Building an innovative, sustainable and expandable data provider architecture that generates additional value for biomedical research

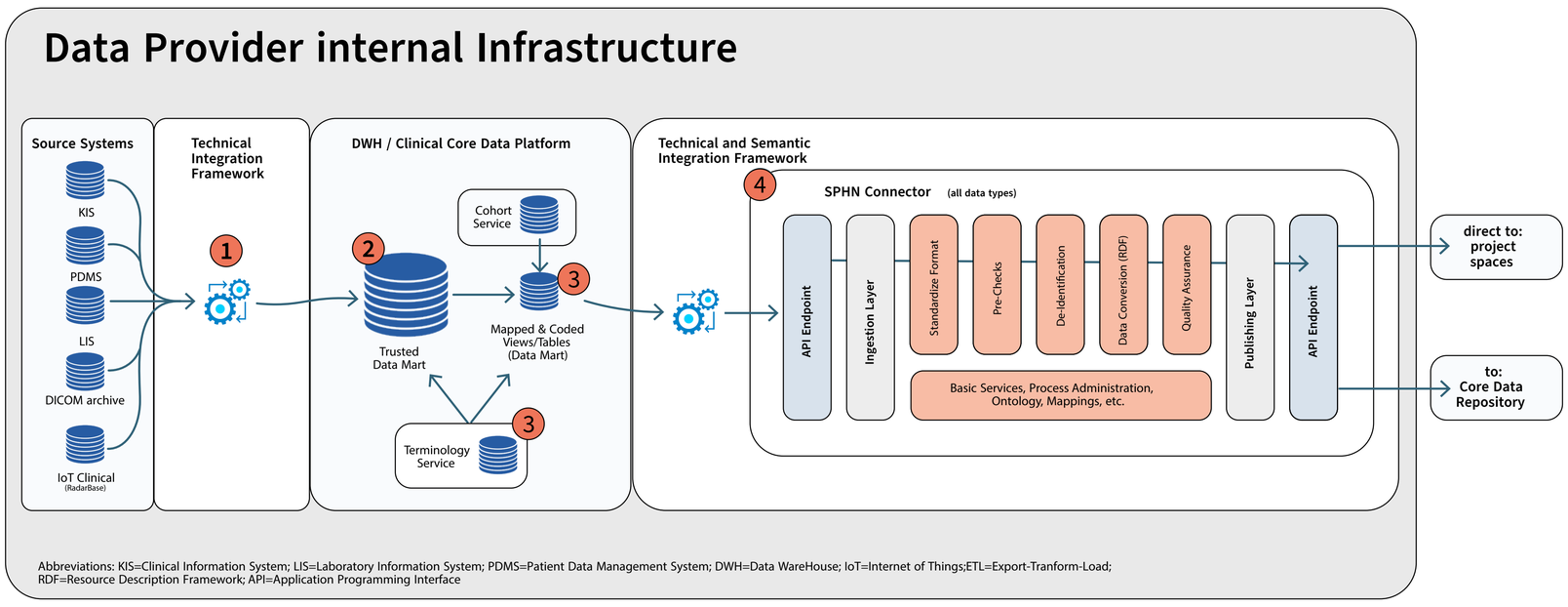
Katie Kalt[2#], Sabine Österle[1], Philip Krauss[3], Michael Müller-Breckenridge[1], Katrin Crameri[1*]
with support of the SPHN ITAC Working group members (in alphabetical order): Reinhard Armand Dietrich[6], Yves Jaggi[4], Markius Obreiter[7], Jean Louis Raisaro[4], Daniel Teixeira[5], Andreas Unterkircher[4].
[1]SIB Swiss Institute of Bioinformatics | Personalized Health Informatics Group, [2]University Hospital of Zurich, [3]Trivadis part of Accenture, [4]Centre hospitalier universitaire vaudois, [5]Hôpitaux Universitaire Genève, [6]Insel Spital, [7]University Hospital Basel, [#]contracted by SIB as lead architect, presenting author, *corresponding author: katrin.crameri@sib.swiss.

In the past, many research initiatives relied on proprietary or common ad-hoc data models such as OMOP or i2b2, and the process of acquiring data from source systems was limited to data that met the requirements of a specific research project. SPHN takes a more generic approach and develops a broad platform compliant with the SPHN Semantic Framework, to describe variously structured data types using international standards. The SPHN IT Architecture (ITAC) Working Group, which includes representation from all five university hospitals, is currently developing common IT infrastructure components to transform, validate, store and deliver data to research projects or shared services. Core components such as the "SPHN connector" or the "Research Core Data Repository" will not only enable faster provisioning of consistently high quality data, but will also allow the development of a data catalogue with information on the availability and provenance of the data. Finally, applied technologies will allow federated analysis and learning by bringing the algorithm to the data, rather than (repeatedly) transfering gigabytes of sensitive data to research projects. Healthcare facilities outside SPHN can also implement the developed infrastructure components for data exchange with limited effort. Although the infrastructure is still under development, SPHN leverages synergies with other health data initiatives and establish the architecture as a cross-organizational platform that could form the basis for a future research data ecosystem and, consequently, can be sustainably managed beyond the SPHN's funding period.

Goals of the architecture

- Development of modular and standardized components for data provisioning to research projects.
- Improvement of data provisioning processes and data quality (e.g., pipeline reproducibility, logging of procedural steps, provision of data validation solutions).
- Ensuring a common data semantics following the SPHN Interoperability
- Framework to describe different structured data types using international standards
- Reducing costs and efforts of data provisioning to research projects
- Allow a common interface across hospitals for service connectors (e.g., for a joint metadata catalogue)

Data Provider internal Infrastructure



Abbreviations: KIS=Clinical Information System; LIS=Laboratory Information System; PDMS=Patient Data Management System; DWH=Data WareHouse; IoT=Internet of Things;ETL=Export-Tranform-Load; RDF=Resource Description Framework; API=Application Programming Interface

The SPHN Connector 4

The SPHN Connector produces RDF files according to the SPHN Interoperability Framework;

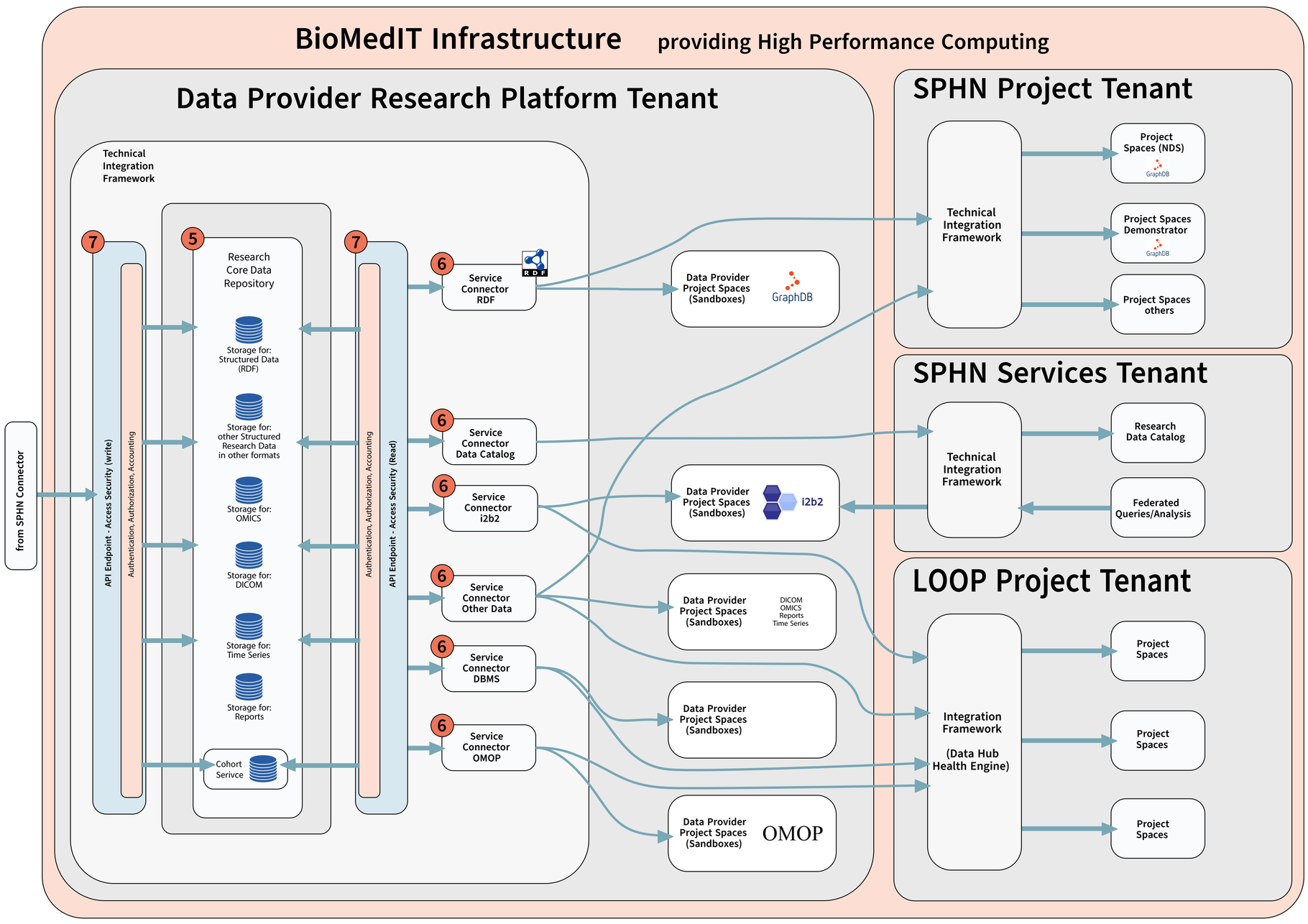
- it allows easy onboarding of new data providers without requiring them to have knowledge in RDF generation and validation
- it enables data to be streamed on a patient basis (i.e., no heavy load bulk transfer necessary)
- de-Identification of data through the SPHN Connector can be activated (in case the data provider doesn't prefer its own de-identification solution)
- it provides a standardized validation process to guarantee correct RDF data compliant with the specification
- RDF output can be transmitted directly to research projects or ingested in the Core Data Repository, which serves as data source for all research services

The steps ofthe SPHN Connector pipeline are:

- ingestion (data must be loaded into the SPHN Connctor)
- conversion of standardized ingested tabular data into JSON
- execute pre-checks to make sure data can be processed
- de-identifies data according to specific rules
- convert data into RDF turtle format
- validate the generated RDF using the SPHN Quality Control Framework
- make data available for download or send it into the Core Data Repository (or directly to project spaces

Research Core Data Repository (CDR) 5

- Stores de-identified and standardized research data for easy and fast access
- Stores links to data at data provider site available for data retrieval on request
- Omics and DICOM data are linked through metadata with other structured data
- Serves as single source for all research data requests
- Provides access via standardized and open API (7)
- Data access is protected based on authorized access



Service Connectors 6

A Service Connector provides an interface between the CDR and the end-user
Services could either:

- provide data in a specific format
- allow retrieval of metadata of structured data and other files
- allow retrieval of key performance indicator (KPI) information about data to be presented in a metadata catalogue
- any party can create its own service adapter and share it with any data provider
- access authorization, privacy and data protection requirements are safeguarded by the CDR-API

FHIR, OMOP and i2b2

- In a future version, the SPHN Connector will accept FHIR resources to reduce the efforts of the data providers even further, given that hospitals have implemented FHIR as the data exchange format in healthcare
- Common data models like i2b2 and OMOP are limited with regards to the data representation in a pure semantic way. This is why we aim to provide data for various data models from the semantic data representation

Data privacy, data protection and de-identification (pseudonymization/anonymization)

- The architecture offers data protection and privacy by design.
- All collected data will be de-identified prior to making it available to reserchers
- Having a generic service API allows data access to be managed centrally for all services.
- Every service and project will need its own DTUA and ethics approval.

Acknowledgements and References

The SPHN Connector is implemented and financed by SPHN.
The Core Data Repositories will be created in collaboration by SPHN and The LOOP.

With Partners:

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