Fact sheet
The SPHN semantic framework – Pillar 2: Data transport and storage format

1 Introduction

Semantic interoperability ensures that information is consistently interpretable by both machines and humans - across projects, systems, countries and over time. To enable the use of health data from clinical routine and other sources for research, SPHN has developed a semantic interoperability strategy [1]. The framework for this strategy is built on three pillars: Pillar 1: Semantic representation [2]; Pillar 2: Data transport and storage, and Pillar 3: Use cases.

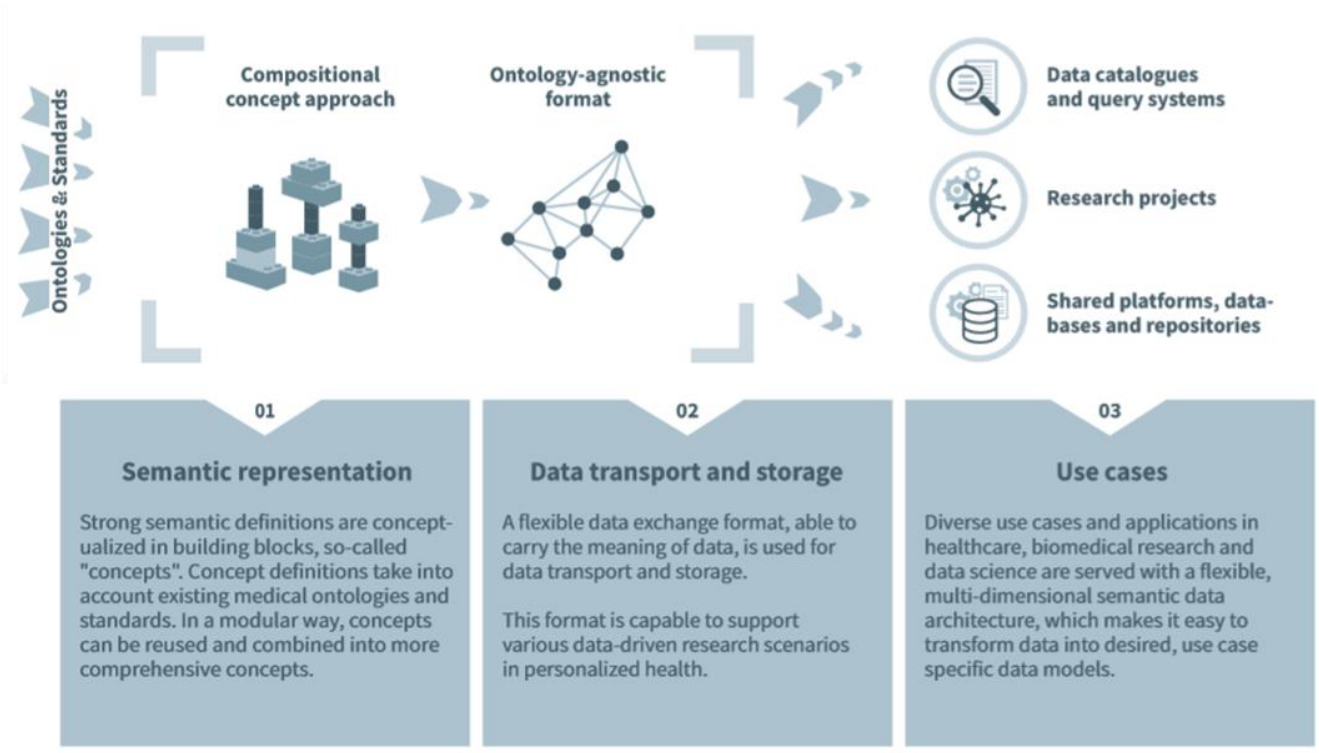

Figure 1: Three pillars Swiss personalized health network (SPHN) semantic framework

In this fact sheet we focus on Pillar 2, a flexible transport and storage format to ensure interoperability with and for various use cases and applications in healthcare, biomedical research and data science. This format needs to be able to carry data and their meaning, and capable to support various data-driven research scenarios in Personalized Health. The format needs to allow referencing of various biomedical ontologies and integration or interlinking of multiple data types.

Swiss Institute of Bioinformatics (SIB)
Personalized Health Informatics Group (PHI)
Elisabethenstrasse 43
CH-4063 Basel, Switzerland

2 Evaluation process

The Task Force “Data Exchange Format” of the Data coordination center (DCC) Working Group HospIT evaluated several solutions for SPHN, such as common data models (e.g., i2b2, and OMOP), data exchange standards/formats (e.g., FHIR), semantic-based descriptive framework (e.g., RDF) and flat files. The following evaluation criteria were used [3]: Coverage of the SPHN dataset; Extensibility; Scalability; Understandability; Adoption cost for researchers; Adoption cost for the hospitals; Supported value sets; Worldwide adoption; Available tools and Stability. Within the process of evaluation, it became clear that the use of existing common data models would have introduced several limitations. First, data models are rigid representations of knowledge (made of tables and relationships between them) that are proper to specific purposes. There is no one-size-fit-all data model that could serve all SPHN use cases. If multiple data models had to be used within SPHN, the cost of mapping from local data sources to each model would be unsustainable for hospitals on the long run. Additionally, the use of a single data model would not allow to fully fulfill the project needs, leading either to a loss of information or to temporary ad hoc individual solutions. Data models, however are hard to extend without losing compatibility with existing mapping tools or analytics built on top of them. As a result, the Task Force recommended to adopt a more flexible and extensible solution for exchanging data and meta-data within SPHN such a descriptive language able to formally represent any current and future SPHN concept without information loss and the need of convoluted mappings. In particular, the Task Force recommends the use of the Resource Description Framework (RDF), a universal solution to model information that is described in a series of semantic concepts such as the SPHN Dataset (Pillar 1). With RDF, SPHN concepts and their instances (i.e., the data) can easily be mapped from/to other data representations or merged with other RDF data sets without losing their semantic. A two-day hackathon in November 2019 resulted in the successful integration of mock data form all five hospitals in RDF. At the moment (spring 2020) several Driver projects are running pilot projects to further test RDF as exchange format.

3 Background information on RDF

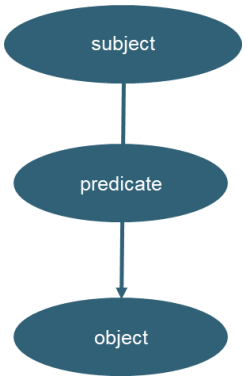

Figure 2: RDF triple

RDF intrinsically models information into a labeled, directed multi-graph where nodes and edges are identified by Uniform Resource Identifiers (URIs). The basic entity in the RDF graph is the “triple” (subject–predicate–object). An RDF graph is made of several triples. One popular and readable syntax and file format used to express data in RDF is “Terse RDF Triple Language” (Turtle). Several converters exist to convert RDF data into common file formats such as XML or JSON. RDF data can be queried using a standardized graph query language called SPARQL that is similar to SQL. One feature of SPARQL is that it allows federated queries across different data sources and therefore allows the easy integration of different data sources and data types.

Swiss Institute of Bioinformatics (SIB)
Personalized Health Informatics Group (PHI)
Elisabethenstrasse 43
CH-4063 Basel, Switzerland

Examples for international projects, which use RDF:

- UniProt
- Yosemite Project (US)
- GOFAIR
- Data.gov Centre for Disease control
- EHR4CR
- European Joint programme on rare diseases
- Wikidata, DBPedia

4 How is it implemented?

SPHN Concepts in RDF

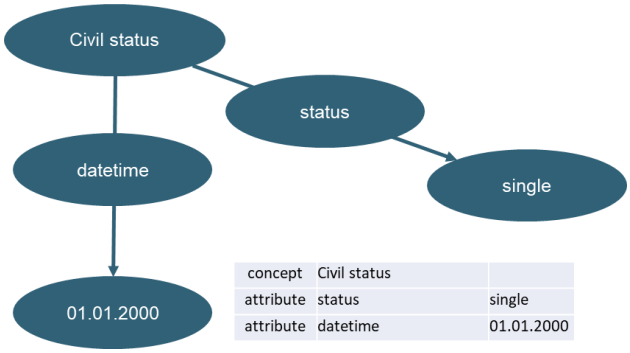

Figure 3: Semantic triple model (subject, predicate and object) for the concept Civil status

The relations between concepts, attributes and the data are expressed in triples made of a “subject” a “predicate” and an “object”. Informally, an RDF triple says that the Civil status (subject) has datetime (predicate) “01.01.2000” (object) and a status (predicated) which is “single” (object). Since RDF does not depend on a specific semantic standard, it allows to use different ontologies, and value sets as defined in the SPHN Dataset.

Workflow of RDF file generation

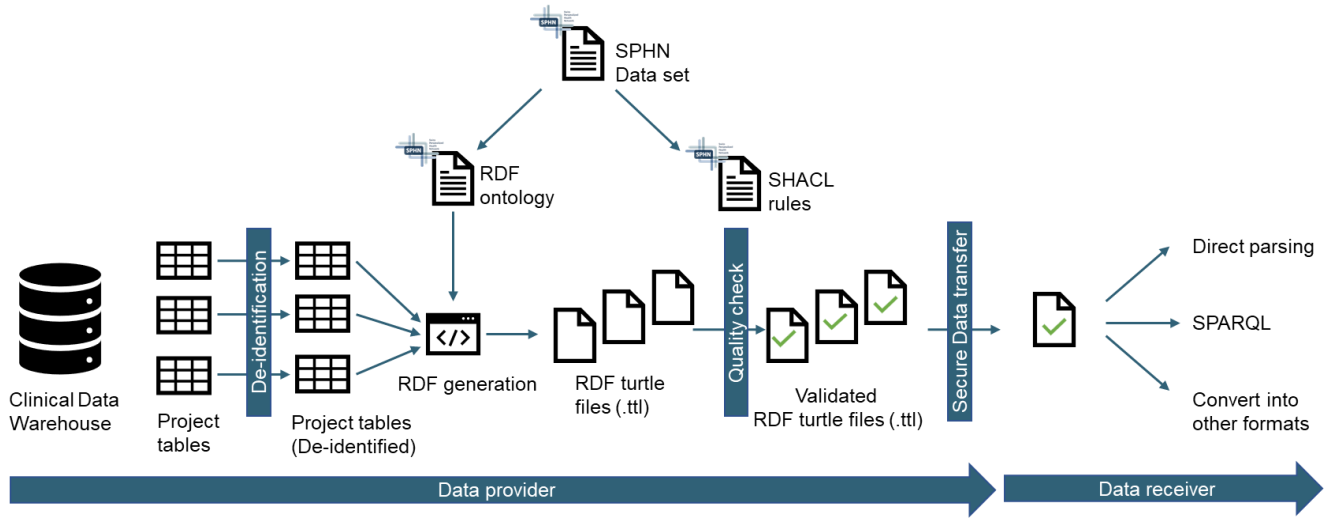

Figure 4: Workflow of RDF file generation by the Data Provider and usage by the Data Recipient

Swiss Institute of Bioinformatics (SIB)
Personalized Health Informatics Group (PHI)
Elisabethenstrasse 43
CH-4063 Basel, Switzerland

From the SPHN Dataset [2], the RDF Ontology (formal representation in RDF of the SPHN concepts and their relationships) and the SHACL rules (specifications for validation of RDF data) are created centrally according to the ruleset [4] developed by the HospIT Task Force. Both are shared with the various Data Providers. Based on the RDF ontology, Data Providers extract and de-identify the data for a specific project, and convert them into the correspondent RDF representation (turtle files). These RDF turtle files are validated with the SHACL rules to ensure that the generated RDF matches the definitions in the RDF Ontology. Afterwards, the validated turtle files are encrypted and sent to the Data Recipient through the BioMedIT infrastructure. After decryption, the Data Recipient can now either directly parse the RDF turtle file (for example with Python) and extract the needed information, or they can load the turtle files into an RDF data store (this could be a simple directory that stores all the turtle files) or a triple store for increased performance and run SPARQL queries on the RDF triples to convert the data into another format e.g., flat files in CSV or relational tables, to use them in Excel, or other analytical tools.

5 Steps from prototype to routine operation

Currently the SPHN RDF Ontology is derived from the SPHN dataset and needs to be updated after every change in the SPHN dataset. As a rapid solution during the evaluation process, the Task Force has developed a set of scripts for MS Excel that generate the RDF ontology in the turtle format from the SPHN Dataset, itself stored in MS Excel. Yet, such a procedure (script) is only a temporary solution at the prototype scale. A more sustainable solution based on more robust tools for ontology management has to be evaluated and adopted, instead.

The SPHN National Steering Board (NSB) endorses the use of RDF as one of the formats of choice for the data transport and storage in all SPHN projects, and recommends to adopt RDF as one of the SPHN approved data standards.

For the DCC, the introduction of RDF would entail the following activities and tasks:

- Maintain the Excel script and later an alternative technical solution that will enable the generation of the RDF Ontology in a user-friendly and easy manageable way;
- Share the RDF Ontology with the Data Providers and Data Recipients via one joint source of truth;
- Create and maintain shared resources (e.g., centrally generated unique resource identifiers, so called URIs, for common value sets and for which no RDF representation is already available)[1];
- Set up a process for change requests in the RDF Ontology and specifications (in alignment with the change request for Semantic concepts).

[1] *According to the advice during the SPHN workshop with international experts, reused sources (such as LOINC) should preferably be provided and maintained by the ontology provider itself. LOINC for example refers to the FHIR terminology server in order to provide such a repository (in XML, JSON).*

Swiss Institute of Bioinformatics (SIB)
Personalized Health Informatics Group (PHI)
Elisabethenstrasse 43
CH-4063 Basel, Switzerland

Additionally, services need to be provided to the Researchers and the Data Providers:

- Provide projects with instructions, trainings and tools on how to create and extend their own concepts to the official RDF Ontology or add new relations;
- Set up a process for projects to create their own concepts and to align and integrate them into the SPHN Data set
- Train and support researchers to formulate SPARQL queries to explore their dataset and select appropriated tools;
- Train and support Researchers and Data Providers with solutions to process RDF or to convert RDF into other formats such as CSV or JSON (and vice versa).

6 References

- [1] SPHN data set
- [2] Semantic strategy 2019
- [3] Task force report “Data Exchange Format”
- [4] RDF specifications and user guide

Swiss Institute of Bioinformatics (SIB)
Personalized Health Informatics Group (PHI)
Elisabethenstrasse 43
CH-4063 Basel, Switzerland