Fact sheet
The SPHN semantic framework – how to achieve data interoperability

1 Introduction

The Swiss Personalized Health Network (SPHN) is a national research infrastructure initiative. The use of clinical data for research purposes is one of the most critical issues in the SPHN endeavor. The complexity of biological processes, of new therapeutical approaches, such as in immunotherapies in oncology, require increasing access to real-word data for personalized health research. Additionally, data is becoming a key factor to leverage research in all sectors of healthcare. Accessibility to data is one important aspect; data should also be findable, interoperable and reusable, according to the four FAIR principles.

Interoperability of the data is based on two important aspects. The first, is about **the semantic**, i.e. the meaning of the data. Addressing this aspect is by far the biggest challenge in interoperability in the world today. The second challenge is about the organization of the handling of the data, notably the diversity of **data types** (see Section 2), resulting in an even larger heterogeneity of data standards, data formats, data processing strategies, and data quality. Especially with regards to the interpretability of large data sets from different sources, advancing data interoperability is a key success factor.

SPHN developed a data-driven semantic framework (see Section 3) with a strong semantic layer and a model-agnostic flexible technical transport and storage pillar that allows the transport of the data with conservation of the semantic meanings.

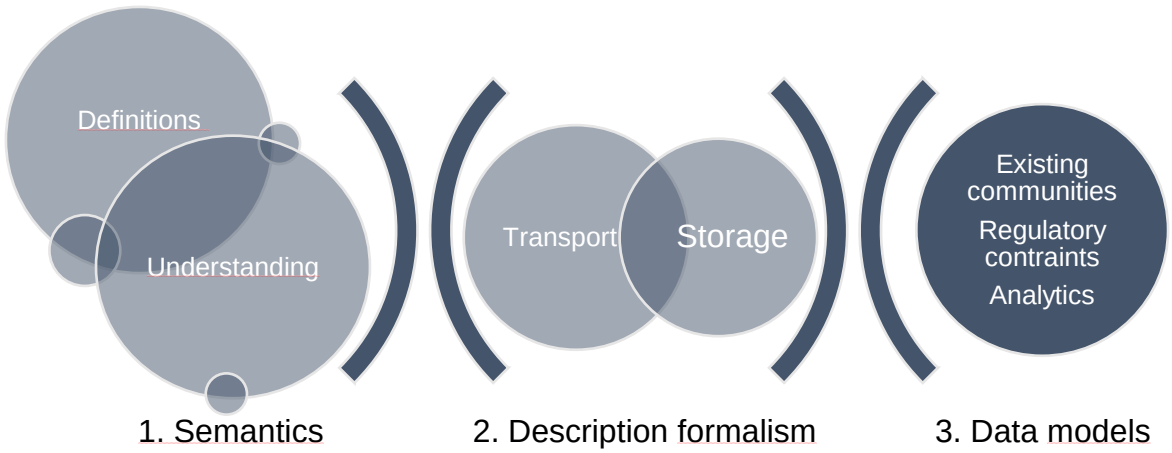

Figure 1: Three pillar SPHN semantic framework.

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

2 Types of data within SPHN

SPHN classified the data used within the SPHN driver projects in the following categories:

Hospital clinical data (routine data)

- Basic routine data (available for all patient e.g. diagnosis, demographics)
- Specific routine data (data used across domains e.g. drug allergy or samples)
- Imaging data
- Clinical data registries, including highly specialized medicine registries
- Molecular or -omics data generated in hospitals (clinical grade)

Clinical research data

- Cohorts
- Clinical study data (clinical research project, clinical trial and clinical observational study data)
- Patient self-reported and wearable device data
- Molecular or -omics data generated in research facilities (research grade)

Healthy citizen data

- Citizen/consumer health data, life style data, social media data, wearable devices

Reference data

- Reference data sets of all kind (environmental data, potential exposure to noxious agents, geographical data, statistical data…)

Projects are very diverse with respect to the types of data they use and vary significantly in the total number of variables they include.

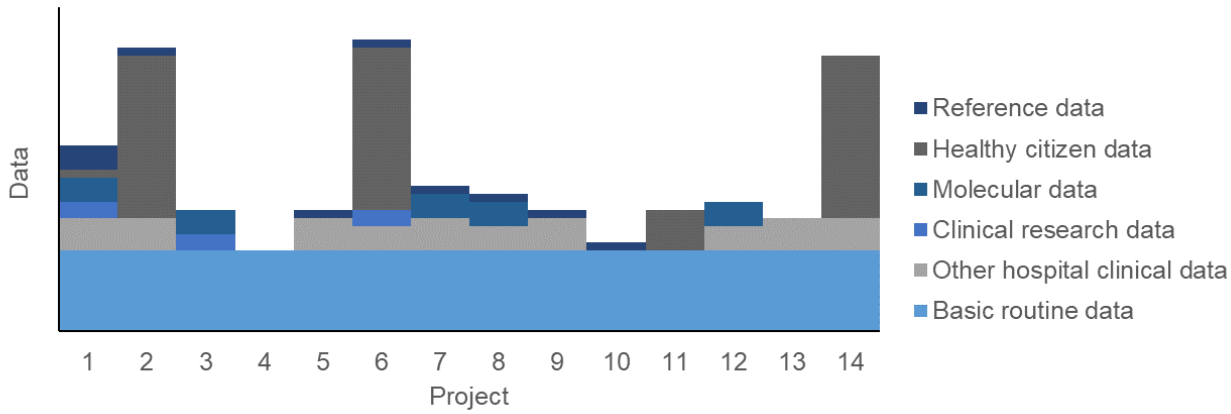

Figure 2: Examples of data types used by the different projects (for the purpose of illustration only; examples do not correspond to actual SPHN projects).

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

3 SPHN Semantic Framework

The semantic strategy of SPHN is based on a very simple set of principles.

First, the meaning of each brick of data used in SPHN must be well defined. The definition of these bricks, the semantics, is not as easy as it seems at the first glance. For example, everybody understands a sentence such as “the patient weights 74 kg”. So, the variable Weight can be defined, with a unit: kg. However, very soon it appears that there are multiple different weights. For example, the “Weight at birth”, “weight before treatment”, “reported weight”, “target weight”, to name a few. And so is it with every variable, including the simplest apparently such as blood glucose level, blood pressure, etc. In order to be able to aggregate, analyze, large cohorts of cases, a very precise definition of each of these bricks of information is required. There are other problems, such as the evolution in time. For example, a positive HIV test 30 years ago was something completely different than today. Or new diseases that must be immediately properly encoded, such as the new coronavirus 2019-nCoV.

Second, how to exchange and store the data. The data used in health is constantly evolving and augmenting. In addition, what is considered as pertinent for a specific situation is also constantly changing. This requires adopting a very flexible way of describing the container that will be used to store and transport the data. To achieve this, the SPHN is evaluating a very strong and commonly used approach that has been adopted and developed to address a similar challenge on the Web. The Resource Description Framework (RDF) belongs to the World Wide Web Consortium (W3C) specifications. It is a general method for conceptual description and modeling of information.

Third, how to communicate with existing communities. There are numerous existing communities that are using and exploiting data. Each community has its own culture and tools of data exchange. For example, the US Federal Drug Administration (FDA) imposes a standard data model named CDISC (Clinical Data Interchange Standards Consortium). In the healthcare system, the imaging industry is based on DICOM, while the electronic health record industry is based on HL7 (among them FHIR). Finally, the research community is often using OMOP or i2b2 data models.

In summary, SPHN is building a strong strategy based on having a strong semantic as first pillar; a very plastic and versatile data transport and storage second pillar; and the capacity to adapt to any specific data model allowing to cooperate with any specific community as third pillar.

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

4 How is it implemented?

In the first phase (until End of 2019) the SPHN Clinical Data Semantic Interoperability (CSI) working group focused on the semantic definition of basic (A concepts) and specific routine data (B concepts) needed to support the SPHN Driver projects and to fulfill the hospital mandate of the collaboration agreements with the hospitals (published in the SPHN dataset V2019.3).

In the SPHN semantic framework, we aim for a compositional approach: Concepts are generalizable building blocks, which can be used in different contexts. Each concept contains all information necessary to understand it, and concepts can be combined to composed concepts, which again can be combined to more complex compositions.

Of course, when defining the concepts in the SPHN semantic framework, we always have to find the right level between abstraction and granularity to optimize the power of expression. If we think of the pulse – for example – the pulse is the rate of the heart. There is also the respiratory rate; so we can abstract it to rate. The concept of rate has two attributes: number of events, e.g. number of heart beats, and the unit of time, (see Figure 3). These attributes describe a rate no matter if it is the rate of the heart or the rate of breathing. The pulse can be central or peripheral, so we have a rate of something which is measured, and an anatomical location where it is measured. Both, heart rate and respiratory rate can be combined to the concept of vital signs.

concept	Rate	number of event per unit of time
attribute	events	number of events, e.g. number of heart beats
attribute	unit	unit of time, e.g. minutes

concept	Heart Rate	frequency of the hearth beats, i.e. the number of time a hearth beats per unit of time
attribute	rate	measured heart rate, and time unit
attribute	datetime	datetime of measurement
attribute	body site	body site where the hearth rate was measured
attribute	method	method of hearth rate measurement (palpation, stethoscope, ecg, etc.)
attribute	physiologic state	finding related physiologic patient state, e.g. resting, exercise
attribute	regularity	regular or irregular heart rate

concept	Respiratory Rate	frequency at which the breathing occurs
attribute	rate	measured respiratory rate, and unit
attribute	datetime	datetime of measurement

Figure 3: Example of the concept “Rate” and its use in the concepts “Hearth rate” and “Respiratory Rate”

Where needed value sets or standards (e.g. LOINC, SNOMED-CT, ICD-10, ICD-O3, CHOP, ATC) are recommended for each attribute. Definitions for concepts, value sets and standards are aligned with national and international sources.

The SPHN clinical data set builds an overarching collection of concepts, which can be combined in a flexible manner. All entries, which are nationally harmonized across SPHN are equipped with the prefix “SPHN” in the RDF

graph. To not limit the project, data that is not nationally harmonized yet can be added to the RDF graph separated by another prefix e.g. “project name”. The flexibility of the RDF allows transferring national harmonized data, in combination with other data (other standards, format etc.). With this approach, the fraction of nationally harmonized data within a data set will increase over time. There are numerous data models, but SPHN must be able to speak with everybody, so the choice of the data model should not be predefined by the transport and storage format, but should be chosen based on the needs of the researcher’s projects, the research setting, or regulatory constraints. Based on these needs, researchers can:

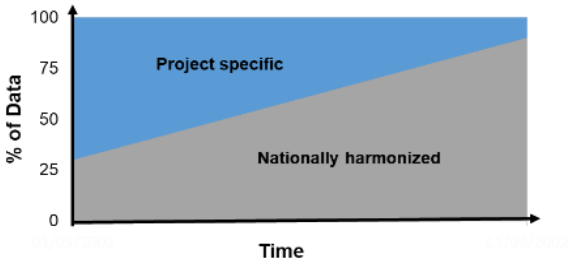

Figure 4: Ratio of project specific to nationally harmonized data within a SPHN project over time.

1. use the RDF files directly as input into their analysis software, e.g. Python
2. extract data into a flat file, e.g. Excel
3. load the data into a data model, e.g. i2b2, OMOP or a data management software, e.g. LabKey or openBIS

5 Appendix

- SPHN data set
- Semantic strategy 2019

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