN RDF Schema	The SPHN RDF schema is the technical exchange format for SPHN data in a FAIR (Findable, Accessible, Interoperable, and Reusable) format. In addition, it provides means to link other existing standard ontologies.
DCC Terminology Service	The DCC terminology service provides SPHN compatible, machine-readable versions of national (CHOP or ICD-10 GM) and international (SNOMED CT, LOINC, ATC, UCUM) terminologies and classifications in RDF-compliant formats.
Template	The template helps projects to develop their project-specific RDF schema. This template provides pre-filled basic metadata annotations and imports the SPHN RDF schema with its external terminologies and other general RDF libraries.
Quality Assurance Framework	The SPHN data quality assurance framework consists of a set of SHACL rules and statistical SPARQL queries to validate the compliance of the RDF data produced to the SPHN schema.
SHACLer	The SHACLer tool allows the adaption of this quality assurance framework in a project specific manner. It is a tool which extracts SHACL rules from an SPHN-compliant project-specific RDF schema.
SPARQLer	The SPARQLer tool automatically extracts one SPARQL queries per concept of an SPHN-compliant RDF schema.

3 Evaluation

The SPHN semantic interoperability strategy and the corresponding ecosystem helps researcher in Switzerland to address the FAIR data principles. While F1, F2, F3, A1, I1, I2, I3 are covered in the SPHN strategy (see Table 2), the duty to fulfill F4 and R1 are in the responsibility of a project (e.g. choice of a license or of the repository).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 September 2021 doi:10.20944/preprints202109.0505.v1

Table 2. List of FAIR criteria addressed in the SPHN semantic interoperability strategy.

FAIR criteria	Addresses in the SPHN strategy
F1. (meta)data are assigned a globally unique and persistent identifier.	In the SPHN RDF schema, data are assigned to a Unique Resource Identifier (URI) with the following namespace https://biomedit.ch/rdf/sphn-resource/.
F2. data are described with rich metadata (defined by R1 below)	Administrative metadata is provided in the SPHN RDF file header (e.g. SPHN RDF schema version used, the extraction date of the data and the identifier of the data provider). Descriptive metadata for all data elements, including which data elements to include, their definition, standards and/or value set to be used are included in the SPHN RDF schema. Additionally, properties of a concept provide additional (meta)data of a data element such as the "method of a measurement".
F3. Metadata clearly and explicitly include the identifier of the data	Metadata mentioned in F2 is part of the schema and therefore linked with the data. *A project can include references to additional metadata in the RDF schema, this needs to be addressed on the individual project level.*
F4. (meta)data are registered or indexed in a searchable resource	*This FAIR criterion needs to be addressed on the individual project level.*
A1. (meta)data are retrievable by their identifier using a standardized communications protocol	Since data is represented using the RDF standard, the W3C query language SPARQL can be used to query the data.
A2. metadata are accessible, even when the data are no longer available	*This FAIR criterion needs to be addressed on the individual project level.*
I1. (meta)data use a formal, accessible, shared and broadly applicable language	SPHN is using the W3C RDF standard as language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles	SPHN is currently using LOINC and SNOMED CT as controlled vocabulary, the system is however flexible to be expanded to other controlled vocabularies.
I3. (meta)data include qualified references to other (meta)data	SPHN requires the reference to existing URIs of the external terminologies provided, when possible, for the annotation of meta(data).

R1. meta(data) are richly described with a plurality of accurate and relevant attributes	*This FAIR criterion needs to be addressed on the individual project level.*

4 Challenges

The development of such a data interoperability ecosystem has presented us with several challenges:
The choice of RDF as an exchange format comes with a cost, as clinical data is generally not encoded in RDF-compliant formats. The data transformation imposes an additional burden on data providers to map raw clinical data to the format requested by the SPHN framework. To support data providers, the DCC has developed comprehensive documentation and guidelines, explaining the strategy and how to represent data following the SPHN framework. For instance, the specification of conventions for the definition of URIs of common resources helps to improve data interoperability between equivalent clinical data coming from different data providers. Furthermore, training events and hackathons have been organized to help facilitate the learning and understanding of semantic web standards to the SPHN community and the interested audience outside our network. A fundamental problem in the clinical environment is that a substantial part of the information is usually not available in a structured form. Besides, the part that is available is largely standardized for billing purposes rather than with the aim of reflecting patient reality. In addition, the use of internationally recognized standard terminologies in healthcare is only emerging. Therefore, data providers need to put a lot of efforts in the structuring of data and the mapping of local codes to standard terminologies, in order to provide understandable, valuable and fit-for-purpose data for researchers. The mapping task is not trivial and requires domain knowledge as well as understanding of the used standard terminology terms to make sure that the semantic meaning is correctly translated from the local codes to the applied standards.
Finally, data validation is a critical step to ensure that the data generated is usable by researchers. The SHACL rules validate the compliance of the data in respect to the RDF schema and additional features such as cardinalities, but in combination with the external terminology they also validate that only valid codes are used to encode the data. Although this might sound trivial, such a control step brings an immense advantage over other systems where codes are represented as strings and not as reference links to external terminologies. The SPHN SHACLer tool allows the projects to easily expand these validations to their schema extensions.

5 Conclusion

The SPHN ecosystem for FAIR data supplies data providers and researchers with the tools and services to make data (more) FAIR. The full adaption of the FAIR principles in the medical domain in Switzerland is however still a long way. SPHN is therefore not only building an infrastructure framework but is also investing in research support, education and training to foster the understanding and implementation of these new technologies.

Acknowledgments. The authors would like to acknowledge the SPHN Working groups: Clinical Data Semantic Interoperability, chaired by Christian Lovis and the RDF Task force of the Hospital IT chaired by Katie Kalt as well as all representatives of the Swiss University Hospitals (HUG, CHUV, USB, USZ and Inselspital) and the SIB Swiss Institute of Bioinformatics for their contributions. Our special thanks goes to Philip Krauss from Trivadis part of Accenture for this contribution to the design and implementation of the ecosystem.

Correspondence. Dr. Sabine Österle (sabine.oesterle@sib.swiss) Personalized Health Informatics Group, SIB Swiss Institute of Bioinformatics, Basel, Switzerland

References

[1] SPHN, "Swiss Personalized Health Network. Report from the National Steering Board 2016 – 2019." doi: doi.org/10.5281/zenodo.4044123.

[2] A. K. Lawrence, L. Selter, and U. Frey, "SPHN - The Swiss personalized health network initiative," *Stud. Health Technol. Inform.*, vol. 270, pp. 1156–1160, 2020, doi: 10.3233/SHTI200344.

[3] M. D. Wilkinson *et al.*, "Comment: The FAIR Guiding Principles for scientific data management and stewardship," *Sci. Data*, vol. 3, pp. 1–9, 2016, doi: 10.1038/sdata.2016.18.

[4] C. Gaudet-Blavignac, J. L. Raisaro, V. Touré, S. Österle, K. Crameri, and C. Lovis, "A national, semantic-driven, three-pillar strategy to enable health data secondary usage interoperability for research within the swiss personalized health network: Methodological study," *JMIR Med. Informatics*, vol. 9, no. 6, pp. 1–10, 2021, doi: 10.2196/27591.

[5] "WC3 RDF." https://www.w3.org/RDF/.

[6] "ICD-10 GM." https://www.dimdi.de/dynamic/en/classifications/icd/icd-10-gm.

[7] Bundesamt für Statistik, *Medizinisches Kodierungshandbuch. Der offizielle Leitfaden der Kodierrichtlinien in der Schweiz*. 2020.

[8] "ATC." https://www.whocc.no.

 doi:10.20944/preprints202109.0505.v1

[9] "SNOMED CT." https://www.snomed.org/.
[10] C. J. McDonald *et al.*, "LOINC, a universal standard for identifying laboratory observations: A 5-year update," *Clin. Chem.*, vol. 49, no. 4, pp. 624–633, 2003, doi: 10.1373/49.4.624.
[11] "UCUM." http://unitsofmeasure.org.