From clinical routine data to FAIR research data

Vasundra Touré[1*], Kristin Gnodkte[1*], Jan Armida[1*], Sabine Österle[1#]

supported by the other members of the PHI Group and collaborators (in alphabetical order): Owen Appleton[1], Jascha Buchhorn[2], Katrin Crameri[1], Patricia Fernandez Pinilla[1], Martin Fox[1], Simone Guzzi[1], Petar Horki[1], Shubham Kapoor[1], Philip Krauss[2], Julia Maurer[1], Michael Müller-Breckenridge[1], Christian Ribeaud[1], members of the SPHN Semantic working group, the SPHN Data exchange task force and the BioMedIT Research support working group.

[1]SIB Swiss Institute of Bioinformatics, Personalized Health Informatics Group, [2]Trivadis—Part of Accenture, *presenting authors, [#]corresponding author sabine.oesterle@sib.swiss

Personalized health research requires large amounts of structured and well standardized clinical (routine) data. Combining these with other health related data (e.g. omics or molecular), enables research on innovative diagnostic and therapeutic approaches. The Swiss Personalized Health Network (SPHN) interoperability framework takes the challenge of "FAIRifying" data by linking different types of data, and make its meaning understandable to both, humans and machines. Representing data as linked data with existing and defined standards in SPHN allows researchers to easily combine subsets of data from different sources as well as leverage the knowledge of the integrated terminologies within their research projects. The SPHN interoperability framework contains all components and tools needed to transform clinical (routine) and other health related data into high-quality FAIR research data.

SPHN semantic concepts

SPHN concepts are defined as generalizable and unambiguous building blocks, which can be used in different contexts to understand a specific data meaning. Concepts can be combined to compose larger and more complex concepts, which again can be combined to more complex compositions. The meaning of a concept is expressed using existing semantic standards (controlled vocabularies) with a so called "meaning binding". The data elements (values) of a concept can be expressed using one or several recommended standards (e.g. ATC, LOINC, SNOMED CT).

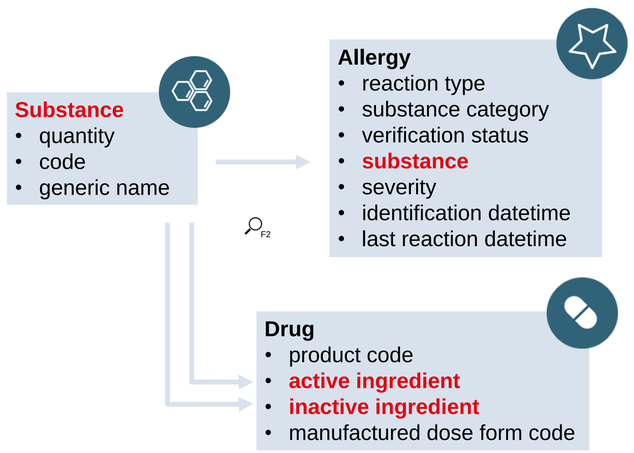
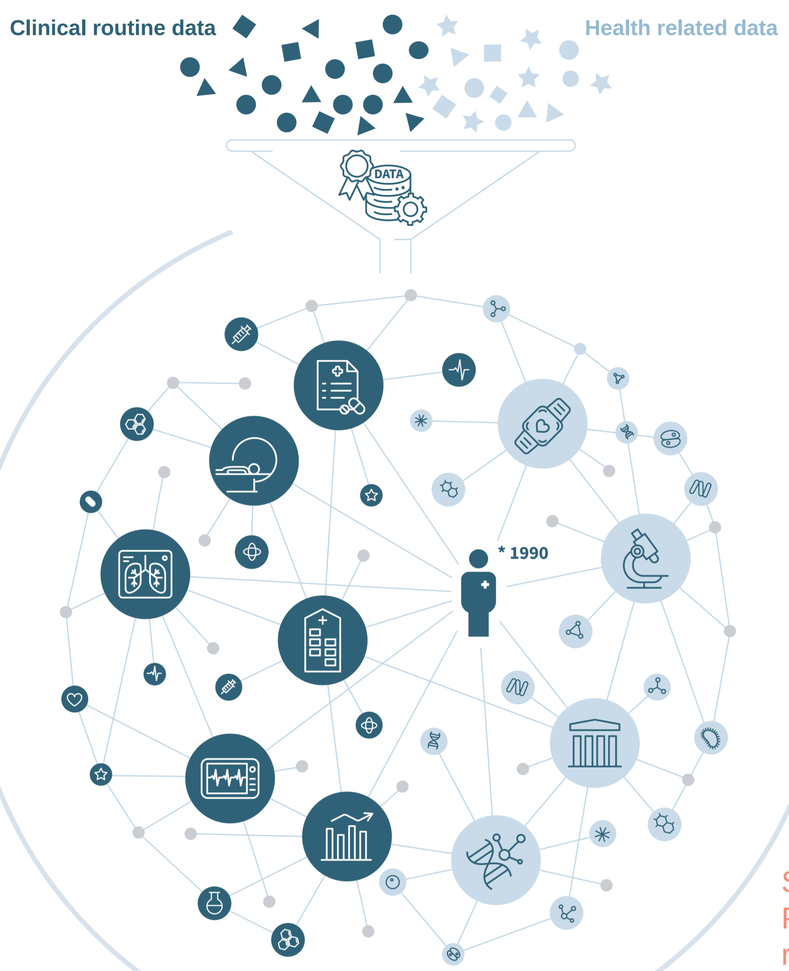


Figure 1. *Substance concept reused in two different contexts: in the Allergy concept as a substance and in the Drug concept as either an active or inactive ingredient.*



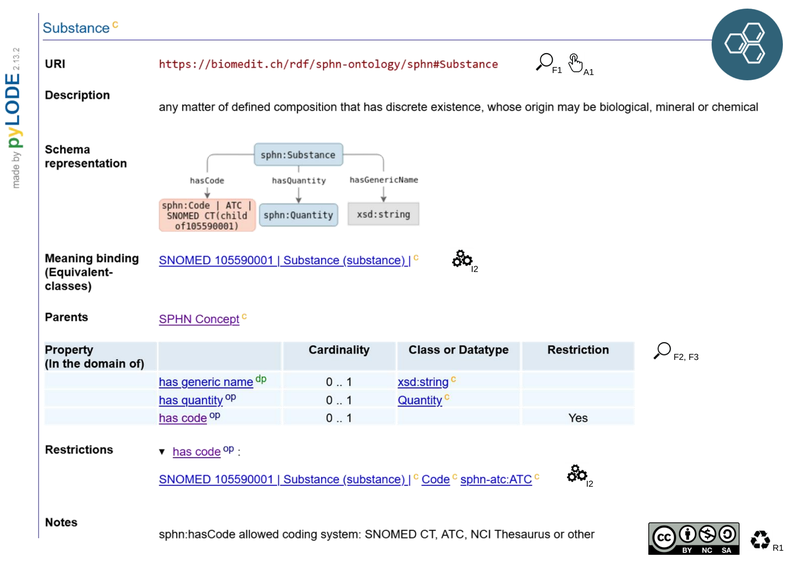


Figure 2. *Visual representation of the Substance concept using pyLODE.*

RDF graph representation

SPHN concepts and external terminologies are encoded in the Resource Description Framework (RDF) for facilitating the transport, sharing and exploration of knowledge through graph structures.

In the context of SPHN, data providers comply with the SPHN RDF schema when delivering health-related data to research projects in a FAIR manner.

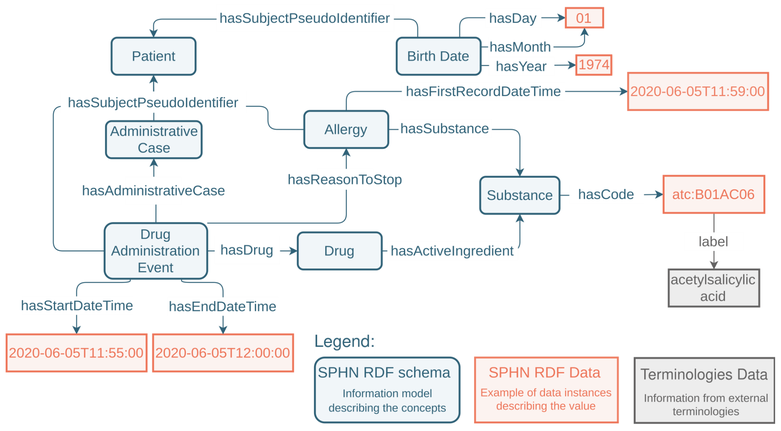


Figure 3. *Patient with a drug substance administered sees the administration being stopped because of an allergy to that substance. The substance is coded in ATC.*

Quality assurance tools

Validation and exploration of SPHN-compliant data by both data providers and data users is facilitated by several tools:

SPARQLer — automatic creation of SPARQL (SPARQL Protocol and RDF Query Language) query files for retrieving metadata connected to each concept.

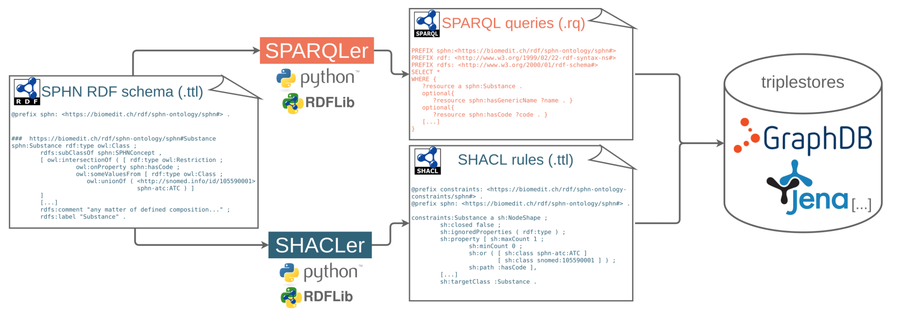


Figure 4. *SPARQLer and SHACLer tools taking as input the SPHN RDF schema.*

SHACLer — automatic creation of a file with SHACL (Shapes Constraint Language) rules.

RDF inference with the use of terminologies

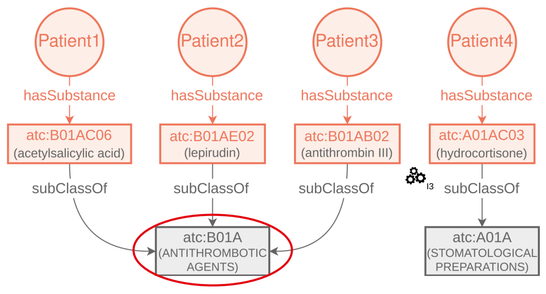


Figure 5. *Simplified example of data representation of Patients with Substances administered. The substances encoded in ATC have hierarchical information which enables to do reasoning on the data.*

The hierarchical structure of ATC enables the data user to query directly for the patients that have been given antithrombotic agents even though the data provided encoded specifically for the given substance. The computer can reason and infer knowledge that acetylsalicylic acid, lepirudin and antithrombin III are antithrombotic agents but not hydrocortisone; and therefore, retrieves patient 1, 2 and 3 only as result.

"Which patients got administered an antithrombotic agent?" F1 A1

SPARQL query

```
PREFIX atc: <https://www.whocc.no/atc_ddd_index/?code=>
PREFIX sphn: <https://biomedit.ch/rdf/sphn-ontology/sphn#>
SELECT ?patient
WHERE {
   ?patient  sphn:hasSubstance*  atc:B01A .
}
```

Results

?patient

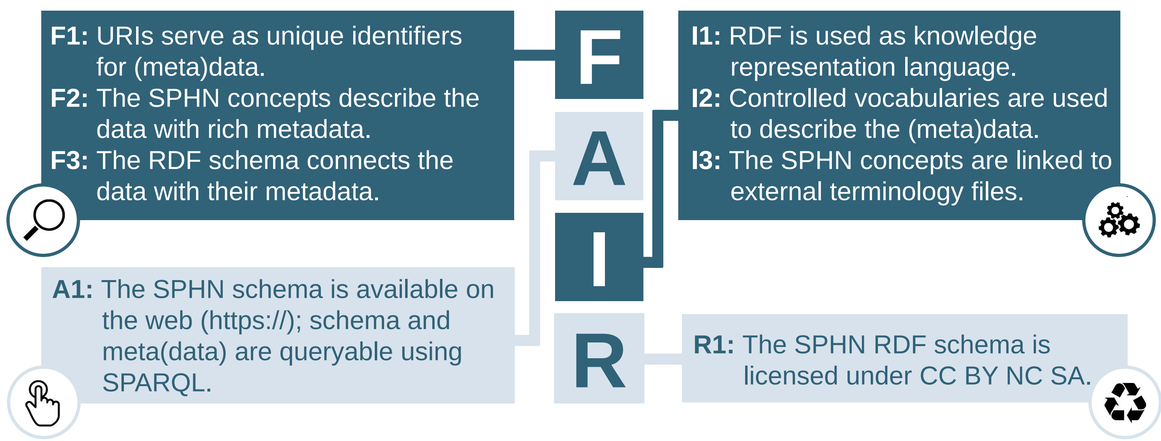
Patient 1, Patient 2, Patient 3

*Note: the * in the query after sphn:hasSubstance indicates to the RDF 'reasoner' to investigate the hierarchies when retrieving data.*

Benefits of SPHN compliant graph data for researchers:

- Exploration and analysis across data types with a single query language SPARQL;
- Metadata enrichment and standardization through external terminologies;
- Innovative research enabled by knowledge inference through linked data.

SPHN compliance to Findable Accessible Interoperable Reusable



Acknowledgements and references

The SPHN semantic framework is implemented in realm of SPHN, a project of SAMW ASSM and SIB Swiss Institute of Bioinformatics

With our partners:

Universitätsspital Basel · HUG Hôpitaux Universitaires Genève · CHUV · INSELSPITAL Universitätsspital Bern Hôpital universitaire de Berne · UniversitätsSpital Zürich · trivadis Part of Accenture

- www.w3.org/RDF; www.snomed.org; www.whocc.no/atc; www.loinc.org;
- Gaudet-Blavignac, *et al.* 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 Inform;9(6):e27591 (2021);
- Österle, *et al.* The SPHN Ecosystem Towards FAIR Data. CEUR Workshop Proceedings, SWAT4HCLS, 3127-1, 19-28 (2021);
- Wilkinson, *et al*. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016).

GitLab · Documentation · Training

SPHN is funded by the Swiss State Secretariat for Education, Research and Innovation (SERI)