The BioMedIT Network Project

Report 2017-2020

PHI Group, March 2021

Given the sensitive nature of health-related information, research using patient data calls for high levels of security and data protection in ICT infrastructure and processes, and requires a corresponding level of expertise in handling sensitive data to fulfil all the stringent legal, regulatory and ethical requirements. The key challenge is to provide researchers with an integrated solution that safeguards data while allowing them to carry out their research work. In close collaboration with the partner institutions, the Personalized Health Informatics (PHI) Group of SIB has setup the BioMedIT network as part of SPHN to provide all authorized researchers in Switzerland with easy access to collaborative analysis of sensitive personal data without compromising data privacy. The network is operational today at three nodes based in Basel, Zurich and Lausanne and offers a variety of central services as well as organizational and technical security measures to guarantee confidentiality, integrity, availability and resilience of the systems.

1 Background

The advent of digital transformation in health care has produced an exponential increase in the amount of information available for each patient. The use of this information in data-driven biomedical research can lead to important changes in medicine. In order to leverage the potential of health-related data through biomedical research, data science and related research fields:

- data needs to be interoperable and available to researchers in various disciplines;
- strong capabilities in clinical bioinformatics and computational service infrastructure are required in order to enable the integration and interpretation of large and rich data sets; and
- big data analyses and machine learning generate vast amounts of data and require high-performance IT infrastructure for computing and storage.

Given the sensitive nature of health-related information, data-driven and Personalized Health research requires special IT infrastructure and services, blending the concepts of

(i) security - to protect the confidentiality of the data and the privacy of research study participants;
(ii) scalability and performance - to be able to adapt to changing needs of users; and
(iii) flexibility and ease of use - to foster cutting-edge biomedical research.

Security measures for Information and Communications Technology (ICT) systems are necessary to protect confidential information from unauthorized use, modification, loss or release. Until relatively recently these requirements were not a major concern for academic computing facilities in Switzerland because those were predominantly tailored towards the handling of (non-sensitive) basic research data.

The need to securely transfer data to the destination of choice as well as to control access to data sets across research teams from different institutions and potentially also across borders has a significant impact on the architecture of IT infrastructure for researchers working on data from multiple sources (hospitals,

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

SIB Swiss Institute of Bioinformatics
Personalized Health Informatics Group (PHI)
Elisabethenstrasse 43
CH-4051 Basel

The Personalized Health Informatics Group manages
the SPHN Data Coordination Centre and the BioMedIT project

research facilities, technology centers, etc.) and operating in multidisciplinary teams. In addition, in the context of nation-wide collaborative research projects, technical interoperability between different IT infrastructure should be implemented to enable execution, with reproducible results, of data analysis workflows executed at distributed locations.

To address the above needs, the BioMedIT network project was funded by the Swiss federal government for the period of 2017-2020 within the framework of the Swiss Personalized Health Network Initiative (SPHN) and in close collaboration with the strategic focus area Personalized Health and Related Technologies (PHRT) of the Swiss Federal Institutes of Technology (ETH) domain. The goal of the BioMedIT network project is the creation and operation of a national, secure IT infrastructure to support computational biomedical research and clinical bioinformatics using confidential data. The BioMedIT network can be used by all Swiss Universities, research institutions, hospitals and other interested partners, wherever there is a need to process sensitive data.

2 Organization and functionality of the BioMedIT network

The BioMedIT network encompasses:

1. the local technical and procedural high-security BioMedIT node infrastructures (BioMedIT nodes),
2. the connection and collaboration between the BioMedIT nodes,
3. the central infrastructure components and procedural solutions provided as a central service,
4. the connections to external partner institutions (e.g. data providers, technology centers, etc.).

The BioMedIT network builds on three legally independent scientific IT competence platforms. In the realm of the BioMedIT network project, all three institutions committed to build a high performance computing infrastructure (in addition to their already existing scientific compute clusters), especially designed for sensitive (confidential) data for Personalized Health and data-driven research (the BioMedIT nodes):

- sciCORE established sciCOREmed in Basel, operated by the University of Basel,
- Core-IT established SENSA in Lausanne, operated by SIB, and
- SIS established LeonhardMed in Zurich, operated by ETHZ.

A BioMedIT node is a local or regional node that provides a secure compute and storage infrastructure for handling (securely storing, managing and processing) sensitive research data (clear text, pseudonymized or coded personal data). As an integral part of the BioMedIT network, each node performs its function within a snowflake architecture (see below) for receiving and routing data, warranting interoperability, interfacing, and collaborating with the other existing BioMedIT nodes.

The BioMedIT network is specifically designed for collaborative research projects using sensitive data, brought together from federated data sources and analyzed by multidisciplinary research teams from different institutions (project clients). The network therefore follows a hub-and-spoke organizational design in which one BioMedIT node serves as the main (destination) node, on which the data is gathered and processed, whereas the other two nodes receive the data from data providers in their proximity and route them to the destination node. This “snowflake” architecture saves many point-to-point connections and contributes to make processes more efficient, especially with regards to multiple data transfers from the same data providing institutions. Data providing institutions are connected to the network in order to enable standardized and secured sharing of sensitive research data over the BioMedIT infrastructure. The preconfigured, secured connections and established protocols for the data flow (from a data providing institution to a project space on the destination node) significantly facilitate transfers of sensitive data throughout Switzerland. Technical interoperability between the BioMedIT nodes enables sharing of data and analysis workflows within the BioMedIT network.

A common Information Security Policy applies to the entire BioMedIT network, defining the necessary organizational and technical measures to guarantee the confidentiality, integrity, availability and resilience of the systems with regard to processing of data .

The computational service infrastructure of the individual BioMedIT nodes can also be used by organizational clients, such as Swiss universities, research institutes, hospitals, service providers or any other interested partners that require a secured workspace (institutional tenant) with a higher level of security (compared to 'standard' infrastructure) in order to securely store, manage or process confidential research data.

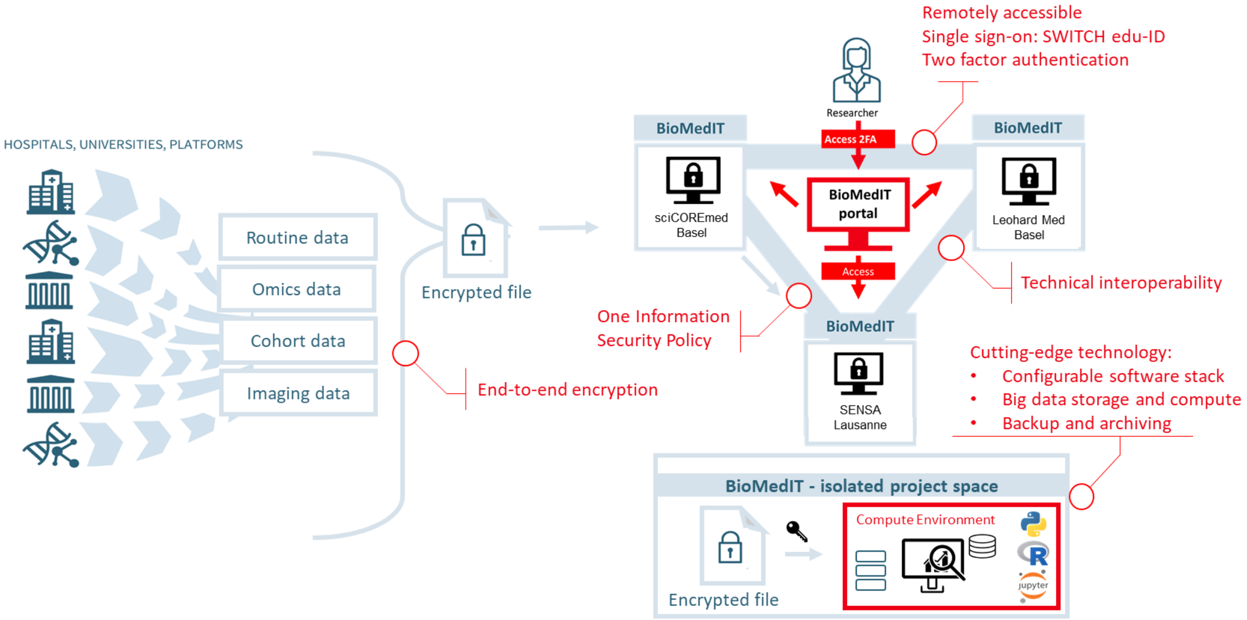


Figure 1: The components, organization, characteristics and context of the BioMedIT network, as of December 2020.

2.1 Governance of the BioMedIT network

SIB's Personalized Health Informatics (PHI) Group is responsible for central infrastructure components (tools, platforms, servers, etc.) and procedural solutions of the BioMedIT network. Under the umbrella of the SPHN Data Coordination Center (DCC), which is mandated to promote the development and implementation of nationwide semantic and technical interoperability standards, PHI operates a central service layer and is responsible for the coordination as well as the management and direction of the BioMedIT network. SIB has the overall responsibility for the BioMedIT network project and handles the finances allocated to the PHI Group, the BioMedIT network associated institutions, and additional partner organizations contributing to the project.

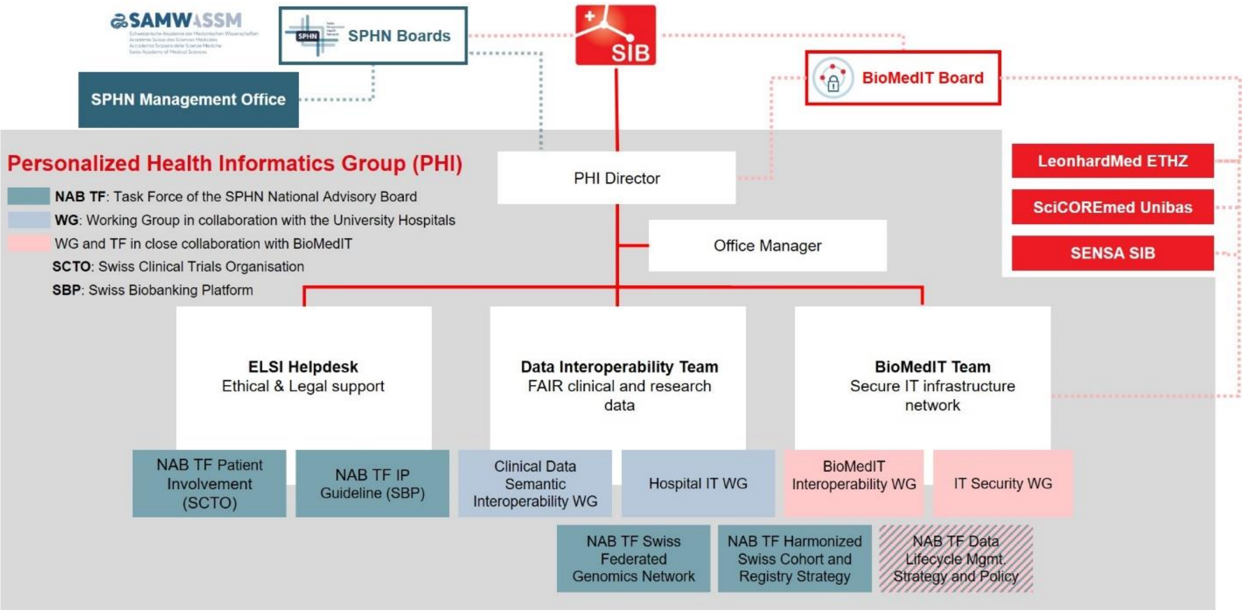

SAMW ASSM
SPHN Boards
SIB
SPHN Management Office
BioMedIT Board
Personalized Health Informatics Group (PHI)
NAB TF: Task Force of the SPHN National Advisory Board
WG: Working Group in collaboration with the University Hospitals
WG and TF in close collaboration with BioMedIT
SCTO: Swiss Clinical Trials Organisation
SBP: Swiss Biobanking Platform
PHI Director
Office Manager
LeonhardMed ETHZ
SciCOREmed Unibas
SENSA SIB
ELSI Helpdesk
Ethical & Legal support
Data Interoperability Team
FAIR clinical and research data
BioMedIT Team
Secure IT infrastructure network
NAB TF Patient Involvement (SCTO)
NAB TF IP Guideline (SBP)
Clinical Data Semantic Interoperability WG
Hospital IT WG
BioMedIT Interoperability WG
IT Security WG
NAB TF Swiss Federated Genomics Network
NAB TF Harmonized Swiss Cohort and Registry Strategy
NAB TF Data Lifecycle Mgmt. Strategy and Policy

Figure 2: The BioMedIT network project is an integral part of SPHN and managed by the Personalized Health Informatics (PHI) Group within the assignment of the SPHN Data Coordination Center (DCC).

The BioMedIT board is the strategic and executive body of the BioMedIT network project. It includes representatives from the BioMedIT nodes, associated institutions as well as representatives from SIB's executive management. Further, technical interoperability in BioMedIT is driven by the BioMedIT interoperability working group (BIWG). The BIWG is composed of employees of the BioMedIT nodes and chaired by a Lead Developer, who is affiliated with the PHI group. The IT security working group addresses IT security and privacy issues specifically relevant in the context of the BioMedIT project, and is composed of representatives of the node institutions and SIB.

3 Achievements of BioMedIT in phase I

In the first phase, the focus of the BioMedIT network project was on the establishment of the local high-security BioMedIT node infrastructure, the connection and coordination between the nodes, the connections to the data providing institutions, and the security architecture of the entire network. A first paper on the advancements of BioMedIT has been published in June 2020. In addition to a selection of central services provided by PHI, two essential BioMedIT tools have been developed by the BIWG: The Secure Encryption and Transfer Tool (sett) for the transport of sensitive data, and the BioMedIT portal – the entry point for researchers to manage and access their projects.

In the summer of 2020 BioMedIT and SPHN reached an important milestone: the first sets of pseudonymized patient data for the SPHN funded Personalized Swiss Sepsis Study were successfully transferred over the BioMedIT network; initially between the two university hospitals in Geneva and Lausanne and the project work space hosted at SIS (ETHZ) in Zurich, and subsequently from the other three university hospitals to ETHZ.

At the end of 2020, there were more than one dozen research projects in preparation or already running on the three BioMedIT nodes (see individual node reports for detailed information). In addition, a longitudinal cohort study of national importance has been in the process of being transferred to the BioMedIT network.

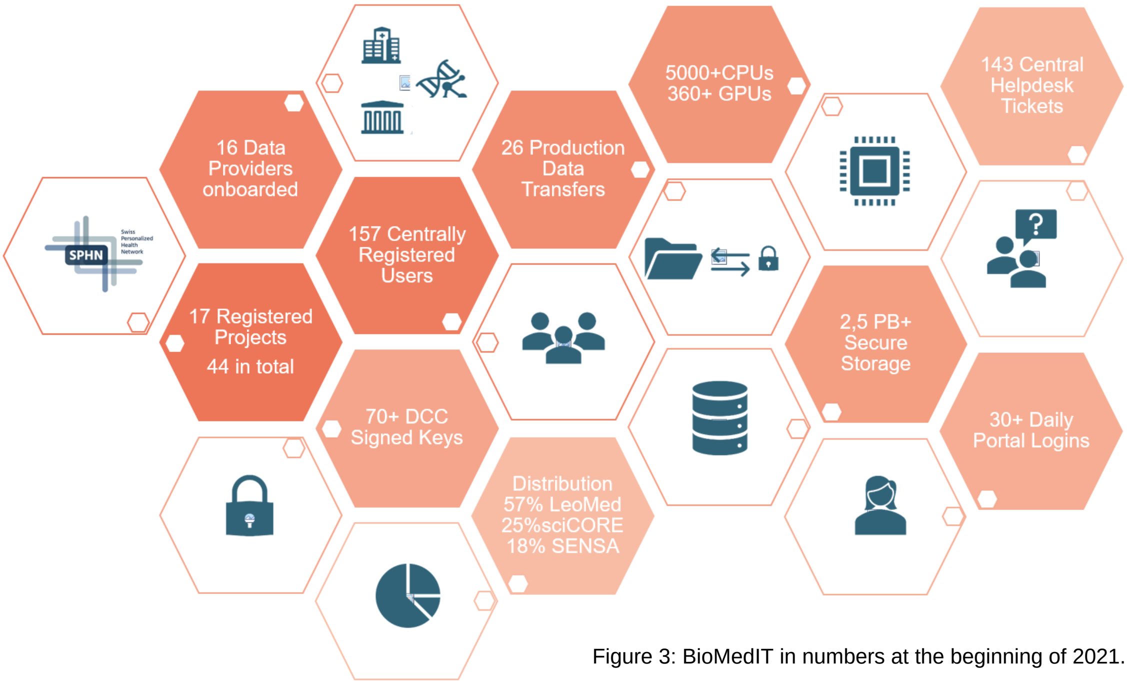


Figure 3: BioMedIT in numbers at the beginning of 2021.

3.1 Infrastructure as a Service

The BioMedIT nodes provide a broad range of secure services to store, manage, process and share biomedical data. They offer state-of-the-art software for data science and specialized tools for data management together with long-standing expertise in scientific IT support for research data management, bioinformatics, HPC and computational analysis. Software, storage and compute capacities are tailored to the needs of individual projects. In mid-2020, the BioMedIT infrastructure featured data storage capacity approaching in total 5 Petabytes, support for data encryption, secure backup, private-cloud and high-performance computing (HPC) leveraging more than 3000 CPU cores and 324 GPUs. Researchers authorized to use the secure BioMedIT network can access the infrastructure only from within trusted IT environments (either from within Swiss university or university hospital networks or via VPN to one of these networks). Depending on the use-case, researchers access project workspaces in the BioMedIT network using the command-line in a terminal or web-based remote desktop technology. By linking to the BioMedIT network, Swiss research institutions and hospitals benefit from a trusted research IT infrastructure without the need to (re)develop costly in-house infrastructure and know-how.

3.2 Security architecture

The three BioMedIT nodes share a common security architecture to transfer, store, manage, analyze and share sensitive (biomedical) data while following the latest technical and legal standards required by Swiss legislation as well as internal regulations of the associated institutions. All BioMedIT nodes comply with the requirements of the SPHN/BioMedIT Information Security Policy. The policy, in conjunction with an Access and Use Policy (AUPs, to be signed by BioMedIT users), related Standard Operating Procedures (SOPs) and Work Instructions (WI), define the necessary organizational and technical measures to guarantee the confidentiality, integrity, availability and resilience of the systems with regard to processing of data. These documents set out the way research data is handled, preventing misuse and malicious damage.

Data security in the BioMedIT nodes is principally based on the concept of allocation of isolated tenants. Private tenancy ensures that data stored on one project cannot be shared – intentionally or by accident – with another project. Shared tenancy, where data is shared between multiple projects, is only permitted in those cases where there is a specific authorization.

Users securely connect to project spaces via the BioMedIT portal. Access to the Internet or other networks outside the BioMedIT project space is strictly controlled and limited to trusted and explicitly white-listed resources.

Encrypted backups of the project data are done on a regular basis.

3.3 Onboarding of 16 national data providers

All providers of data to projects in the BioMedIT network are linked to a single BioMedIT node, irrespective of where the project is hosted. This “snowflake” design was chosen to minimize the amount of onboarding efforts needed for each data provider. Data providing institutions are securely connected to the network through both, administrative and technical measures, to enable secured sharing of sensitive research data over the BioMedIT infrastructure. Each Data Provider has one landing zone at a particular node to where encrypted and signed data packages are sent, generally via Secure File Transfer Protocol (SFTP) from whitelisted IP addresses. The same method is used for data transfers which take place internally between the BioMedIT nodes.

Upon data transfers, the DCC and the partner BioMedIT node provide the necessary information in order to allow the establishment of data transfer connection between the data provider and the partner node. The data providing institution is responsible for the installation of the required software (see next section), the generation of keys used for the encryption and decryption of data and those ones needed to connect to the secure servers in the BioMedIT node, the transfer of the encrypted data as well as readiness with regards to the network settings at the data provider side. Every public key generated to be used in the process for data transfer is further verified and signed by the DCC to confirm the identity of the user and the key integrity. The partner BioMedIT node presents the data provider with further information for sending data into the BioMedIT network.

All the above tasks are checked and verified by the DCC and the partner BioMedIT node. As of end of March 2021, a total of 16 institutions have been successfully on-boarded to the BioMedIT network.

3.4 Secure and standardized data transfers

Transferring sensitive personal data from a Data Provider (e.g. from a clinical data management system of a University Hospital) into the BioMedIT network requires data to be encrypted and sent using secure and standardized methods. Within the BioMedIT network, data transfers are carried out in a “snowflake” manner, the BioMedIT node hosting the project serves as the main (destination) node on which the data is gathered and processed while the other two (transfer) nodes receive the data from data providers in their proximity and route the data to the destination node.

Institutions that act as data providers are securely connected to the BioMedIT network via a landing zone to where encrypted and signed data packages are sent. Transfers are carried out using Secure File Transfer Protocol (SFTP) and connection to the landing zone server may only be made from whitelisted IP addresses. The same method is used for data transfers which take place internally between the BioMedIT nodes.

Coordinated by PHI, the BIWG developed and maintains **sett (Secure Encryption and Transfer Tool)**, an easy-to-use, open source tool to support the full process of complex data packaging and secure data transfer with both a graphical user interface (GUI) and a command line interface (CLI). sett is a wrapper around standard packages such as GPG, tar, SFTP & Liquid files and supports four modules for key management, data packaging and encryption, data transfer, and data decryption and unpacking. The tool includes a range of validity, integrity and error checks, which are all carried out transparently – with minimal user input required, making sett quite intuitive to use. Moreover, all operations are logged and stored locally on the user's computer. With the release of version 1.0 in the beginning of 2020, the tool was made available for public use, also outside of BioMedIT. The sett user documentation brings together all relevant information for data providers and data recipients. The codebase and the data packaging specification are made public on gitlab.com. As of March 2021, sett has registered 447 downloads.

More functionalities, such as efficient data packaging for large data sets, and sett tool benchmarking with various data providers (Swiss University Hospitals and Technology centers) are planned for the course of 2021.

3.5 Accessibility of the network to users

The **BioMedIT portal** is the one stop-entry to the BioMedIT network. To gain access to the BioMedIT portal and all three BioMedIT nodes, users must create a SWITCH edu-ID account, a persistent identity reinforced by two-factor authentication that enables access to all federated services.

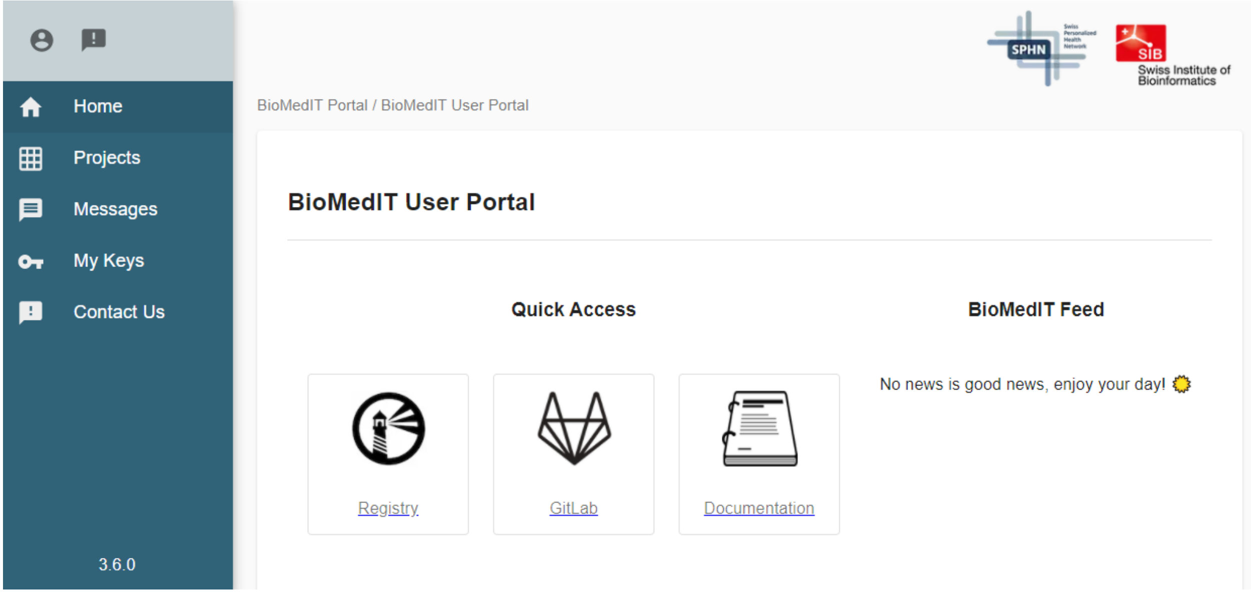
Access to the portal and many other resources is planned via the future biomedit.ch landing page (to be launched in Q2 2021), a site open to the general public.

Using the BioMedIT portal:

- Researchers have an overview of their BioMedIT projects resources, i.e. their project environments and git repositories;
- Researchers can access their projects via a virtual desktop or a virtual terminal session from the portal;
- Researchers can view and manage their GPG keys (required for encryption of data) registered with the BioMedIT key server;
- Data managers can view the status of current data transfers or request a new data transfer to one of their projects on BioMedIT;
- Data providers can get notifications for their data transfers;
- Project leads can manage access and permissions to project members over the portal.

At the beginning of 2021, 17 SPHN/PHRT projects, 26 data transfer requests and 157 users were registered in the BioMedIT portal (see Figure 3 in page 5).

More functionalities such as segregated views on data transfers, integration of new resources, access to the Federated Query System, a Container registry, etc. are planned for the course of 2021. The portal is developed as an open source tool that is currently also used in two of the BioMedIT nodes for their local project (i.e. projects outside of SPHN or PHRT that are mostly run by local research groups of the home institutions of the nodes).



No news is good news, enjoy your day!

Figure 4: Snippet of BioMedIT portal – one sign-in, one-stop entry for resources and projects on the BioMedIT Network.

3.6 Central services

BioMedIT provides a central **key server** where public GPG keys of BioMedIT users can be stored. Public GPG keys are required to encrypt and sign data during secure data transfers in the BioMedIT network. To increase the level of trust, all public keys are verified and approved ("signed") by the SPHN Data Coordination Center (DCC) before their use. The key server is based on the open-source community SKS-key server technology. A comprehensive user guide on key management is also available online. As of March 2021, the BioMedIT key server hosts 83 public GPG keys of BioMedIT users.

BioMedIT **WebProtégé** provides an easy-to-use ontology development environment to collaborate and share with other BioMedIT users. It currently supports editing of OWL 2 and OBD terminologies and is based on an open-source application developed by Stanford University. BioMedIT WebProtégé enables users to access it through their web browsers without the need to download or install any software. The service was launched on 11/2020 and has 24 registered users (as of March 2021). Developments such as integrating the service to the BioMedIT portal and the user guide are planned for 2021.

BioMedIT provides its registered users a BioMedIT **code repository service** (based on Git). Using this service, BioMedIT users can create, collaborate and share their application codes with other BioMedIT users. The service is based upon the GitLab community edition and supports a federated SWITCH edu-ID login needing two-factor authentication. The service was launched in September 2018 and has 158 registered users (as of March 2021).

3.7 National processes, documentation and training

BioMedIT hosts a full repository of documentation on the BioMedIT wiki space (confluence). A special task force with representatives of all three nodes and the PHI Group has elaborated and consolidated the first set of processes relevant for the smooth running of the BioMedIT network procedures (such as Data Transfer, User Management, Data Provider Management, Project Lifecycle), and a documentation specialist has documented the processes in the form of Standard Operating Procedures (SOPs). In addition, a number of work instructions (e.g. for key signing, for data transfers, portal user management), user guides as well as technical reference documents have been produced.

At the moment, the two SOPs for Data Transfers and User Management have been approved, released and applied. The challenge regarding the consolidated definition and implementation of national processes is to overcome (or overrule) existing local practices that – for a successful and efficient service – need to be streamlined and adapted accordingly. Often, the processes have to be designed according to the node with the strictest or most complex restrictions (influenced by the home institution policies of the nodes), making those processes less lean than aspired.

BioMedIT offers the SPHN/BioMedIT Data Privacy and IT Security Training course. In this online course users are introduced to data privacy regulations, relevant laws and aspects of information security, all important considerations when dealing with confidential human data.

3.8 Partners of the BioMedIT network

In order to further test-drive, leverage and develop the SPHN infrastructure from a clinical research perspective, SPHN teamed up with the Personalized Health and Related Technologies (PHRT) program of the ETH Domain. PHRT's goals include improving the efficiency and quality of personalized health and precision medicine by providing a range of individual diagnostic and therapeutic strategies for patients.

In addition to joint calls for proposals by SPHN and PHRT Switzerland – which led to projects partly funded by SPHN (for infrastructure building) and PHRT (for the research part) – BioMedIT also allocated funds to two of the PHRT technology centers: The Health 2030 Genome Center and the Clinical Proteotype Analysis Center at ETHZ. These centers are the basis of a Swiss multi-omics pipeline. Using the same clinical biospecimens, data is generated and analyzed in a coordinated manner on each of the genomic, transcriptomic and proteotype levels.

The funding by BioMedIT has allowed the implementation of:

a) interoperable data management systems;
b) processes to ensure data quality;
c) development of robust, portable, and reproducible analysis workflows suitable for application in personalized health research projects;
d) the assurance of interoperability between PHRT platforms and SPHN projects.

The ultimate goal of this cooperation is to directly connect the Swiss multi-omics pipeline via BioMedIT to hospital-based clinical data from the clinical data management systems of the hospitals (Clinical Data Warehouses), maintaining a high level of data security while allowing interoperability between basic science and clinical science.

3.9 BioMedIT (co-)funded projects

In contrast to SPHN, which is funding projects to develop and test new technologies, methods and infrastructures at single or joint sites and making them available to other institutions after proof of concept, BioMedIT is co-funding infrastructure development projects of national importance.

During the first phase, in accordance with the wishes of the SPHN National Steering Board and the approval of the BioMedIT board, the following projects were co-funded by BioMedIT:

- The SVIP-O project establishes the Swiss Variant Interpretation Platform for Oncology.
- The IDEAL project develops an open-source software, enabling management for identifying data of patients included in clinical trials, cohorts or registries at the interface with Clinical Data Warehouses in university hospitals.
- The Beacon and SchemaBlocks development and implementation within the framework of SPHN is a driver project of the Global Alliance for Genomics and Health (GA4GH).

3.10 Funding streams and expenditures in phase I (2017 – 2020)

The BioMedIT network project has been provided with close to CHF 17’900’000 from SERI for phase I. As outlined above, BioMedIT funds have been distributed to the following recipients:

- the three BioMedIT Nodes (sciCORE, SIS, Core-IT),
- SIB Central Services (in Lausanne),
- SIB PHI Group (in Basel),
- PHRT technology partners (the Health 2030 Genome Center and the Clinical Proteotype Analysis Center at ETHZ),
- external project partners.

Due to a late start of BioMedIT, not all funds provided by the SERI for phase I could be spent in the first four years. A total of approximately CHF 6’200’000 was transferred to phase II as provisions with specific allocation. For the above mentioned funding streams, the BioMedIT network project has spent a total of CHF 11’700’000. The individual allocations and expenditures are outlined in table 1.

4 Outlook to phase II (2021-2024)

The underlying BioMedIT nodes and the main components of the BioMedIT network as well as the necessary central services are functional today and can be used for data transfers and research within the framework of SPHN and PHRT. Still, some areas lack functionality, automation, technical components, and specific offerings for use-cases with non-clinical or special clinical data types.

Therefore, the following challenges and issues will be addressed in phase II:

- Availability of state-of-the-art data science and bioinformatics tools and services for regular as well as special use-cases (e.g. multi-omics projects, imaging projects, etc.);
- Support for FAIR data management, data integration, and a semantic web strategy (e.g. RDF) as an innovative solution for the entire data life cycle;
- Compatibility of (special) project requirements with security standards;

- Information security standards: clarify the responsibility and accountability of the affiliated institutions with regard to the security level required for different data types and processes;
- Further advance the BioMedIT portal for the benefit of PIs and users (e.g. by adding dashboard functions, link addition, etc.); addition of resources, integration of (internal and external) services;
- Re-use of existing data or data sets in BioMedIT: FAIR data repositories and respective processes, guidelines and policies;
- Scalability and sustainability plan for BioMedIT after 2024;
- Professionalization of products and services;
- Automation of processes, automation of selection of services (e.g. a user can choose the necessary software and tools for interoperable and reproducible computational workflows without significant time lags);
- International access to BioMedIT;
- Governance: framework agreements for the use of the BioMedIT network with the associated partner institutions; harmonization of processes and (legal) documents between the nodes where useful and feasible; together with SPHN, define the process for re-use and safeguarding of data on BioMedIT;
- Implement as far as possible uniform standards and processes for local and national projects; integration of current node specific functions into the national context where possible and of benefit to the wider BioMedIT network environment;
- Provision of a source of truth for all information on BioMedIT (website), and offer online training materials for a broad range of topics and tools.

The overarching goal of the second phase of the BioMedIT network project will be to establish BioMedIT as the national, fully-managed infrastructure platform for sensitive research data in Switzerland. The platform will be based on modern large-scale computing advances and will make state-of-the-art data analysis tools accessible to the Swiss research community. The project administration of collaborative research projects on a national scale will be further shifted from the BioMedIT nodes to the BioMedIT portal, and the federated compute and storage resources provided by the BioMedIT nodes will be presented as one cloud resource to their users, with a central access and project management point (the BioMedIT portal), a unified account across all cloud sites (SWITCH edu-ID) and a simple and swift project setup across the federation.

The BioMedIT Roadmap 2021-2024 for the further development of BioMedIT will be available in August 2021.

BioMedIT services – a secure IT network for the responsible processing of health-related data in Switzerland. What can we do for you?

Table 1: Funding streams and expenditures of BioMedIT in 2017-2020

Funding Streams	Balance
Node Funding (sciCORE, SIS, Core-IT) for the following deliverables: – Provide at least 100 TB of secure storage and at least 500 CPU cores of state-of-the-art computing capability – Fulfil the security standards outlined in the SPHN Information Security Policy – Provide research projects with highly secure project spaces – Follow all relevant Standard Operating Procedures of BioMedIT – Support snowflake architecture for data transfer and routing to other BioMedIT Nodes – Establish and maintain direct networking connections with regional university hospital(s) and other data providers	CHF 2’800’000
Additional Security Milestone to secure data at rest: 1. Provide mechanisms to users for encrypting data on the node's on-line storage system. The encryption environment enables use of project-specific keys and, when applicable, the associated key management systems. Provisioning of the encryption environment can be a built-in solution (e.g. storage vendor-provided software) or a solution providing an equivalent level of encryption-based security (CHF 300’000 per node) 2. Reimbursement of costs for professional penetration testing within the first phase of BioMedIT (ceiling limit of expenses per node: CHF 40’000). The report of the test is provided to the BioMedIT board. 3. Establishment of a dataset inventory (also referred to as ‘records of processing activity’) which a BioMedIT node maintains in its function as a Processor in accordance with the requirements of GDPR’s article 30 (CHF 10’000 per node).	CHF 1’050’000
BioMedIT Interoperability WG (2019-2021): 1 FTE per node per year for three years, starting in 2019	CHF 1’350’000
Provision 2021-24: Security Officers: 1 FTE per node per year for four years, starting in 2021	CHF 2’400’000
External Funding – Co-funding of the PHRT Technology Centres (the Health 2030 Genome Center and the Clinical Proteotype Analysis Center at ETHZ), for (i) the implementation of interoperable data management systems, (ii) processes to ensure data quality, (iii) development of robust, portable, and reproducible analysis workflows suitable for application in personalized health research projects, and (iv) the assurance of interoperability between PHRT platforms, SPHN projects and BioMedIT (CHF 900’000 per center, timeline: 2018/2019-2021); – Co-funding of the SVIP-O project to establish the Swiss Variant Interpretation Platform for Oncology, aiming at providing a centralized, joint and curated database for clinical somatic variants coming from Swiss hospitals and related institutions, allowing (i) to assist clinicians in their task of interpreting variants in a harmonized way, but also (ii) to provide an unparalleled source of data for research, with well-characterized and clinically validated annotations on variants detected in patients (CHF 949’000, timeline: 2018-2021); – Funding of the IDEAL project to develop an open-source software system that is freely available to Swiss university hospitals, enabling to manage identifying data of patients included in clinical trials, cohorts or registries in a uniform and interoperable way (at the interface with the Clinical Data Warehouses) in each hospital (CHF 500’000, timeline: 2020-2021); – Financial support to Beacon and SchemaBlocks development and implementation in the framework of SPHN being a driver project of the Global Alliance for Genomics and Health (GA4GH) (CHF 125’000, timeline: 2020-2021).	CHF 3’374’000
Provision 2021-24: Metadata catalogue activities, EGA or equivalent	CHF 2’280’930
Provision 2021-24: RDF/data integration node support	CHF 1’800’000
Central Services – PHI: covering personnel, consultancy and running costs (CHF 1’600’000 for four years) – SIB: Support Services covering work delivered by Core-IT, Executive Management, the Legal, Human Resources, Finance, and Communication departments, and IT Security (CHF 1’000’000 for four years)	CHF 2’677’520
Total	**CHF 17’732’450**

Annex 1: Report of the sciCORE node

The sciCOREmed environment is functional and hosts productive research projects since Spring 2020. It adheres to the national mandate in terms of IT security (SPHN Information Security Policy) and identity management (platform accessible to any researcher with a Switch EduID identity).

In the period 2017-2020, sciCORE has contributed actively in to the establishment of the BioMedIT mandate, by attending and/or hosting various BioMedIT workshops on the design and implementation of the infrastructure, both at the node level and at the national interoperability level (e.g. technical specification of the data "snowflake"). sciCORE has contributed to the SPHN Information Security Policy as well as to multiple documents specifying operational aspects of the federated service. Collaborators at sciCORE have contributed in the preparation of the training material for usage of the BioMedIT environment, as well as in actually providing lectures at multiple sites (Zürich, Basel, Lausanne) and online.

In the second half of 2017 and early 2018 various technical options for the implementation of the local BioMedIT node were evaluated, considering classical HPC service management and two cloud technologies. The implementation strategy selected for sciCOREmed is a local cloud based on the OpenStack technology. A consulting company (StackHPC, UK) was hired to support the establishment of the implementation plan. From Summer 2018 to December 2019 the core IT infrastructure of sciCOREmed was gradually established, including training of local staff on the OpenStack technology. Several workshops with StackHPC and in which other BioMedIT nodes were invited were organized to share know-how. The design of the infrastructure was consolidated in Autumn 2018 and the hardware was installed in early December 2018. During 2019, the infrastructure has been configured, and different solutions were evaluated for the back-end management systems, in particular as regards auditing and monitoring for which the system Wazuh was finally selected. Proof-of-concept usage of the infrastructure has been conducted in late Fall 2019 (with projects from SwissTPH and Kantonspital St-Gallen). In early 2020, productive-level services were deployed including the virtual desktop environment (Guacamole). In this environment projects are managed as OpenStack tenants subject to virtual network isolation. Access to sciCOREmed tenants via the BioMedIT national portal using 2-factor authentication-enabled SwitchEduID identity is fully functional. The first productive project in this new environment has been an urgent COVID-19 research project (ILGE), a collaboration between ETH and Universitätspital Basel (USB). The ILGE project is de facto a spin-off of the SPHN PSSS project and reused largely the legal framework developed in PSSS. sciCORE has also been active in the BioMedIT security working group, contributing to documents and review processes (e.g. penetration testing).

In the period 2018-2020, sciCORE has been active in providing the technical perspective in the preparation and review of contract templates for SPHN collaborative projects. sciCORE collaborators have contributed to various SPHN working groups, including the Data Lifecycle Management and the Cohort and Registry Strategy task forces.

sciCORE is also supporting the central services of the Data Coordination Center since 2017, by hosting e.g. the BioMedIT project management suite (Atlassian wiki, task manager, gitlab), the metadata query system Clinerion, as well as a series of technical services (e.g. ID management (keycloack), BioMedIT portal, etc). sciCORE has actively supported technology evaluation of potential new SPHN technologies (e.g. distributed computing in the European collaboration of the Personalized Health Train) and was often pioneering the adoption of the national services in production.

In the 2nd half of 2020, several SPHN and non-SPHN projects have started being productive in the sciCOREmed environment, with a ramp-up of the number of supported projects planned at the beginning of the next funding period.

Expected challenges for the next period will be in the refinement of management processes (including cost-efficiency considerations), improved usability of the infrastructure, improved data management support, increased interoperability of computing resources (distributed computing), support to another OS than Linux, etc.

Annex 2: Report of the SIS node

Note: An extended report of the SIS node is available on request at the PHI Group

Overview

The BioMedIT node Zurich represented by the Scientific IT Services (SIS) of ETH Zurich, provides and operates the Leonhard Med platform and its services for secure transfer, storage, management and analysis of confidential research data at ETH, and at a national level for the SPHN and PHRT programs. Leonhard Med has been used in production since January 2018 and it is currently hosting securely 26 research project spaces with authorized access for more than 200 users. The achievements of the SIS node during the reporting period 2017-2020 are presented below.

Leonhard Med Security Architecture and Policies

The Leonhard Med security architecture was designed and developed to fulfill the current technical and legal standards required by Swiss legislation. In brief, project spaces in Leonhard Med are strictly separated (i.e., multi-tenant project separation concept with two levels of security (“elevated” and “high”), access of authorized users requires two-factor authentication (2 FA), access to the Internet from Leonhard Med is strictly controlled, secure data transfer procedures are required, data are securely stored and backed up, user activities on the system are logged and monitored. For compliance, all Leonhard Med users must abide by the Leonhard Med Acceptable Use Policy and all computer administrators working within the Leonhard Med platform must abide by the IT Security Guidelines for the Administration of Computer Systems within Leonhard Med. Leonhard Med follows the frameworks and standards for IT service management and information security management established at the IT Services of ETH, which are ISO/IEC 20000-1 and ISO/IEC 27001 certified and is in progress of structuring systematically the service processes following the FitSM framework.

Leonhard Med Platform Technical Specifications and Services

Leonhard Med is a multi-tenant high performance computing platform, with command line and remote desktop user access, virtualization infrastructure available and with active developments for private cloud services. To date, the Leonhard Med features a data storage capacity of more than 2000 TB, secure encrypted backup and high performance computing leveraging more than 4500 CPU cores and 360 NVidia GPUs. Support for providing user-controlled encryption for data-at-rest is far advanced. More than 2477 software libraries are available in Leonhard Med for various applications, such as bioinformatics, statistics, simulations, visualization, data science and machine learning tools. Additionally, solutions and support are provided for secure software installation by users, for containerized workflows, data management software tools, databases or for web-based applications in Leonhard Med. Leonhard Med can be used as a secure HPC cluster (Lustre as fast parallel file system and LSF as batch queue system) for large scale data processing and analysis and as a secure interactive workstation (both via command line and remote desktop) for processing and analysis of smaller scale data and/or requiring graphical user interface applications (e.g. RStudio, MATLAB, Juypter Notebooks, openBIS, custom web-based applications etc.).

The current Leonhard Med standard services are described in a service-level Service Level Agreement and comprise the following items: secure tenant creation, storage, backup and compute nodes allocation, user accounts setup, default enabled software library collection and technical expert support. In addition to the standard services, Leonhard Med custom services and expert support are provisioned for secure data transfer (e.g., BioMedIT process and tools, SFTP, VPN over IPsec, LiquidFiles, SETT), container support for computational workflows or applications (e.g., Docker, Podman, Singularity), databases (e.g. Postgres, MongoDB), data management tools (e.g. openBIS, RDF engine), development and deployment of custom web-based applications and data privacy and IT security awareness training (with SPHN/BioMedIT).

Leonhard Med Customer and User Base

Leonhard Med customer and user base steadily increased and cover research groups at Swiss academic institutions (i.e., ETHZ, UZH), multi-center research projects (SPHN, PHRT, the Tumor Profiler project, upcoming LOOP translational data driven precision medicine projects), clinical research facilities (NEXUS at ETHZ) and upcoming digital clinical trials (Digital Trial Intervention Platform, dTIP at ETHZ). While currently focused on research with confidential biomedical data, Leonhard Med supports any research project handling sensitive person data (i.e., confidential research data) and thus subject to Swiss data protection legal framework (FADP, HRA and HRO). SIS currently provides Leonhard Med services and/or is project partner for a total of 10 SPHN/PHRT, infrastructure development and platform projects (PRECISE, PSSS, PWS, SVIP-O, PHRT-MS, PHRT-MMA, SHFN, IMAGINE, SOCIBP, SwissPKcdw), 1 ETH-Industry consortium (Tumor Profiler), 1 SDSC funded project (MIDATA), 1 BRCCH (CGLBRCCH), 12 ETH research groups and clinical research facility (Prof. Gunnar Rätsch, Prof. Karsten Borgwardt, Prof. Joachim Buhmann, Prof. Julia Vogt etc. and NEXUS) and 1 UZH research group (Prof. Michael Krauthammer).

SIS BioMedIT Node Zurich Activities

SIS was and continues to be an active partner in collaborative activities with SPHN, PHRT, SDSC and BioMedIT, such as: BioMedIT Interoperability Working Group, BioMedIT IT Security Working Group, Task Force Data Management, SPHN Information Security Policy, SPHN Data Transfer and Use Agreement, SPHN/BioMedIT Data Privacy and IT Security Training course, SPHN Data Life Cycle Management Task Force. In particular, as part of the BioMedIT Interoperability Working Group (BIWG) SIS is actively contributing to: (i) development of the SETT tool, (ii) design, development and implementation of the BioMedIT secure data transfer service (i.e., security risk-based assessment, landing zone and operational process definition, implementation and operation, data provider and project data manager technical onboarding, service oversight and user support), (iii) BioMedIT portal frontend development and (iv) containerized computational workflows testing (Docker, Podman, Singularity). Furthermore, in the framework of BioMedIT, SIS designed and is developing the Filesystem Encryption Tool (FET), which is a command-line tool for user-controlled encryption of data on top of non-encrypted file systems for data at rest.

SIS proactively engaged in outreach activities targeted to the biomedical research community: co-developer of the BioMedIT/SPHN Data Privacy and IT Security Training course, regular RDM workshops at ETH, member and mentor for the MERH CAS workshop and Fresh Ideas for cancer Care, co-author of the BioMedIT white paper and speaker at Swiss Research Data Day, HPC.ch, SIB days, Personalized Health Conference, One Health meets Sequencing, ELIXIR Bioinformatics Industry Forum, information events on secure IT infrastructure services for the SPHN/PHRT, and LOOP programs.

Outlook

In alignment with the BioMedIT goals, SIS will focus on the consolidation of the current infrastructure and services and on the continual service improvement of Leonhard Med: enhance the user experience and expand the portfolio of customizable services, cover the full data life cycle (including secure long term data repositories), extend the domains to new use cases (e.g. digital clinical trials, data driven translational research).

Annex 3: Report of the Core-IT node

Overview

At the start of the BioMedIT project in 2017, the BioMedIT node was under the leadership of the Vital-IT group who had already experience in providing sensitive data infrastructures for the university hospitals in Lausanne (CHUV) and Geneve (HUG). That was a head-start for the project in terms of security and infrastructure knowledge which later resulted in a common security architecture that was done in collaboration with sciCORE and SIS as well as an external security company.

In Q4/2018 the node leadership was given to SIB's CTO and the newly created Core-IT group which runs IT infrastructure and services at SIB. A new BioMedIT infrastructure (compute cluster, firewall, encrypted storage system) was acquired and put in production in 2019. Since this time, the infrastructure has been operational and accessible to SPHN, BioMedIT and other projects. This was particularly true due to a very good collaboration with the BioMedIT network in (regular node head meetings were introduced in 2020) and the technical specifications of the BioMedIT node architecture which resulted in transfers of real data between hospitals and BioMedIT nodes. There was a particularly strong collaboration with sciCORE on the selection and operation of OpenStack as the virtualization system used for secure project spaces. Both, sciCORE and Core-IT, worked with the same consulting company StackHPC to operate and maintain the infrastructure. Since SIS/LeonhardMed are also moving to OpenStack instead of OpenNebula, an even stronger link will be made in the next phase of BioMedIT.

Technical and security aspects

Core-IT put a particular focus on advanced security for data stored in the BioMedIT node. Therefore, a modern storage system was chosen that allows encryption on the fly and at rest, i.e., sensitive personal data is always encrypted when accessible on the BioMedIT node Romandie, also referred to as SENSA (https://sensa.sib.swiss).

This strong security focus of the node was also expressed in several other aspects:

- Node personnel lead the creation of the first SPHN Information Security Policy that has been in place since 2018. That effort shaped the BioMedIT node architecture and operations adopting the principle of "security by design".
- A next step was to get a formal security review of BioMedIT's software and hardware architecture. This was done with a leading Swiss IT security company that approved the design. Furthermore, the groundwork for specific security assessments of the BioMedIT nodes was done, and the actual assessments are planned for 2021.
- In 2019 a "Data Privacy and IT security" training was put in place in collaboration with all three BioMedIT node, representatives from Swiss university hospitals, IT security and legal experts. The training was initially done as a classroom training, including scientific users and system administrators. The training is also available as an online training, and therefore ensured a continuous training during travel-restricted times.
- The node provides a key server to store PGP keys that are used for all data transfers in the BioMedIT network, from data providers to the respective project spaces at the nodes.

Other contributions of the node:

- Node personnel coordinated the activity to have a PAM model for SWITCHaai and edu-ID based access to SSH and SFTP servers. This was done in close collaboration with sciCORE (where the development was done), SIS as well as ELIXIR.
- Node personnel contributed to the various activities in the BioMedIT interoperability group, in particular, by contributing to the development of the secure data transfer tool (sett).
- In collaboration with the Clinical Bioinformatics group of SIB and SIB's legal department, a web application was developed to act as a GDPR compliant "record of processing". This application is in production at the node and was also proposed to other BioMedIT nodes.

Scientific projects

Given the early experience of the node with sensitive data infrastructures maintained for CHUV and HUG, the node acts as a Processor (according to GDPR) on a variety of SPHN-funded projects such as SOIN, SPO, SVIP and soon SACR. Additionally, several Swiss and EC-funded projects are supported, including SPSP, the SwissBioBanking Platform, IMMUCAN, etc.

The demands for new projects, additional storage space as well as CPU and GPU power are growing steadily.

Outlook for the next phase of BioMedIT 2021-2024

In 2020, the University of Lausanne (UNIL) has expressed interest in the BioMedIT node and its operation since they already started to provide a similar IT infrastructure for UNIL and CHUV. In 2021, negations will continue between SIB and UNIL to operate the node jointly and to provide a higher resource capacity. In fact, extending capacity for more scientific projects is a major goal for the 2nd phase of BioMedIT. That is particularly needed since there is a continuously growing demand for secure infrastructures for scientific projects.