Swiss Personalized Health Network

SPHN

The SPHN SHACLer

Automatic generation of SHACL rules based on the SPHN RDF Schema

Dr. Vasundra Touré, Scientific Coordinator
Personalized Health Informatics, SIB Swiss Institute of Bioinformatics

A project of

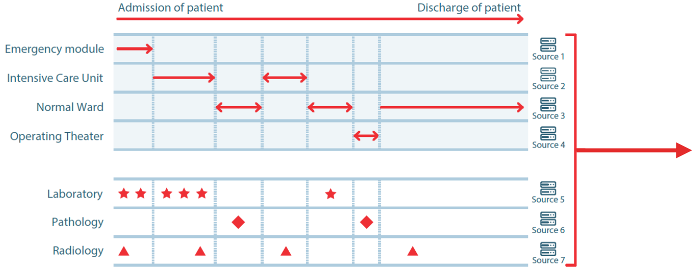
SAMW ASSM
Schweizerische Akademie der Medizinischen Wissenschaften
Académie Suisse des Sciences Médicales
Accademia Svizzera delle Scienze Mediche
Swiss Academy of Medical Sciences

SIB
Swiss Institute of Bioinformatics

30.05.24

Swiss Personalized Health Network at a glance

Clinical routine data

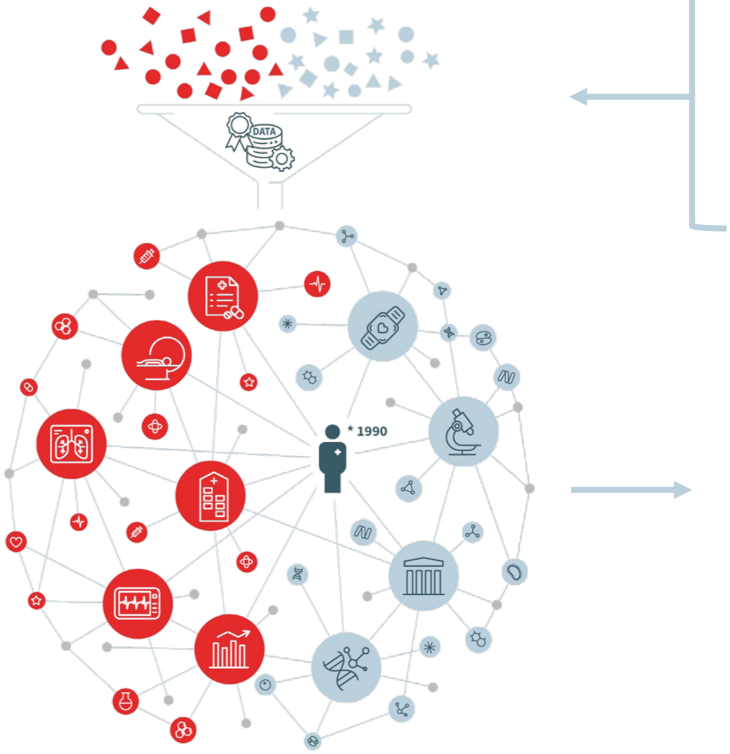


→ Data capture along the patient path through the hospital.

▲ ★ Data capture in specified hospital facilities

Other health-related data

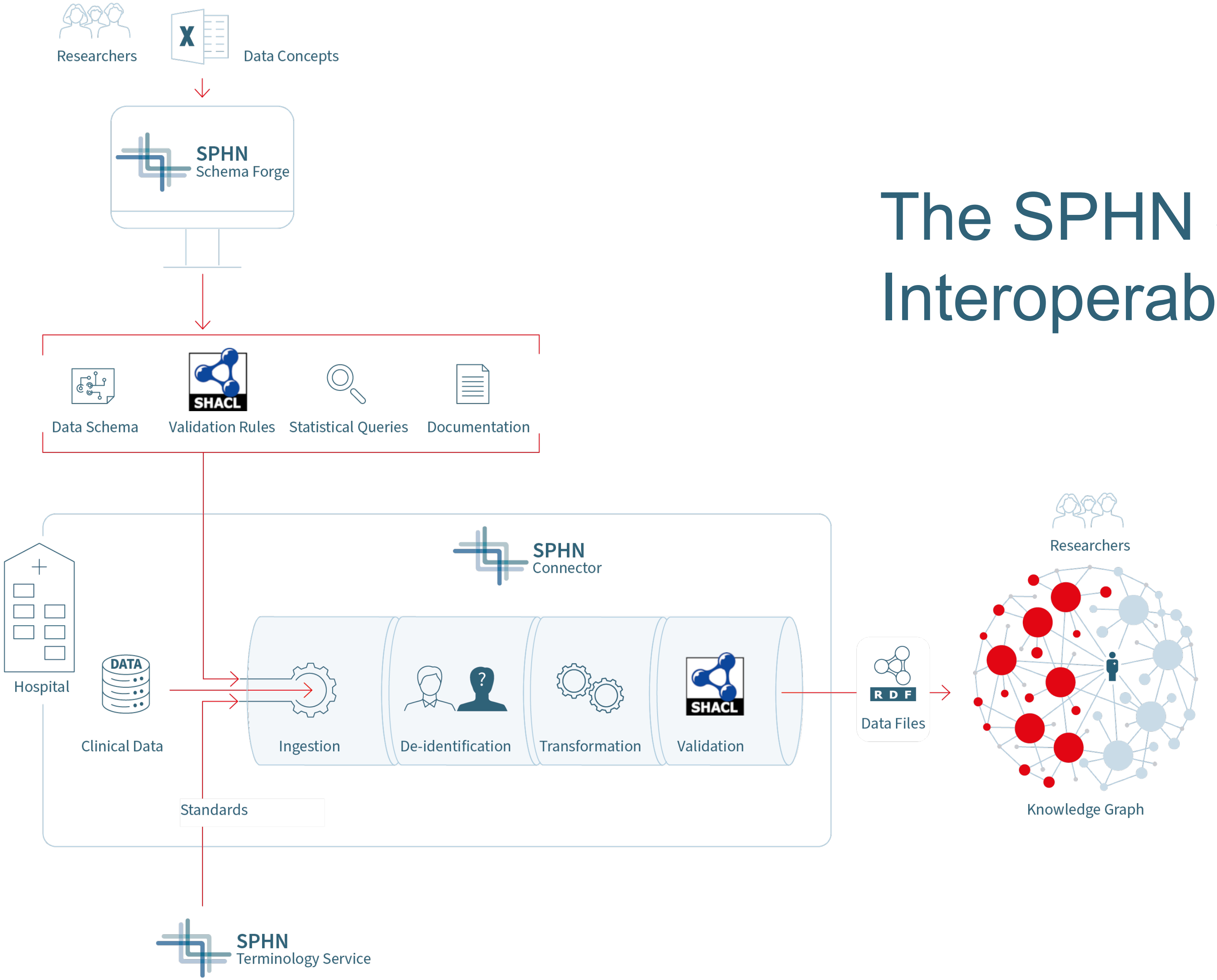
- Clinical Trial Data
- Registry Data
- Patient Reported Outcomes
- Sensor Data
- Medical Imaging
- etc.



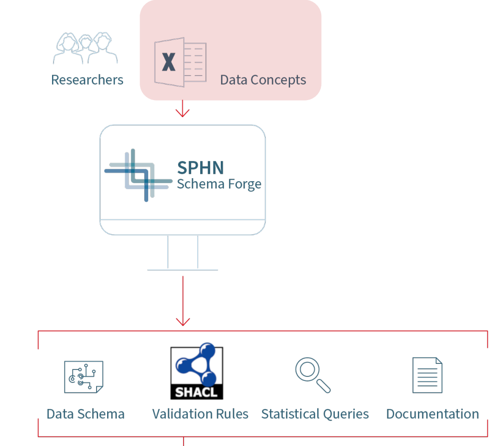
Integrated FAIR research data

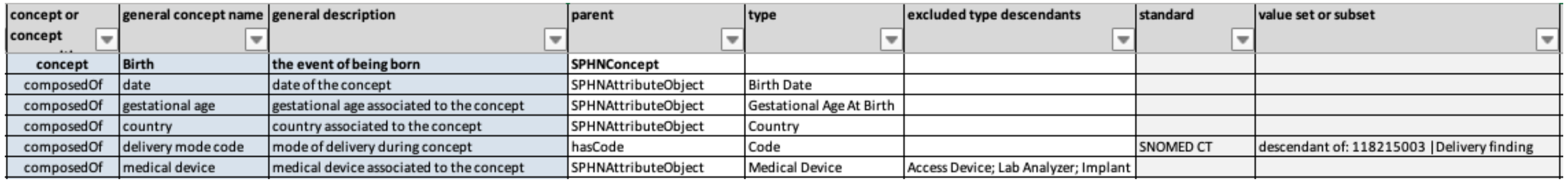
“From clinical routine data to FAIR research data”

The SPHN Semantic Interoperability Framework

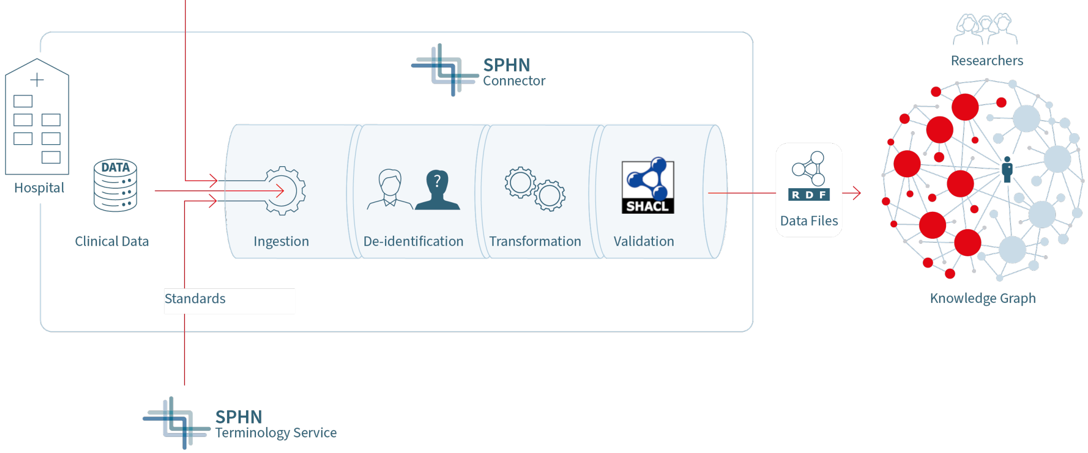


The SPHN Semantic Interoperability Framework



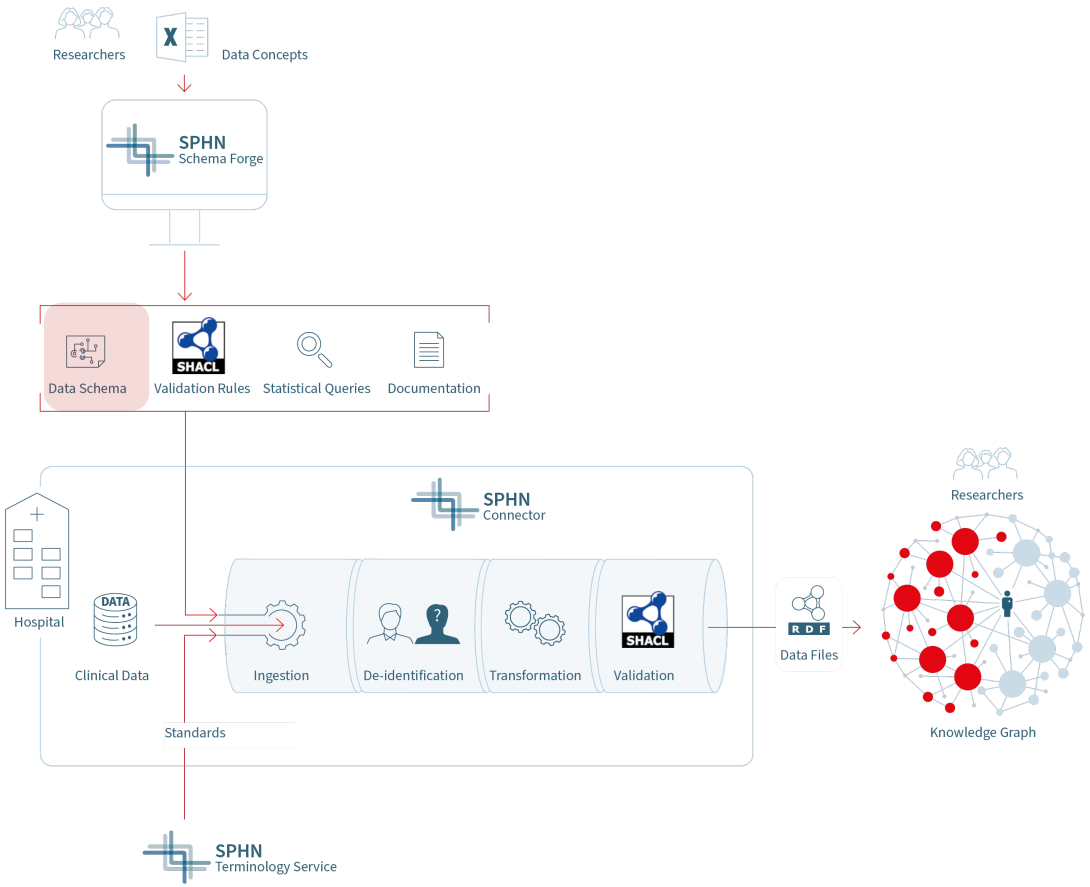


concept or concept	general concept name	general description	parent	type	excluded type descendants	standard	value set or subset
concept	**Birth**	**the event of being born**	**SPHNConcept**				
composedOf	date	date of the concept	SPHNAttributeObject	Birth Date			
composedOf	gestational age	gestational age associated to the concept	SPHNAttributeObject	Gestational Age At Birth			
composedOf	country	country associated to the concept	SPHNAttributeObject	Country			
composedOf	delivery mode code	mode of delivery during concept	hasCode	Code		SNOMED CT	descendant of: 118215003 |Delivery finding
composedOf	medical device	medical device associated to the concept	SPHNAttributeObject	Medical Device	Access Device; Lab Analyzer; Implant		

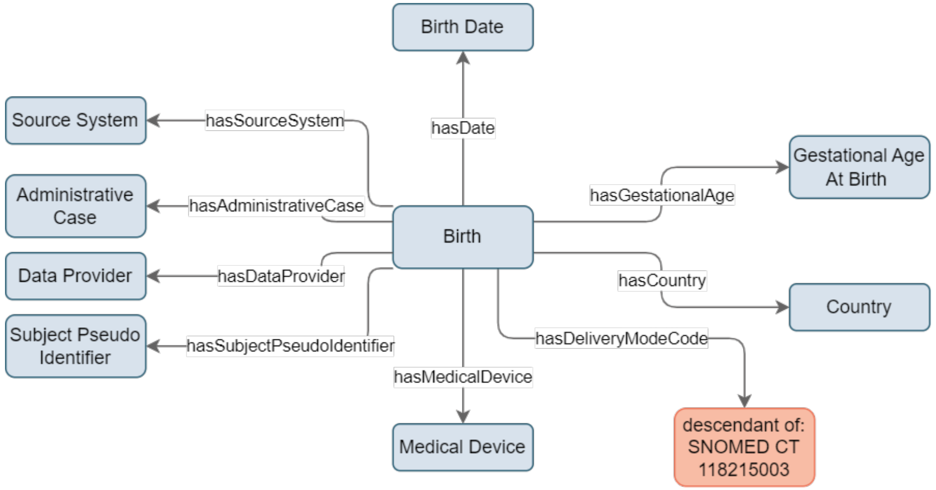


The SPHN Dataset
Semantics defined in Excel

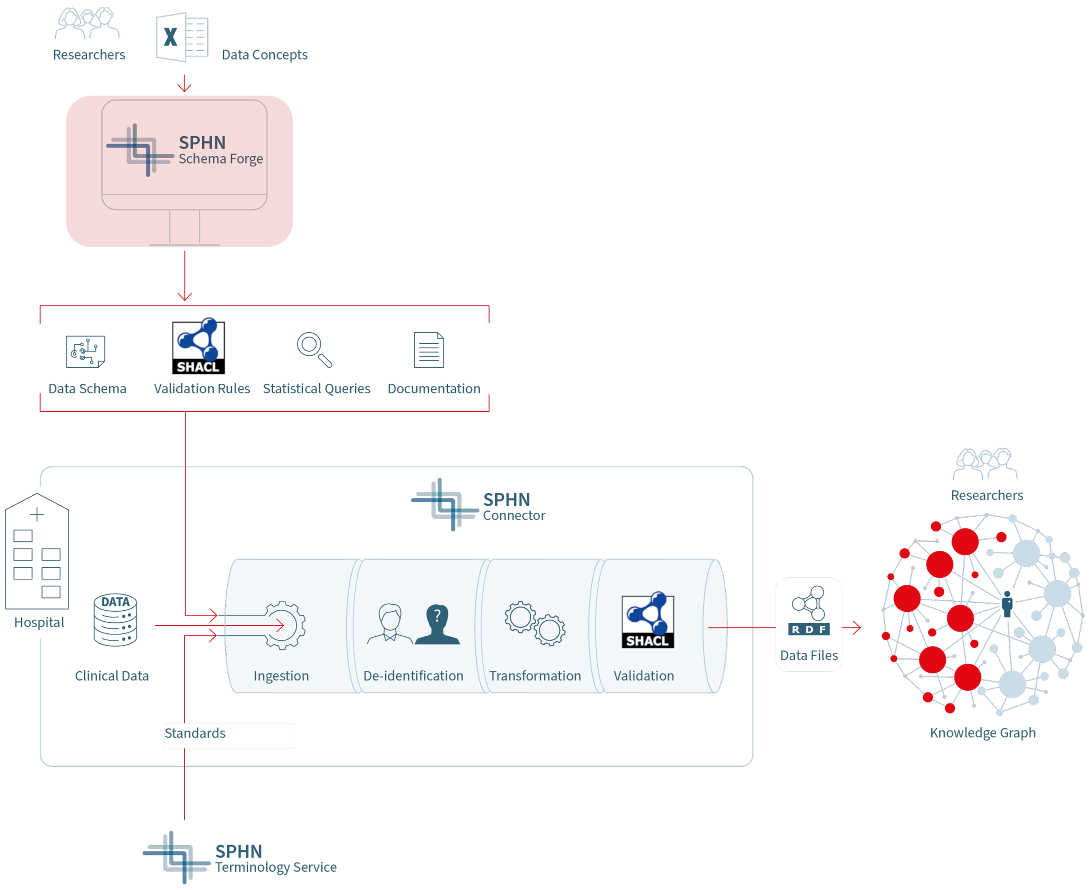
The SPHN Semantic Interoperability Framework



SPHN RDF Schema is the blueprint

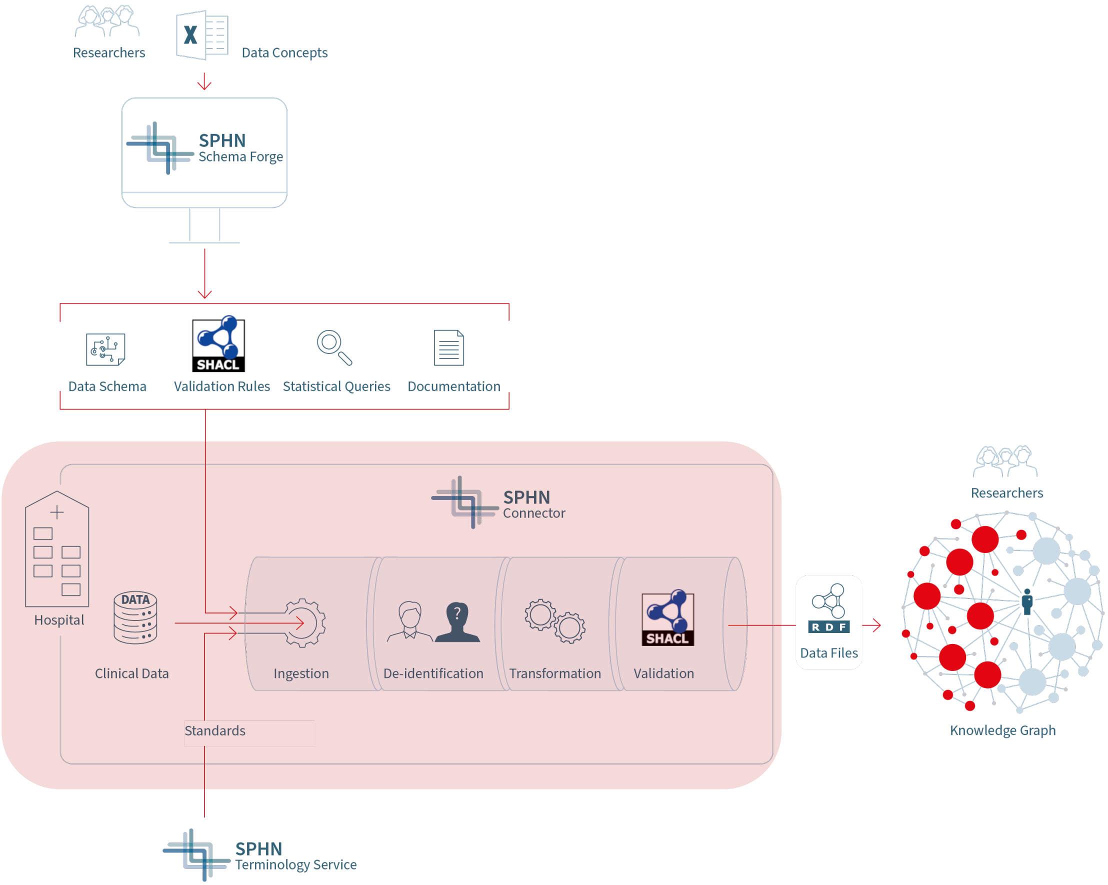


The SPHN Semantic Interoperability Framework



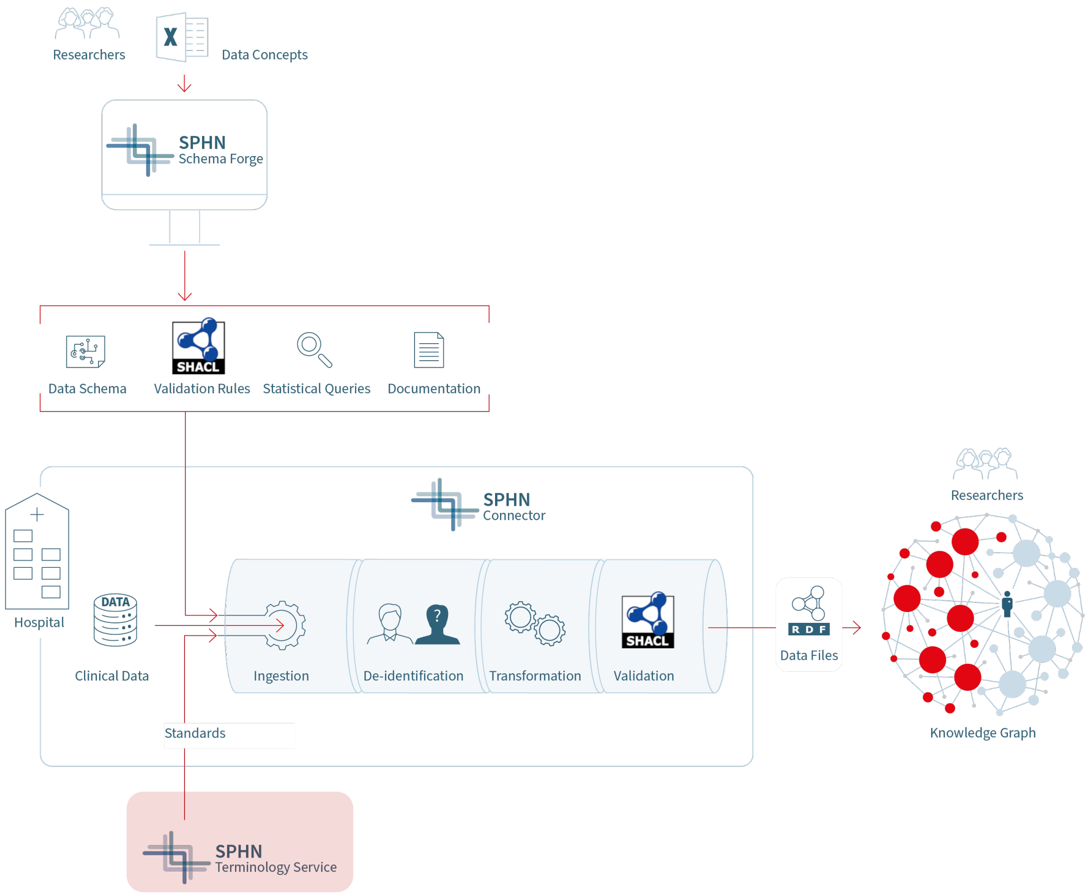
- SPHN Schema Forge builds the RDF Schema

The SPHN Semantic Interoperability Framework



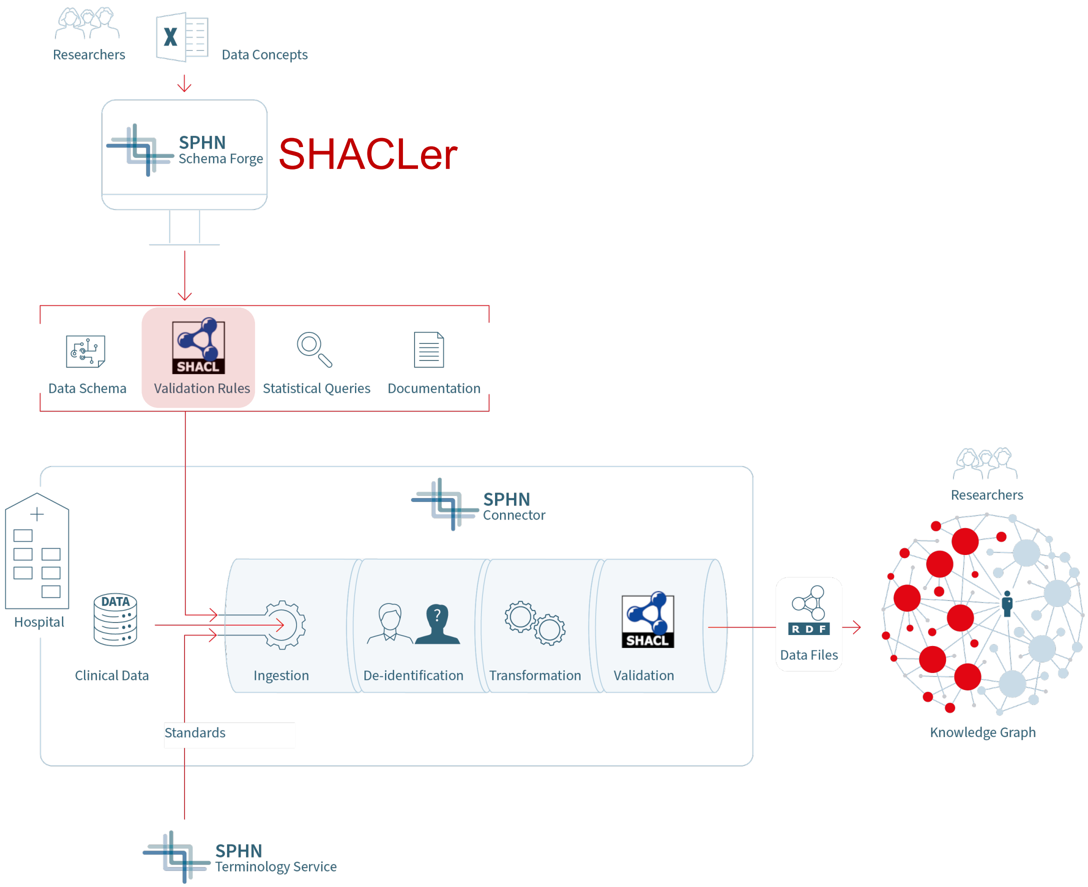
- SPHN Schema Forge builds the RDF Schema
- SPHN Connector builds validated data in RDF

The SPHN Semantic Interoperability Framework



- SPHN Schema Forge builds the RDF Schema
- SPHN Connector builds validated data in RDF
- SPHN Terminology Service generate RDF versions of external terminologies

The SHACLer – Validator for SPHN-related data



SHACL rules automatically built with the SHACLer (Python script, uses rdflib)

→ SHACLer integrated in the SPHN Schema Forge

git.dcc.sib.swiss/sphn-semantic-framework/sphn-shacl-generator

SHACLer

Why put so much effort on validation?

Many stakeholders involved, with different systems generating data

One single point of truth needed to check and validate data

Data validity & severity levels

The generated SHACL constraints (384) raise different levels of severity

ERROR (184)

Violation of the schema definitions (e.g. cardinality, value sets, ranges)

Start Datetime > End Datetime

Data validity & severity levels

The generated SHACL constraints (384) raise different levels of severity

ERROR (184)

Violation of the schema definitions (e.g. cardinality, value sets, ranges)

Start Datetime > End Datetime

WARNING (198)

Instances violating recommended naming conventions

Unversioned code instance which had a change in the meaning

Data validity & severity levels

The generated SHACL constraints (384) raise different levels of severity

ERROR (184)	WARNING (198)	INFO (2)
Violation of the schema definitions (e.g. cardinality, value sets, ranges)	Instances violating recommended naming conventions	Old versioned code which is still valid
Start Datetime > End Datetime	Unversioned code instance which had a change in the meaning	Old versioned code which is not valid anymore

General information about SHACLs

- Shapes are open for the SPHN RDF Schema but closed for project-specific schemas

- Projects are free to extend the SHACLs to add finer rules relevant to their projects (their own pipeline / manual)

SHACL rules generated with the SHACLer

Examples of SHACL rules built for data validation

Disclaimer

The SHACLs snippets in upcoming slides may be simplified for the purpose of readability

Find more here:

sphn-semantic-framework.readthedocs.io/en/latest/sphn_framework/schemaforge.html#shacler

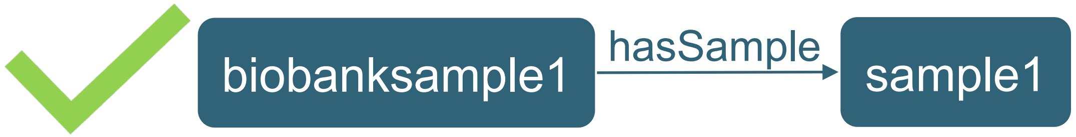
SHACLer Documentation

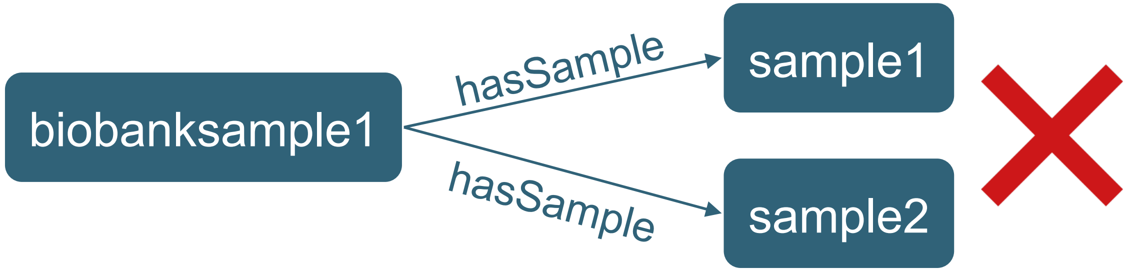
Cardinality Constraints

```
constraints:Biobanksample a sh:NodeShape ;
        sh:property [ sh:class :Sample ;
                sh:maxCount 1 ;
                sh:minCount 1 ;
                sh:path :hasSample ];
        sh:targetClass :Biobanksample .
```

Cardinality Constraints

```
constraints:Biobanksample a sh:NodeShape ;
        sh:property [ sh:class :Sample ;
                sh:maxCount 1 ;
                sh:minCount 1 ;
                sh:path :hasSample ];
        sh:targetClass :Biobanksample .
```





Class Constraints

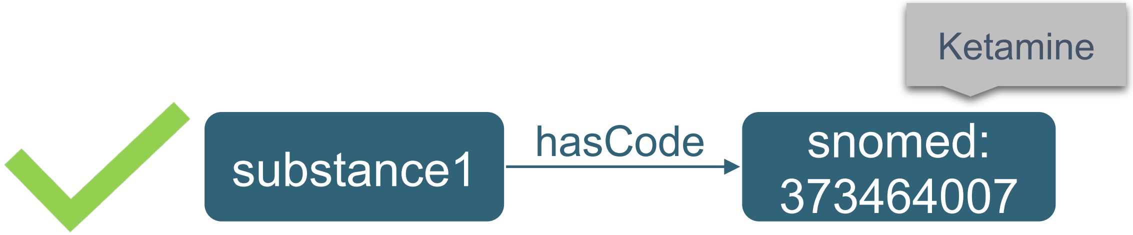
```
constraints:Substance a sh:NodeShape ;
    sh:property [
        sh:or ([ sh:class :Code ] [ sh:class sphn-atc:ATC ] [ sh:class snomed:105590001 ]);
        sh:path :hasCode ] ;
    sh:targetClass :Substance .
```

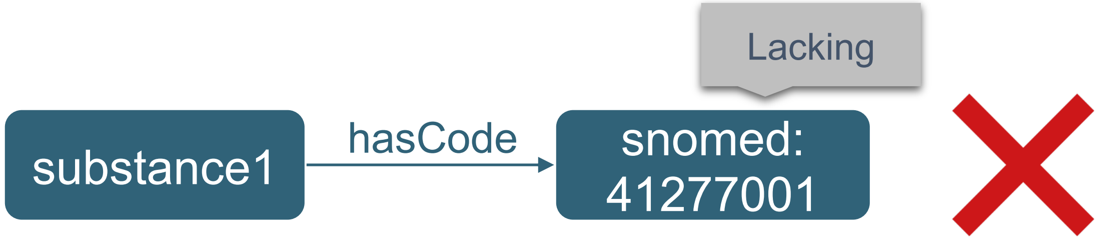
Substance

Class Constraints

```
constraints:Substance a sh:NodeShape ;
    sh:property [
        sh:or ([ sh:class :Code ] [ sh:class sphn-atc:ATC ] [ sh:class snomed:105590001 ]);
        sh:path :hasCode ] ;
    sh:targetClass :Substance .
```

Substance





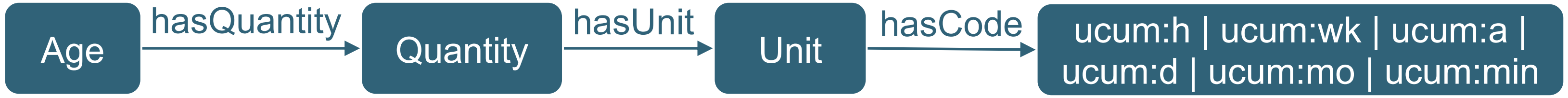
Class Constraints with SPARQL expression

skos:scopeNote *"sphn:hasCode no subclasses allowed"* gets interpreted in the SHACLer as:

```
constraints:AdministrativeSex a sh:NodeShape ;
    sh:property [sh:or (  [ sh:class snomed:261665006 ] [ sh:class snomed:703117000 ] [ sh:class snomed:74964007 ]
                          [ sh:class snomed:703118005 ] ) ;
                 sh:path :hasCode ] ;
    sh:sparql [ a sh:SPARQLConstraint ;
            sh:message "No descendents (all subclasses) of the specified codes are allowed" ;
            sh:select """SELECT ?this (sphn:hasCode> as ?path) (?class as ?value)
                          WHERE {
                                 ?this sphn:hasCode/rdf:type ?class .
                                 FILTER( ?values IN ( snomed:261665006, snomed:703117000, snomed:74964007, snomed:703118005 )) .
                                 FILTER (?class NOT IN ( ?values ) ) .
                                 FILTER NOT EXISTS { ?values rdfs:subClassOf+ ?class .}
                         }""" ] ;
    sh:targetClass :AdministrativeSex .
```

Sequence path constraints

owl:restriction:



Age hasQuantity Quantity hasUnit Unit hasCode ucum:h | ucum:wk | ucum:a | ucum:d | ucum:mo | ucum:min

Gets interpreted as:

```
constraints:Age a sh:NodeShape ;
sh:property [ sh:class :SubjectPseudoIdentifier ;
            sh:maxCount 1 ;
            sh:minCount 1 ;
            sh:path :hasSubjectPseudoIdentifier ],
            [ sh:in ( ucum:h ucum:wk ucum:a ucum:d ucum:mo ucum:min ) ;
            sh:maxCount 1 ;
            sh:minCount 1 ;
            sh:path ( :hasQuantity :hasUnit :hasCode ) ] ;
    sh:targetClass :Age .
```

Validity of old versioned codes

```
constraints:OldVersionedCodeHasBeenValid a sh:NodeShape ;
    sh:severity sh:Info ;
    sh:sparql [ a sh:SPARQLConstraint ;
            sh:message "The versioned code is not valid anymore due to code meaning change." ;
            sh:select """PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
                        PREFIX sphn: <https://biomedit.ch/rdf/sphn-schema/sphn#>
                        SELECT  ?this (rdf:type as ?path) (?type as ?value)
                        WHERE {
                            ?this rdf:type ?type .
                            }""" ] ;
    sh:target [ a sh:SPARQLTarget ;
            sh:select """PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
                        PREFIX sphn: <https://biomedit.ch/rdf/sphn-schema/sphn#>
                        PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
                        SELECT  ?this
                        WHERE {
                            ?type sphn:hasMeaningValidityInCurrent ?validity .
                            FILTER(?validity = false) .
                                     ?this rdf:type ?type .
                            }""" ] .
```

ATC 2023 vs ATC 2022
ATC 2022 vs ATC 2021
ATC 2021 vs ATC 2020

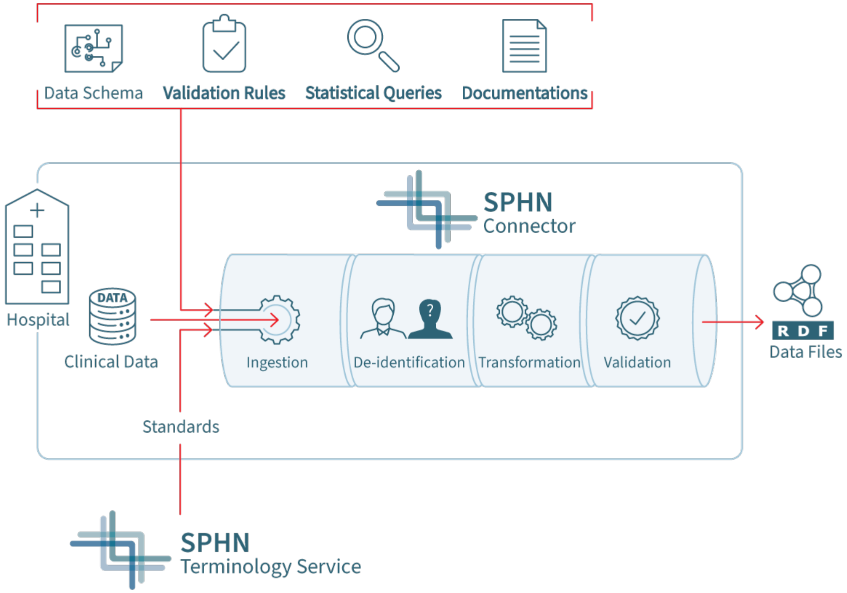
sphn_atc_2023-2016-1.ttl

Checks if an old versioned code used, which has been valid, is not valid anymore

Data validation in practice

Generated SHACLs rules are integrated in the SPHN Connector pipeline

Possibility to filter for SHACL rules based on severity levels (error, warning, info)



Ex: 270K patients (with 63K having violations)

1/3 of time is spent on validation (i.e. 2days)

Data validation in practice: performance

Benchmark on 100 mock patients:

	Structural check (Schema compliance)	Syntactic check (Naming convention)	Terminology check (Meaning change)
Mean time (10 runs)	50s	80s	1000s
Type of severity	Error	Warning	Info

Conclusion

✓ SPHN SHACLs checks:

— compliance with RDF Schema

— compliance with naming convention

— validity of codes

✗ SPHN SHACLs do not check:

— clinical correctness

Is it the end of PhD students manually validating data? Maybe not. But we are getting there...

Acknowledgements

The PHI Group:

Katrin Crameri, **Deepak Unni**, Harald Witte, Jan Armida, Judit Kiss Blind, Julia Maurer, Michaela Egli, Michael Müller-Breckenridge, Owen Appleton, Patricia Fernandez Pinilla, Petar Horki, **Sabine Österle**, Shubham Kapoor, Sergio Guarino, Simone Guzzi, Vasundra Touré

Accenture: **Philip Krauss**

The SPHN NSB and NAB, Task Forces & WGs

The BioMedIT Board and workforces @ ETHZ, Unibas, Unil/SIB

The Hospital workforces @ USZ, USB, CHUV, Insel, HUG

The SPHN Management Office:

Thos Geiger, Liselotte Selter, Sarah Vermij, Christine Remund, Gaudenz Metzger

@SPHN_ch

Vasundra.Toure@sib.swiss

dcc@sib.swiss I info@sphn.ch

www.sphn.ch I www.sib.swiss/phi

www.BioMedIT.ch

Back up slides

Foreseen improvements

- SHACL targetClass entails

```
SELECT DISTINCT ?this
WHERE { ?this rdf:type/rdfs:subClassOf* ?targetClass .
}
```

- SPARQLTarget

Why not SHACL as model instead of RDF Schema?

- Complexity of the SPHN semantics
  - Semantics are intertwined – not disjoint
  - Inheritance has challenges in SHACL
  - SHACL has some simplifications assumptions which do not fit SPHN

- Audience/Stakeholders of SPHN

Swiss Personalized Health Network

Why not use FHIR, openEHR, OMOP, XYZ format?

- Tool stack from the Semantic Web technologies
- SPHN aims to be FAIR (Findable, Accessible, Interoperable, Reusable)
- Inference, validation and analytics capabilities
- Triplestores (open source & commercial)
- Complex components can be represented and interconnected