Semantic Interoperability at National Scale: The SPHN Federated Clinical Routine Dataset

Jan Armida[1,†], Vasundra Touré[1,†], Philip Krauss[2], Deepak Unni[1], Harald Witte[1], Davide Chiarugi[1], Andrea Brites Marto[1], Julia Mauer[3], Thomas Geiger[3], Henning Beywl[4], Marc Daverat[5], Xeni Deligianni[6,7], Dominique Furrer[4], Mathias Gassner[8], Matthias Joos[9], Katie Kalt[9], Janshah Veettuvalappil Ikbal[4], Helena Peic Tukuljac[8], Gaëlle Vuaridel-Thurre[10], Solange Zoergiebel[10], Sabine Österle[1*]

[1] *Swiss Personalized Health Network, SIB Swiss Institute of Bioinformatics, Basel, Switzerland*
[2] *Accenture AG, Basel, Switzerland*
[3] *Swiss Academy of Medical Science, Bern, Switzerland*
[4] *Inselspital, Bern University Hospital, Bern, Switzerland*
[5] *University Hospital of Geneva, Geneva, Switzerland*
[6] *University Hospital Basel, Basel, Switzerland*
[7] *University Children's Hospital Basel, Basel, Switzerland*
[8] *University Children's Hospital Zurich, Zurich, Switzerland*
[9] *University Hospital Zurich, Zurich, Switzerland*
[10] *University Hospital Lausanne, Lausanne, Switzerland*

Abstract

Over the past eight years, the Swiss Personalized Health Network (SPHN) has established a national federated framework enabling semantically interoperable health-related data, with a primary focus on hospital clinical routine data. Rather than centralizing patient-level information, hospitals locally perform semantic coding and standardization and store SPHN-compliant data in dedicated triple stores. To promote discoverability, descriptive metadata and summary statistics derived from these local datasets are then centralized in the SPHN Metadata Catalog, which follows the SPHN Metadata Catalog Schema and aligns with European Health Data Space metadata standards.

As of 2025, the SPHN Federated Clinical Routine Dataset encompasses information from more than 800,000 patients who provided broad consent, covering the period from 2018 to present. Across the first six participating hospitals, the infrastructure holds over 12.5 billion (10^9) RDF triples mapped to 125 SPHN

SWAT4HCLS'26: 17th International Semantic Web Applications and Tools for Health Care and Life Sciences Conference, March 23-26, 2026, Amsterdam, The Netherlands

* Corresponding author
† These authors contributed equally.

jan.armida@sib.swiss (J. Armida); vasundra.toure@sib.swiss (V. Touré); philip.krauss@accenture.com (P. Krauss); deepak.unni@sib.swiss (D. Unni); harald.witte@sib.swiss (H. Witte); andrea.britesmarto@sib.swiss (A. Brites Marto); julia.maurer@sib.swiss (J. Maurer); t.geiger@sphn.ch (T. Geiger), davide.chiarugi@sib.swiss (D. Chiarugi); dominique.furrer@insel.ch (D. Furrer); henning.beywl@insel.ch (H. Beywl); janshah.veettuvalappilikbal@insel.ch (J. Veettuvalappil Ikbal); marc.daverat@hug.ch (M. Daverat); xeni.deligianni@usb.ch (X. Deligianni); matthias.joos@usz.ch (M. Joos); mathias.gassner@kispi.uzh.ch (M. Gassner); katie.kalt@usz.ch (K. Kalt); helena.peic@kispi.uzh.ch (H. Peic Tukuljac); gaelle.vuaridel-thurre@chuv.ch (G. Vuaridel-Thurre); solange.zoergiebel@chuv.ch (S. Zoergiebel); sabine.oesterle@sib.swiss (S. Österle)

0009-0007-4250-3983 (J. Armida); 0000-0003-4639-4431 (V. Touré); 0000-0002-3583-7340 (D. Unni); 0000-0002-4421-3580 (H. Witte); 0009-0005-4104-256X (A. Brites Marto); 0000-0002-0731-2758 (J. Maurer); 0000-0002-5409-4204 (D. Furrer); 0000-0002-1683-008X (H. Beywl); 0009-0002-4190-8659 (J. Veettuvalappil Ikbal); 0000-0001-9968-223X (X. Deligianni); 0009-0007-3673-8583 (K. T. Kalt); 0000-0002-9940-5678 (M. Gassner); 0000-0002-3988-4822 (H. Peic Tukuljac); 0000-0001-7895-0143 (G. Vuaridel-Thurre); 0000-0003-3248-7899 (S. Österle)

semantic concepts including demographics, diagnoses, procedures, medications, laboratory results, vital signs, clinical scores, allergies, microbiology, intensive care data, oncology, and biological samples.
This federated approach ensures that health data remain FAIR (Findable, Accessible, Interoperable, and Reusable) while safeguarding patient privacy by avoiding centralizing information. In this paper, we present the design, implementation, and scope of the SPHN Federated Clinical Routine Dataset, and its role in supporting data discoverability for research and clinical applications.

Keywords
Federated datasets, Metadata, Knowledge Graph, RDF, Terminology Usage, Real World Data

1. Introduction

Hospitals generate large amounts of clinical data as part of everyday patient care, yet leveraging this information for research remains a major challenge. A central barrier is semantic interoperability: heterogeneous data sources often use different terminologies, formats, and standards. Furthermore, language differences between institutions, such as clinical notes in German, French, or Italian in Swiss hospitals, add difficulty to harmonizing data across sites. Without standardized representation, integrating multi-center clinical data for meaningful analysis becomes slow, error-prone, and frequently not feasible at all.

The Swiss Personalized Health Network (SPHN) seeks to address these challenges within Switzerland's national health ecosystem [1]. SPHN promotes standardized data representation and integration, fostering FAIR (Findable, Accessible, Interoperable, Reusable) [2] clinical data across healthcare institutions. A key component of this effort is the SPHN RDF Schema [3], a formal, semantic data model for representing clinical concepts consistently across sites [4]. It leverages well-established international standards such as Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT, [5]), Logical Observation Identifiers Names and Codes (LOINC, [6]), International Statistical Classification of Diseases and Related Health Problems, 10th Revision, German Modification (ICD-10-GM, [7]) and Anatomical Therapeutic Chemical (ATC) classification system [8] to enhance interoperability. Swiss hospitals actively contribute to this national endeavor to construct and deliver clinical routine data according to SPHN specifications, ensuring a standardized and semantically rich dataset suitable for research use.

Centralizing patient-level clinical routine data across multiple hospitals poses significant privacy and legal challenges, as sensitive personal health information is subject to data protection regulations and institutional governance policies. To address these constraints, each participating hospital generates a standardized local extract of its routine clinical data following the SPHN semantic specifications. Collectively, these local extracts form the "SPHN Federated Clinical Routine Dataset" (SPHN FedData). Importantly, only aggregated descriptive metadata about these local datasets is stored centrally, while patient-level data remains securely stored within each hospital. This centrally managed metadata enables researchers to explore data distributions without revealing sensitive information. This supports cross-institution discoverability of data while upholding privacy and governance regulations. For in-depth research projects, access to local datasets can be requested through the respective hospital's data access procedures.

In this paper, we introduce the SPHN FedData, describe its semantic representation, available metadata, and provide a quantitative characterization of the dataset using metadata-derived statistics.

2. Methods

2.1. Data specification

The SPHN FedData contains data from patients available in the Clinical Data Platforms of the participating German and French speaking hospitals, currently University Children's Hospital Zurich (KISPI), University Hospital Basel (USB), University Hospital Bern (INSEL), University Hospital Geneva (HUG), University Hospital Lausanne (CHUV) and University Hospital Zurich (USZ), for patients who have signed the broad consent (referred to as general consent in Switzerland) and who have had a clinical case from 2018 onward. The broad consent allows for secondary use of patient level health data for not yet defined research projects. For these patients, all the data available in the clinical data platform that was mapped to the SPHN RDF Schema 2025.1 [9] is included in the SPHN FedData. Individual time series data for vital signs are not included in the RDF directly, and only metadata should be included as part of the concept "Time Series Data File". The participating hospitals are referred to as H1, H2, H3, H4, H5, and H6 throughout the remainder of this paper.

2.2. Data generation and storage

In SPHN, clinical information is modeled as RDF knowledge graphs following the SPHN RDF Schema [3], it consists of well-defined concepts (i.e., classes), each described and linked by a set of attributes (i.e., predicates). Data is generated via local data pipelines using the SPHN Connector [10], a standardized component of the hospital's internal Extract, Transform, Load (ETL) pipelines. The SPHN Connector accepts data from the clinical data platform, optionally performs de-identification (this step can also be carried out using hospital-specific tools) according to defined rules, validates the extracted elements against the SPHN RDF Schema, and transforms the resulting dataset into semantically interoperable, SPHN-compliant RDF data. Coding of the data is performed by the hospitals as part of these local pipelines. The resulting data is securely stored in a local Virtuoso [11] instance at each hospital. Each individual dataset and its metadata are regularly updated, and data from patients who have revoked their consent are promptly removed by the hospitals. The data analyzed for this pilot study adheres to the SPHN RDF Schema 2025.1 [9] and represents a snapshot as of the metadata from 18.11.2025.

2.3. Metadata extraction

Qualitative metadata follows the SPHN Metadata Catalog Schema [12] and includes general information about the SPHN FedData such as access conditions provided by the participating hospitals. Quantitative metadata was extracted by each hospital using a containerized script and comprised counts of all concepts and predicates, as well as counts of each coded element for the 18 terminologies provided in RDF in SPHN [13]: ATC, Swiss Classification of Hospital Operations (CHOP, [14]), Evidence & Conclusion Ontology (ECO, [15]), Ontology of bioscientific data analysis and data

management (EDAM, [16]), Experimental Factor Ontology (EFO, [17]), European Medical Device Nomenclature (EMDN, [18]), Genotype Ontology (GENO, [19]), Genomic Epidemiology Ontology (GenEpiO, [20]), HUGO Gene Nomenclature Committee (HGNC, [21]), ICD-10-GM, International Classification of Diseases for Oncology, 3rd Edition (ICD-O-3), LOINC, Ontology for Biomedical Investigations (OBI, [22]), Oncotree [23], Orphanet Rare Disease Ontology (ORDO, [24]), SNOMED CT, Sequence Ontology (SO, [25]), Unified Code for Units of Measure (UCUM, [26]). For codes from coding systems that are not part of the set of SPHN terminologies but may still be used within hospitals, the SPHN RDF Schema allows these elements to be structured using the SPHN "Code" concept. These are stored in text fields, making them semi-structured. Both the qualitative and quantitative metadata were converted into RDF by a customized script and made available on the SPHN Metadata Catalog [27].

2.4. Metadata analysis

The metadata of the SPHN FedData was analyzed using SPARQL queries and R scripts. A report of the analysis is available in R Markdown (https://git.dcc.sib.swiss/sphn-semantic-framework/outreach/sphn-feddata-analysis). The analysis primarily focuses on counts of core concepts. Core concepts are defined as entities directly linked to the Subject Pseudo Identifier concept (i.e., the patient) whereas non-core concepts (e.g., codes, quantities, values) are typically reused and therefore linked within a core concept, providing contextual and metadata information for that core concept. For example, Body Site is modeled as a non-core concept and is reused in measurements and other procedures which are themselves core concepts. The 2025.1 release of the SPHN RDF Schema (hereafter referred to as '2025.1') comprises 168 SPHN concepts including 75 core concepts (not including "SPHN Concept" and "Measurement" which are non-instantiable concepts). To provide a clear overview of the domain coverage within the SPHN FedData and avoid inflation of counts due to the inclusion of lower-level instance data, the analysis was restricted only to the core concepts. To improve the readability of figures, core concepts provided in the SPHN FedData were further grouped into 8 specific domains, as specified in Table 1.

Table 1. SPHN core concepts available in hospitals grouped into specific domains. Sixty out of the total seventy-five core concepts defined in the SPHN RDF Schema have been grouped into higher-level categories to facilitate readability in the analysis' plots.

Domains	Core concepts in SPHN FedData
Administrative	Administrative Case, Consent, Follow Up, Healthcare Encounter, Insurance Status, Resuscitation Directive
Clinical condition	Access Device Presence, Allergy, Assessment Event, Billed Diagnosis, Diagnosis, Implant Presence, Nursing Diagnosis, Transplant Presence

Demographics	Administrative Sex, Age, Birth, Civil Status, Death, Home Address, Nationality
Laboratory	Antimicrobial Susceptibility Lab Test Event, Biobank Sample, Isolate, Lab Test Event, Microbiology Biomolecule Presence Lab Test Event, Microbiology Microscopy Lab Test Event, Microorganism Identification Lab Test Event, Sample, Sample Processing
Measurement	Blood Pressure Measurement, Body Height Measurement, Body Mass Index, Body Position, Body Surface Area, Body Temperature Measurement, Body Weight Measurement, Cardiac Index, Cardiac Output Measurement, Circumference Measurement, Fluid Balance, Fluid Input Output, Gestational Age At Birth, Heart Rate Measurement, Nutrition Intake, Oxygen Saturation Measurement, Respiratory Rate Measurement, Time Series Data File
Medical procedure	Billed Procedure, Electrocardiographic Procedure, Imaging Procedure, Organ Support, Radiotherapy Procedure
Medication and treatment	Drug Administration Event, Drug Prescription, Oxygen Administration Event
Oncology	Oncology Diagnosis, Oncology Surgery, Tumor Grade Assessment Event, Tumor Stage Assessment Event

3. Results

This section summarizes the main findings regarding the SPHN FedData, focusing on its (i) metadata accessibility and exploration, (ii) SPHN concepts coverage, and (iii) terminology coverage across participating hospitals. The results illustrate the scale and diversity of the federated datasets, providing insight into the data standardization progress achieved within SPHN.

3.1. Metadata accessibility and exploration

All descriptive metadata of the SPHN FedData is publicly accessible through the SPHN Metadata Catalog (https://fdp.dcc.sib.swiss) and can be explored through the SPHN Schema Scope (https://schemascope.dcc.sib.swiss). Figure 1 presents a screenshot of the schema graph, displaying the classes and their instance counts. Together, these two complementary components provide transparent access to the structure and scope of the dataset while ensuring that no patient-level information is disclosed. The open accessibility of the metadata supports the FAIR data principles (Findable, Accessible, Interoperable, and Reusable) and promotes metadata exploration and collaboration within the research community.

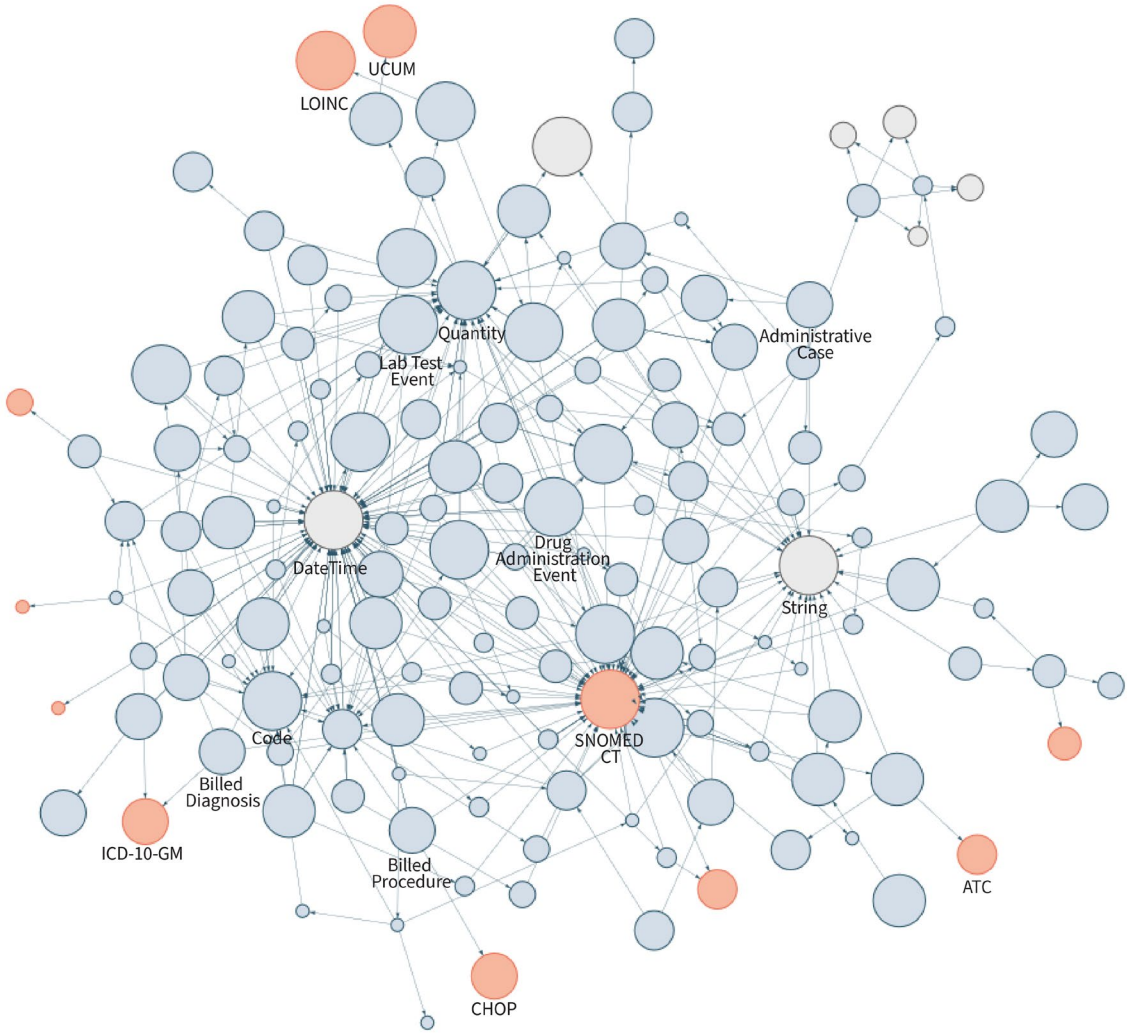

Figure 1: Overview of the SPHN FedData metadata. The Figure shows the SPHN FedData metadata visualized in the SPHN Schema Scope. Blue nodes represent concepts; grey nodes indicate datatypes, and orange nodes correspond to code systems. In the view shown, the node size is proportional to the number of instances per class (scaled in steps). Edges between nodes reflect attributes defined in the SPHN RDF Schema [3].

3.2. SPHN concepts coverage

The federated dataset currently includes data from six hospitals**,** covering approximately 800,000 patients who provided broad consent and had an administrative case between 2018 and 2025. The local RDF databases collectively contain 1.21×10^{10} RDF triples (see Table 2), representing a substantial amount of data.

Table 2. General key statistics for individual hospital datasets. The table reflects the number of Subject Pseudo Identifier (i.e., "Patients"), triples, SPHN concepts, and core concepts available at each individual hospital (columns H1-H6). Aggregate total numbers across all hospitals (H1-H6) are

also provided in column “Total”. SPHN concepts include all concepts defined in 2025.1 while the core concepts represent a subset which only includes concepts directly linked to a Subject Pseudo Identifier.

	H1	H2	H3	H4	H5	H6	Total
Patients	99,344	113,194	230,242	106,075	145,366	120,863	815,084
Triples [x10^9]	1.5	2.5	3.4	0.6	2.4	2.1	12.1
SPHN concepts	61	92	89	73	86	101	125
SPHN core concepts	26	45	42	34	38	45	60

Across all sites, the dataset comprises 125 out of the 168 defined SPHN concepts, including 60 of the 75 core concepts, and covers a wide range of domains including demographics, diagnoses, procedures, medications, laboratory (including microbiology and samples), measurements (e.g., vital signs), and allergies. Variability can be observed between hospitals on the core concepts delivered as only 18 of them are common to all six hospitals (Figure 2). The list of 18 core concepts provided across all six hospitals are: Administrative Case, Administrative Sex, Age, Assessment Event, Billed Diagnosis, Billed Procedure, Birth, Body Height Measurement, Body Temperature Measurement, Consent, Death, Drug Administration Event, Drug Prescription, Healthcare Encounter, Heart Rate Measurement, Lab Test Event, Oxygen Saturation Measurement and Sample. This highlights differences in local data capture and availability and may have implications for multi-site analysis as it may limit comparability for certain domains.

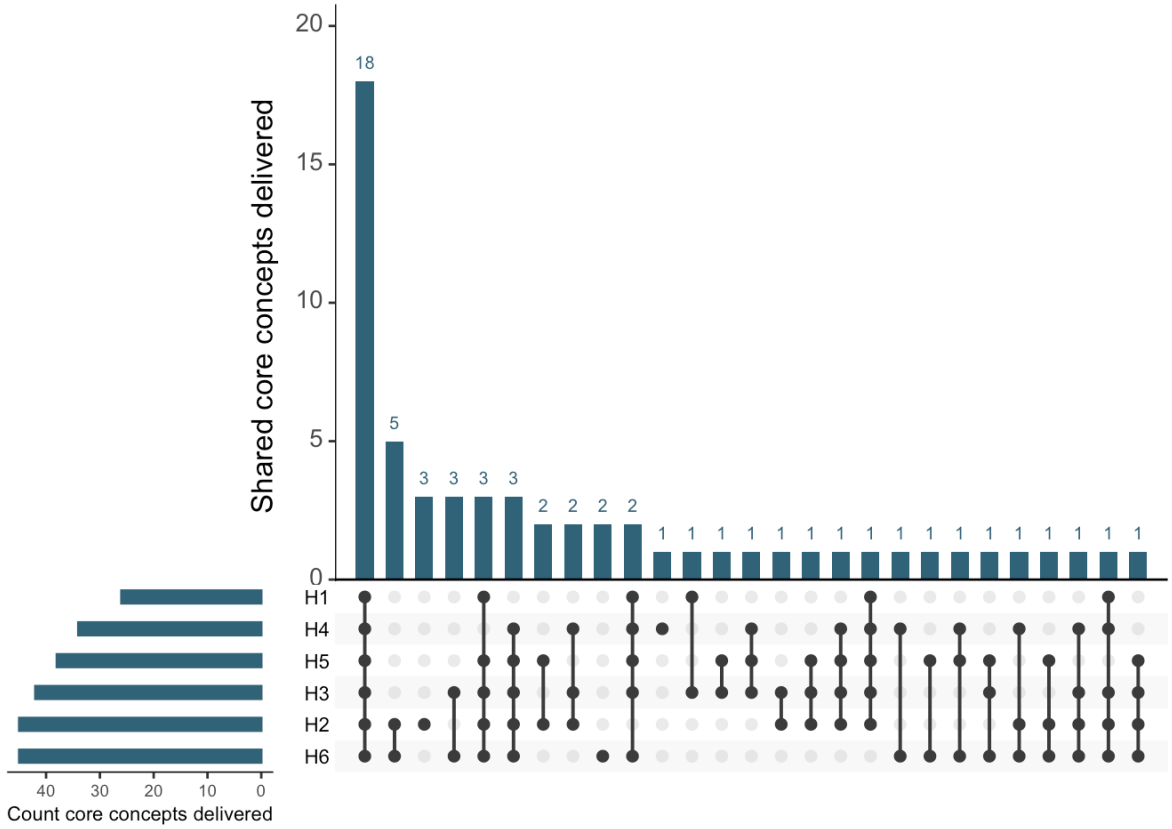

Figure 2. Overlap of shared core concepts across hospitals (H1-H6). Vertical bars show the number of core concepts shared among hospitals while horizontal bars show the total number of core concepts delivered by each hospital. The dotted lines indicate the number of hospitals and specify which hospitals provide the same core concepts.

Figure 3 shows the distribution of core concepts counts, clustered into specific domains, per SPHN concept and hospital. Across the six centers, the delivered core concepts vary in domain coverage by hospital, though the most frequently instantiated concepts are generally consistent. The top 3 are Lab Test Event, Drug Administration Event and Oxygen Saturation Event each with more than 30 million instances. Less frequent domains, such as oncology (e.g., only 2126 instances of Oncology Diagnosis), or not yet delivered domains such as omics, represent emerging areas where data harmonization is still ongoing.

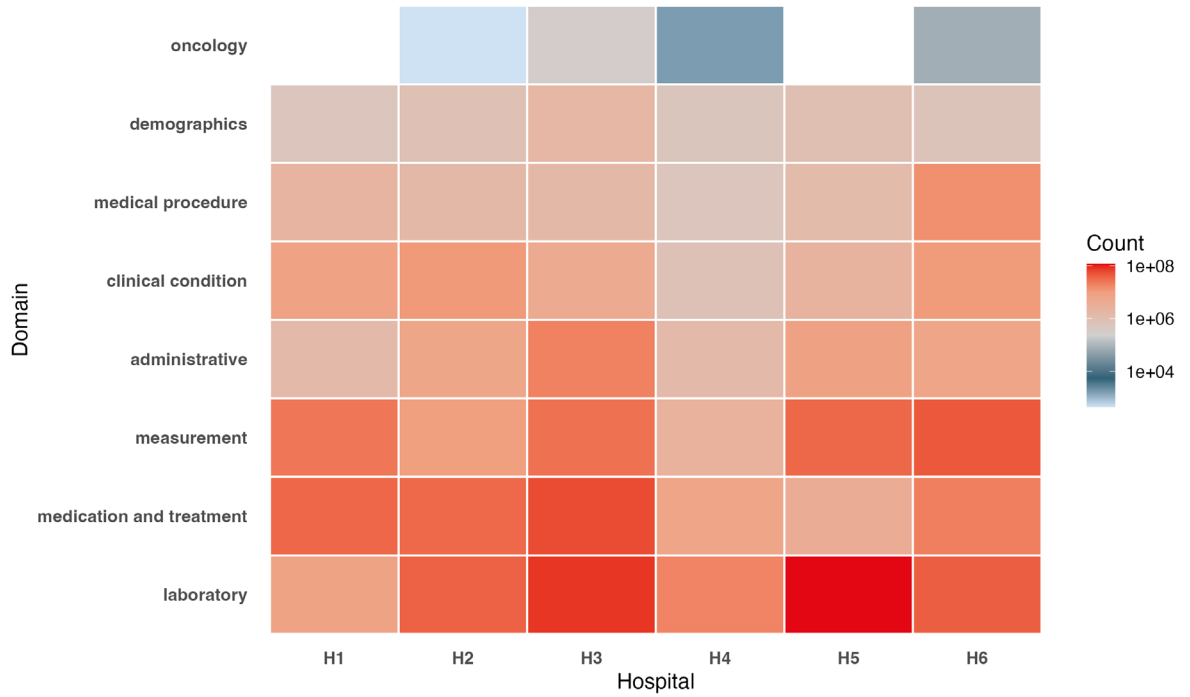

Figure 3: Coverage and total number of instances of core concepts domains delivered across each hospital. Red colored cells show a high number of instances available (e.g. in laboratory and medication and treatment domains) while blue colored cells show a lower number of instances available (e.g. in oncology). White cells indicate that this domain was not instantiated by the hospital, hence it was not available in the data.

3.3. Terminology coverage

In SPHN, a set of 18 terminologies are supported within the SPHN RDF Schema, enabling hospitals to represent data using standardized and uniquely identifiable codes. Across all participating sites, approximately 750 million instances of terminology codes were identified, with 41,140 distinct ones, originating from 7 out of the 18 available terminologies. Table 3 provides detailed information of 5 selected terminologies, LOINC, SNOMED CT, ICD-10-GM, CHOP, and ATC, which were chosen for their importance and relevance within the dataset. CHOP and ICD-10-GM have already been in use prior to the SPHN initiative for coding billing information and consequently predominate in coding across hospitals. They are followed by a significant usage of LOINC and SNOMED CT, which notably mainly started to be implemented as a standard since the beginning of SPHN.

Among the used terminologies, we observe the following:

- ICD-10-GM and CHOP are used primarily for billing diagnoses and procedures, comprising approximately 30% and 35% of all the terminology-coded instances, respectively. Their coding exhibits considerably higher cross-hospital consistency, indicating a mature and stable coding practice with well-defined usage due to harmonized billing processes.

- SNOMED CT serves as a cross-domain ontology for detailed clinical concepts and phenotype definitions, contributing to approximately 16% of all code occurrences.

- LOINC contributes to approximately 9% of code occurrences, and it is used for coding laboratory tests. However, it presents a mixed coding pattern, with a predominance of site-specific implementations and laboratory test procedures, not shared across hospitals.

- ATC codes describe administered or prescribed active substances or substance groups within medications, contributing to approximately 5% of code occurrences.

- Codes from other coding systems, including ORDO (used by two of the six hospitals) and UCUM contribute to approximately 5% of all terminology-coded occurrences.

A cross-hospital comparison of terminologies usage, visualized in Figure 4, reveals substantial overlap for CHOP and ICD-10-GM codes, reflecting their established and harmonized use in administrative and billing workflows. In contrast, LOINC and SNOMED CT exhibit site-specific variability, suggesting local customization of laboratory catalogs and heterogeneous adherence to variable implementation guidelines. Collectively, these findings indicate that administrative terminologies (ICD-10-GM, CHOP) are well standardized across institutions, yet clinical and laboratory terminologies (LOINC, SNOMED CT) remain heterogeneous and represent priority areas for ongoing interoperability enhancement within the SPHN framework.

Table 3. Terminology usage of individual hospital datasets. Count of distinct codes across the main terminologies available at each hospital (H1-H6) and combined across all hospitals. Identical codes delivered in multiple hospitals are only counted once in the column "Total".

	H1	H2	H3	H4	H5	H6	Total
SNOMED CT	3,077	2,119	669	871	1,603	2,683	6,518
LOINC	1,108	1,359	1,426	220	699	890	3,693
ATC	0	1,331	1,682	843	511	88	1,996
ICD-10-GM	10,378	9,105	10,894	6,124	9,044	10,104	12,540
CHOP	10,675	7,837	10,009	4,938	7,613	10,054	14,244

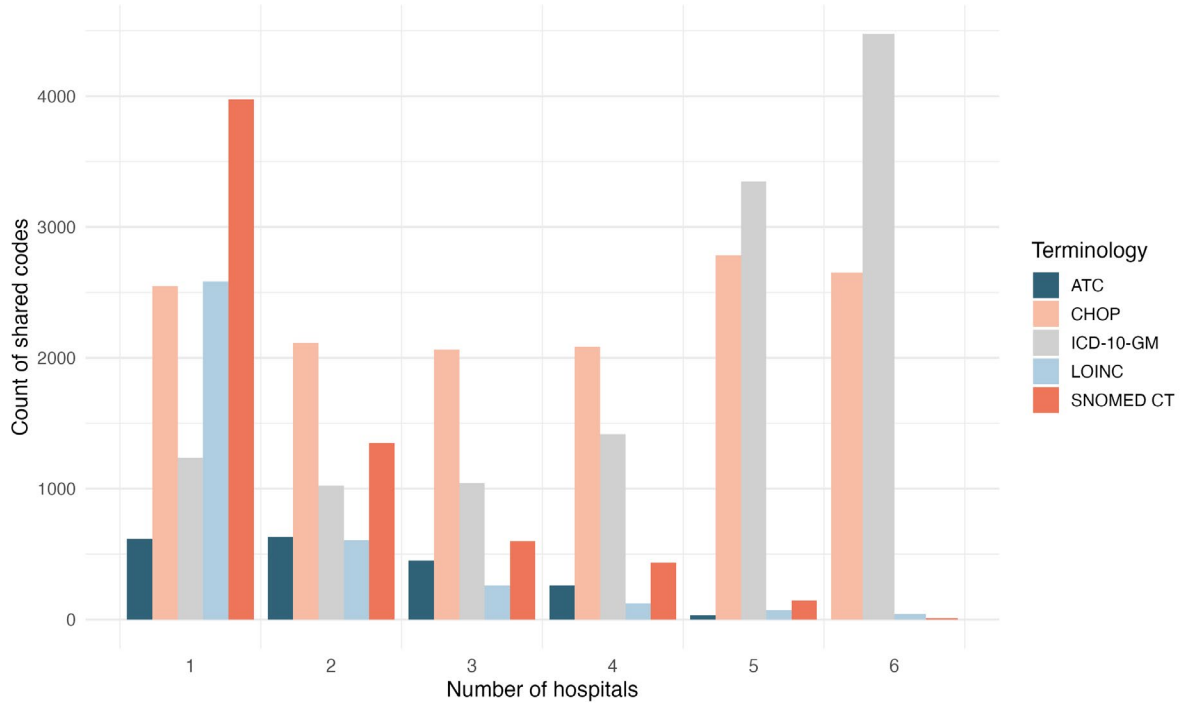

Figure 4. Distribution of shared codes across hospital for each terminology. The x-axis represents the number of hospitals sharing each code, while the y-axis shows the count of codes shared by that many hospitals.

4. Discussion

The SPHN FedData demonstrates the value of maintaining semantic interoperability at a national scale. By standardizing datasets using 125 SPHN concepts (including 60 core ones) across hospitals and aligning them with international terminologies, researchers and infrastructure developers can understand and query hospital data in a uniform way. Data is validated within the SPHN Connector, ensuring that all elements are semantically compliant with the SPHN RDF Schema definitions. Semantic consistency is critical for national-scale data-driven projects, cohort discovery, and cross-hospital feasibility assessments, where the ability to interpret data unambiguously directly impacts the reliability and efficiency of research projects. However, semantic compliance does not guarantee the accuracy of the underlying data values; these still need to be carefully checked by the hospitals in their pipelines.

One of the main strengths of the SPHN FedData lies in its metadata-driven discoverability. The SPHN FAIR Data Point and the SPHN Schema Scope visualization provide descriptive metadata, concept-level summaries, and code-level counts without exposing sensitive personal information. The federated approach supports the understanding of hospital data content while inherently featuring data privacy-preservation and supports future data availability for research. Additionally, the method is also scalable: as new hospitals are onboarded and data volume increases, researchers can continue to explore the dataset using metadata alone, thereby reducing the need for repeated feasibility requests to individual hospitals.

Furthermore, it supports equitable access to health data, ensuring that researchers across institutions can benefit from interoperable datasets to generate scientific knowledge for the broader public good. At the same time, the importance of hospitals' data governance structures and data-access procedures remains fully acknowledged. These frameworks enable hospitals to safeguard patient rights and ensure that any access to health data, whether at the patient level or not, is conducted responsibly and in full compliance with ethical and legal regulations. By streamlining access pathways and reducing administrative complexity, such structures also help avoid opportunity costs that arise when valuable data remain underused. In turn, more efficient data utilization prevents delays in research, innovation, and decision-making that could otherwise limit potential health and societal benefits.

Despite these advantages, several technical challenges emerged during implementation. Analysis of the metadata reveals variability in the relative completeness of different concepts across hospitals. Some domains, such as billing information, medication and vital signs, are available across all hospitals, whereas others, such as oncology-related concepts, remain sparsely represented. This reflects the fact that such information is often not yet available in structured and coded form at scale in the hospitals. Furthermore, differences in local coding practices contribute to semantic heterogeneity: the same clinical concept may be represented with different terminologies (e.g. substances were provided both in ATC and SNOMED CT) or codes within a single terminology (e.g. laboratory tests in LOINC with or without method metadata, using varying units; different level of granularity when coding with SNOMED CT). A clear contrast is observed between terminologies guided with institutional alignment (e.g. CHOP, ICD-10-GM) and those depending on local implementation decisions (e.g. LOINC, SNOMED CT, ATC). In addition to this semantic variability, differences in the number of triples produced per hospital arise from the varying of delivered concepts, the occasional inclusion of timeseries data (which increases the number of triples), and hospital-specific practices regarding instance reuse. The federated architecture preserves flexibility and local control, but it also requires continuous curation, mapping maintenance, and cross-site alignment to ensure that the semantic layer remains coherent and consistent over time. A national alignment is therefore necessary for improving data interoperability. Finally, despite the wide use of terminologies, the SPHN FedData also contains more than 50 million instances of the "SPHN Code" concept. However, we lack information about these instances to determine further details. These instances can on the one hand represent codes from coding systems locally used in a specific hospital, thus functioning as structured free text rather than semantically defined terminology codes. This emphasizes the continued need for mapping locally used terms and codes to standardized terminologies. On the other hand, instances of the "SPHN Code" concept can originate from sources which are currently not provided in RDF by SPHN, but which are used in hospitals to code data. This includes, for example, the Global Trade Item Numbers (GTIN, [28]), or unique identifiers for medicinal products.

5. Conclusion

The SPHN Federated Clinical Routine Dataset represents a major milestone in showcasing the clinical routine data available across multiple Swiss institutions for research. With the metadata-driven data discovery approach, researchers can rapidly explore and assess the scope and availability of clinical routine data without the need to start lengthy (legal and contractual) procedures to access this

information. This approach facilitates evidence-based feasibility assessment and allows researchers to initiate data access requests with a clear expectation of semantic interoperability and alignment with specific research requirements.

Crucially, the initiative is designed for growth and continuity. Additional hospitals, clinical domains, SPHN concepts, and standardized terminology usage are expected to be incorporated over time. In parallel, ongoing harmonization efforts and iterative feedback loops will progressively improve data quality. As mapping coverage increases and semantic alignment matures, subsequent iterations of the SPHN Federated Clinical Routine Dataset will offer greater depth, consistency, and analytical utility. With sustained alignment across institutions, and reinforcement of national terminology governance (particularly in domains such as laboratory tests, medical devices, and medicinal products), this infrastructure has the potential to become a cornerstone for scalable, reproducible, and privacy-preserving clinical research in Switzerland.

Acknowledgments

This work was funded by the State Secretariat for Education, Research, and Innovation (SERI) via SPHN. The authors would like to thank all the teams in the hospitals, which contributed to making this data available in SPHN format. The authors would also like to thank all colleagues involved, especially from USZ, Amanda Ramirez Ramos, Lirui Zhang and Daniele Vaccaro, at KISPI Beat Bangerter, at CHUV Yves Jäggi, at INSEL Pedram Bürgin, at HUG Marouan Borja and at USB Bram Stieltjes and Rita Achermann. Further the authors would like to thank the SPHN partners, at the BioMedIT nodes and SPHN projects, which contributed to the SPHN RDF Schema, the IICU consortia especially Nora Toussaint and team, the SPO consortia especially Sylvain Pradervand and team, the LUCID consortia, especially Julien Ehrsam, Oksana Riba Grognuz and team and the SwissPedHealth consortia, especially Xeni Deligianni and Fabien Belle.

Declaration on Generative AI

During the preparation of this work, the authors used GPT-4 to fix formulations, grammar, and spelling checks. After using these services, the authors reviewed and edited the content as needed and took full responsibility for the content of this publication.

References

[1] 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 Inform*, vol. 9, no. 6, p. e27591, Jun. 2021, doi: 10.2196/27591.

[2] M. D. Wilkinson *et al.*, “Comment: The FAIR Guiding Principles for scientific data management and stewardship,” *Sci Data*, vol. 3, no. 1, pp. 1–9, Mar. 2016, doi: 10.1038/SDATA.2016.18.

[3] “The SPHN RDF Schema.” Accessed: Nov. 17, 2025. [Online]. Available: https://www.biomedit.ch/rdf/sphn-schema/sphn

[4] V. Touré *et al.*, “FAIRification of health-related data using semantic web technologies in the Swiss Personalized Health Network,” *Sci Data*, vol. 10, no. 1, p. 127, Mar. 2023, doi: 10.1038/s41597-023-02028-y.
[5] G. Benson Tim and Grieve, “SNOMED CT,” in *Principles of Health Interoperability: SNOMED CT, HL7 and FHIR*, Cham: Springer International Publishing, 2016, pp. 155–172. doi: 10.1007/978-3-319-30370-3_9.
[6] “LOINC - Logical Observation Identifiers Names and Codes.” Accessed: Nov. 17, 2025. [Online]. Available: https://loinc.org/
[7] “ICD-10-GM - International Statistical Classification of Diseases and Related Health Problems, 10th revision, German Modification.” Accessed: Nov. 17, 2025. [Online]. Available: https://www.bfarm.de/EN/Code-systems/Classifications/ICD/ICD-10-GM/_node.html
[8] “ATC - The Anatomical Therapeutic Chemical .” Accessed: Nov. 17, 2025. [Online]. Available: https://atcddd.fhi.no/
[9] “The SPHN RDF Schema - Release 2025.1.” Accessed: Nov. 18, 2025. [Online]. Available: https://www.biomedit.ch/rdf/sphn-schema/sphn/2025/1
[10] V. Touré *et al.*, “SPHN Connector - A scalable pipeline for generating validated knowledge graphs from federated and semantically enriched health data,” Nov. 2025, doi: 10.21203/RS.3.RS-7930982/V1.
[11] “OpenLink Virtuoso.” Accessed: Nov. 18, 2025. [Online]. Available: https://virtuoso.openlinksw.com/knowledgegraph/
[12] “SPHN Metadata Catalog Schema.” Accessed: Nov. 19, 2025. [Online]. Available: https://sphn.gitlab.io/sphn-metacat-schema/
[13] P. Krauss, V. Touré, K. Gnodtke, K. Crameri, and S. Österle, “DCC Terminology Service—An Automated CI/CD Pipeline for Converting Clinical and Biomedical Terminologies in Graph Format for the Swiss Personalized Health Network,” *Applied Sciences 2021, Vol. 11, Page 11311*, vol. 11, no. 23, p. 11311, Nov. 2021, doi: 10.3390/APP112311311.
[14] Bundesamt für Statistik (BFS), “CHOP - Schweizerische Operationsklassifikation.” Accessed: Nov. 18, 2025. [Online]. Available: https://www.bfs.admin.ch/bfs/de/home/statistiken/gesundheit/nomenklaturen/medkk/instrumente-medizinische-kodierung.assetdetail.32128591.html
[15] M. Giglio *et al.*, “ECO, the Evidence & Conclusion Ontology: community standard for evidence information,” *Nucleic Acids Res*, vol. 47, no. D1, pp. D1186–D1194, Jan. 2019, doi: 10.1093/NAR/GKY1036.
[16] J. Ison *et al.*, “EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats,” *Bioinformatics*, vol. 29, no. 10, pp. 1325–1332, May 2013, doi: 10.1093/BIOINFORMATICS/BTT113.
[17] J. Malone *et al.*, “Modeling sample variables with an Experimental Factor Ontology,” *Bioinformatics*, vol. 26, no. 8, pp. 1112–1118, Apr. 2010, doi: 10.1093/BIOINFORMATICS/BTQ099.
[18] European Commission, “EMDN - European Medical Devices Nomenclature.” Accessed: Nov. 18, 2025. [Online]. Available: https://health.ec.europa.eu/medical-devices-topics-interest/european-medical-devices-nomenclature-emdn_en
[19] “GENO - Genotype Ontology.” Accessed: Nov. 18, 2025. [Online]. Available: https://github.com/monarch-initiative/GENO-ontology
[20] “GenEpiO - The Genomic Epidemiology Application Ontology.” Accessed: Nov. 18, 2025. [Online]. Available: https://github.com/GenEpiO/genepio
[21] R. L. Seal *et al.*, “Genenames.org: the HGNC resources in 2023,” *Nucleic Acids Res*, vol. 51, no. D1, pp. D1003–D1009, Jan. 2023, doi: 10.1093/NAR/GKAC888.
[22] A. Bandrowski *et al.*, “The Ontology for Biomedical Investigations,” *PLoS One*, vol. 11, no. 4,

Apr. 2016, doi: 10.1371/JOURNAL.PONE.0154556.
[23] R. Kundra *et al.*, “OncoTree: A Cancer Classification System for Precision Oncology,” *JCO Clin Cancer Inform*, vol. 5, no. 5, pp. 221–230, Dec. 2021, doi: 10.1200/CCI.20.00108.
[24] “ORDO – Orphanet Rare Disease ontology .” Accessed: Nov. 17, 2025. [Online]. Available: https://sciences.orphadata.com/ordo/
[25] K. Eilbeck *et al.*, “The Sequence Ontology: a tool for the unification of genome annotations,” *Genome Biology 2005 6:5*, vol. 6, no. 5, pp. R44-, Apr. 2005, doi: 10.1186/GB-2005-6-5-R44.
[26] Inc. Regenstrief Institute, “UCUM - The Unified Code for Units of Measure.” Accessed: Nov. 18, 2025. [Online]. Available: https://ucum.org/ucum
[27] “SPHN FAIR Data Point.” Accessed: Nov. 17, 2025. [Online]. Available: https://fdp.dcc.sib.swiss/
[28] GS1, “GTIN - Global Trade Item Number.” Accessed: Nov. 18, 2025. [Online]. Available: https://www.gs1.org/standards/id-keys/gtin