SPHN Data Management Plan (DMP) Guidelines

Based on the SNSF requirements, with specific recommendations for SPHN/BioMedIT projects

Table of contents

Version 1, December 2021

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

PHI | Personalized Health Informatics Group

www.sphn.ch | dcc@sib.swiss

1 Aim

These guidelines are intended for all researchers who work with sensitive health-related data and are required to submit a data management plan (DMP), as required by most funding agencies. The document offers general recommendations for the use of **sensitive biomedical data** to help researchers to plan the entire life cycle of their data, with **particular focus on SPHN-funded projects** using health-related data for research purposes and data **processing on the BioMedIT infrastructure**. The document highlights the SPHN and BioMedIT standards and services, which support secure and FAIR data management. The guideline was developed by the Personalized Health Informatics group of the SIB Swiss Institute of Bioinformatics following the guidelines of the SNSF template "Data Management Plan – content of the mySNF form" and closely aligning with the DLCM template.

2 Document structure

Following the original SNSF guidelines, the core of the SPHN Data Management Plan Guidelines is divided in the following four chapters:

1. Data collection and documentation,
2. Ethics, legal and security issues,
3. Data storage and preservation, and
4. Data sharing and reuse.

Content directly extracted from the original SNSF guidelines is presented in *italic* at the beginning of each sub-chapter. In addition to the SNSF general guidelines, each chapter contains three sections, namely: (i) SPHN Recommendations, (ii) SPHN Specifics [displayed in blue text boxes], and (iii) SPHN Examples, addressing the planning of the data life cycle with SPHN/BioMedIT projects in mind.

The section SPHN Recommendations expands on the points provided by SNSF and serves as a guide for the researcher. Here, researchers can find helpful information required to address the specific topics, with data-driven research (relying on the further use of health-related data) as central focus.

The section SPHN Specifics, on the other hand, details SPHN/BioMedIT related frameworks, procedures and infrastructures that an SPHN-funded project might need for their DMP. Finally, the SPHN Example section provides short paragraphs illustrating real-world use cases within the SPHN landscape. These texts can serve as inspiration for new projects in the DMP preparation process.

3 Guidelines (SNSF questions)

1. Data collection and documentation

1.1 What data will you collect, observe, generate or reuse?

- *What type, format and volume of data will you collect, observe, generate or reuse?*
- *Which existing data (yours or third-party) will you reuse?*

Briefly describe the data you will collect, observe or generate. Also mention any existing data that will be (re)used. The descriptions should include the type, format and content of each dataset. Furthermore, provide an estimation of the volume of the generated data sets.

SPHN Recommendations:

For this section, please cover the following points for each dataset in your project:

- **Cohort:** Describe the characteristic of the group(s) of patients to be included in the project (e.g., number of patients, inclusion and exclusion criteria as age or diagnosis).
- **Data type:** Describe the type(s) of data you plan to collect, observe, generate or reuse (e.g., laboratory data, diagnoses, whole-genome sequencing (WGS), intensive care unit (ICU) monitoring data, outcome data, patient-reported outcome measures (PROMs)) as well as the purpose of the dataset(s) within the context of the project.[1]
- **Data origin and availability:** Describe the origin(s) of the data. This may be routinely collected clinical data, data collected specifically for this study or reuse of (an) existing dataset(s) (e.g., from a cohort study, a registry or a repository). Provide a list outlining the data availability per data source e.g., hospital. For reused data, please add a reference to the original source.
- **Data format:** Description of the data format(s) used for raw, curated and processed data. To maximize the reuse potential of the data, open standard formats are preferred.
- **Data volume:** Estimate the total volume of your data in gigabyte (raw, curated and processed data).

[1] An overview of different data streams and data types can be found in the annex of this document.

SPHN Specifics

SPHN defined a list of recommended formats based on the knowledge acquired by the SPHN Driver projects in different fields of clinical research:

Data type	Recommended format
Clinical routine and research data	RDF (.ttl) - structured clinical data, Text (.txt) – clinical reports.
Genomic and transcriptomic data	FASTA (.fasta) – Raw nucletide sequences, FASTQ (.fastaq) – Nucleotide sequences with quality scores, SAM/BAM (.sam /.bam) – Representation of aligned sequences, CRAM (.cram) – Alternative to SAM/BAM to store aligned sequences providing significantly better compression, VCF (.vcf) – Storage of gene sequence variations.
Multimedia data	DICOM (.dicom) – Medical imaging.

SPHN Examples:

Example 1: “The project will include data of 100 patients from the University Children’s Hospital Basel (UKBB) and 100 patients from the University Children’s Hospital Zurich (Kispi), with the following inclusion criteria: female, 2 to 20 years old with leukemia. The project will use the following data types:

1. **Clinical routine data:** These datasets will include demographics, diagnosis, and laboratory values. For this project, the data will be collected in clinical routine and extracted from the data management infrastructures of UKBB and Kispi (more details can be found in the table below). The data will be delivered in the Resource Description Framework (RDF) turtle (ttl.) format. We anticipate that the clinical routine data collected will amount to approximately 100 MB.

Data type	**Concepts**	**Comment**	**Data source**	**Availability**	
				USB	**USZ**
Demographics	Birth date		CDW	Available	Available
	Gender		CDW	Available	Available
Diagnosis	FOPH Diagnosis		CDW	Available	Available
	Diagnosis	Extracted by NLP	Discharge letter	50 de-identified reports available	Reports needed to be de-identified before sharing
Laboratory results	Tumor marker	Additional information on machine and kit needed	Lab information system	Lab results available, additional information needs to be extracted from the source system	Available
	….				

2. **Whole-genome sequences (WGS)**: Clinical-grade WGS will be generated in the Genome Center (Health2030 Genome Center) from biobank samples obtained from the Swiss Pediatric Hematology and Oncology Biobank Network of the University Children’s Hospital Zurich. WGS raw sequences will be stored in FASTQ format and alignments as Binary Alignment Map (BAM) files. The genomic data generated is expected to total 4 - 5 GB.”

1.2 How will the data be collected, observed or generated?

- *What standards, methodologies or quality assurance processes will you use?*
- *How will you organize your files and handle versioning?*

Explain how the data will be collected, observed or generated. Describe how you plan to control and document the consistency and quality of the collected data: calibration processes, repeated measurements, data recording standards, usage of controlled vocabularies, data entry validation, data peer review, etc. Discuss how the data management will be handled during the project, mentioning for example naming conventions, version control and folder structures.

SPHN Recommendations:

For this section, please cover the following points for each dataset in your project:

- **Data generation/collection:** Describe how the data will be generated/collected. Depending on the data types used, this information shall be requested from the data providing institutions.
- **Controlled vocabularies:** Explain which controlled vocabularies you use within your project, and where you use them (cf. FAIR criteria I2: "Data use vocabularies that follow FAIR principles") to semantically define the meaning both for humans and machines. If your chosen controlled vocabulary is not FAIR, please describe how you plan to make it (more) FAIR (for guidelines for making vocabulary FAIR, see[2]). Please consider controlled vocabulary for concepts (meaning binding), value sets (value set binding), and the data point.
- **Methodology used to represent your data:** Describe the language or data model chosen for knowledge representation (cf. FAIR criteria I1: "Data use a formal, accessible, shared, and broadly applicable language for knowledge representation").
- **Data linkage:** Describe how your data is linked or linkable to other data (cf. FAIR criteria I3: "(Meta)data include qualified references to other (meta)data").
- **Data quality and consistency:** Describe the processes in place to ensure data quality and consistency. This may include the calibration processes, repeated measurements, data recording standards, data validation, data peer review, etc.

[2] https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1009041

- **Data management:** Describe the data management of your project. This includes naming conventions, version control systems and the use of unique identifiers (cf. FAIR criteria F1: “(Meta)data are assigned a globally unique and persistent identifier”).

Ethics and legal (consent, legal contracts, etc.) issues are addressed in chapter 2.

SPHN Specifics:

SPHN developed a data-driven semantic framework composed by a strong semantic layer and a descriptive language solution which is both flexible and easy to expand. This solution allows SPHN projects to directly address important FAIR criteria such as (F1, A1, I1, I2, I3), thus achieving a high degree of interoperability across projects, systems, countries and over time.

The interoperability framework relies on the use of controlled vocabulary (e.g., SNOMED CT and LOINC) and additional data standards (see examples in the table below) to represent the data (FAIR criteria: I2). The datasets are described and structured using a graph-based data representation in RDF (WC3 standard, FAIR criteria: I1) supported by a quality control framework.

Clinical data type	Standard
Diagnosis	ICD-10-GM
General clinical terms	SNOMED CT
Medication	ATC
Laboratory test	LOINC
Procedures	CHOP
Multimedia	DICOM
Units	UCUM
Oncology diagnosis	ICD-O3

In RDF, each data point is assigned to an Internationalized Resource Identifier (IRI), which allows their unique identification (FAIR criteria: F1). All SPHN concepts include a definition of the concepts and its properties, value set and recommended standards including a naming and versioning convention. All SPHN concepts have human and machine resolvable IRI on the web under https://www.biomedit.ch/rdf/sphn-ontology/sphn. The use of IRIs also facilitates the creation of cross-references to other (meta)data (FAIR criteria: I3). The representation of the data in RDF ensures compatibility with other W3C standards. These include standards such as SPARQL, a communication protocol for querying the data (FAIR criteria: A1) and SHACL, a specification for the validation of graph-based data. An SPHN SHACL set was developed to provide a validation mechanism for the data to improve data quality and consistency. Versioning of the ontology is handled by a version IRI provided by Data coordination center (DCC) for the SPHN RDF schema and provided by the project for the project ontology.

Find out more about the SPHN Interoperability Framework on: https://sphn.ch/network/data-coordination-center/the-sphn-semantic-interoperability-framework/

SPHN Examples:

Example 2: “All clinical routine data is collected in clinical routine and stored in the clinical data warehouses of USB and USZ. Additional clinical data within this project will be collected according to the study protocol [Insert description or reference to the study protocol]. The data will be represented according to the SPHN semantic framework recommendations, following the SPHN RDF schema. During the project, we will extend the current RDF schema to cover additional oncology data (see new concepts below in the table). For the oncology related concepts, we will use “NCI Dictionary of Cancer Terms” or/and SNOMED CT as controlled vocabulary to describe our concepts. We will use the following standards to represent our data: cancer diagnosis with ICD-O3, gene names with HUGO, and genomic variants with HGVS. To validate the data produced, we will use the SPHN quality framework and extend it with additional validation rules and constraints for our new oncology concepts and (e.g., reference ranges for laboratory values). The WGS data will be generated by DNA Sequencing Platform of the Health 2030 Genome Center using standardized sequencing protocols. The whole genome sequencing service of the Genome center is ISO 15189:2013 accredited by the Swiss Accreditation Society (accreditation number STS 0714).”

Example of a new concept:

	Name	**Description**	**Type**	**Standard**	**Meaning binding international standard**
Concept	Gene				SNOMED CT 82256003 | Human gene (substance) |
Composed of (property)	id	unique gene id according to HGNC nomenclature	Code	HGNC	
Composed of (property)	alternative symbol	alternative symbol(s) for the gene (e.g., Uniprot).	string		

1.3 What documentation and metadata will you provide with the data?

- *What information is required for users (computer or human) to read and interpret the data in the future?*
- *How will you generate this documentation?*
- *What community standards (if any) will be used to annotate the (meta)data?*

Describe all types of documentation (README files, metadata, etc.) you will provide to help secondary users to understand and reuse your data. Metadata should at least include basic details allowing other users (computer or human) to find the data. This includes at least a name and a persistent identifier for each file, the name of the person who collected or contributed to the data, the date of collection and the conditions to access the data. Furthermore, the documentation may include details on the methodology used, information about the performed processing and analytical steps, variable definitions, references to vocabularies used, as well as units of measurement. Wherever possible, the documentation should follow existing community standards and guidelines. Explain how you will prepare and share this information.

SPHN Recommendations:

For this section, please cover the following points for each dataset in your project:

- **Descriptive metadata:** Describe metadata that depict contextual information of your dataset (e.g., which data is included in the dataset, what is the meaning of a variable) as well as data elements itself (e.g., with which method the measurement was performed, at which temperature the sample was stored). Describe how this metadata information is made machine readable and how you will preserve the definitions of their meaning in the future.
- **Administrative metadata**: Describe metadata that provide technical information about the data resource and its accessibility (e.g., when the resource was created, who the author is, governance processes).
- **(Meta)data standards:** Describe which community standards (if available) have been used to annotate the (meta)data. If possible, provide metadata using controlled vocabulary.
- **Additional documentation:** Provide all necessary information about the methods used to generate the data (e.g., SOPs, description of the software used including its version, and important parameters). As a rule, favor the use of existing community standards and guidelines.

These points answer the following FAIR criteria:

- F1. (Meta)data are assigned a globally unique and persistent identifier,
- F2. Data are described with rich metadata,
- F3. Metadata clearly and explicitly include the identifier of the data they describe, and
- R1. (Meta)data are richly described with a plurality of accurate and relevant attributes.

SPHN Specifics

In SPHN, the project specific RDF schema includes several descriptive and administrative metadata. In the RDF header, administrative metadata is provided with the SPHN ontology version used, the extraction date of the data and the identifier of the data provider (Figure 1).

```
resource:CHE_108_907_884-DataRelease_1620055600 a sphn:DataRelease ;
    dct:conformsTo <https://biomedit.ch/rdf/sphn-ontology/psss/2021/3> ;
    sphn:hasExtractionDate  "2021-05-02"^^xsd:date ;
    sphn:hasDataProviderInstitute resource:CHE_108_907_884-DataProviderInstitute .
```

Figure 1: Extract of a project's RDF Schema depicting various administrative metadata.

The project ontology includes the descriptive metadata for all data elements, including which data elements to include, their definition, standards and/or value set to be used. Additional properties of a concept provide additional (meta)data of a data element. In the example below, the property "method" within to the concept "Heart Rate" is associated to a second concept, "Method description", that provides more information on the exact method used to measure the heart rate. In a similar way, the property "physiological state" is used to describe if the patient was, for example, standing or sitting when the measurement was taken (Figure 2).

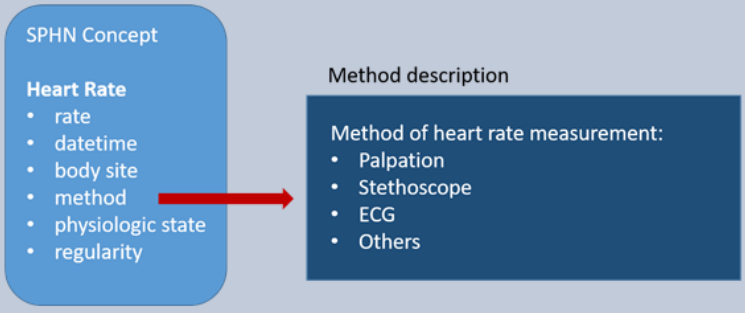

Figure 2: Example of an SPHN Concept "Heart Rate", linked to a specific instance of the property "method".

SPHN Example:

Example 3: “For our project we will provide the administrative metadata following the SPHN recommendation. We will expand the administrative metadata with a new class “DataUse”, which will include the data use restrictions according to Data Use Ontology (DUO) from GA4GH. All administrative metadata will be included as classes in the RDF turtle (.ttl) files. For the descriptive metadata, we will expand the properties of the corresponding classes in the SPHN RDF schema. Additionally, detailed information about the standardized sequencing protocols used to generate the WGS data and other experimental procedures will be provided as text files.”

2. Ethics, legal, and security issues

2.1 How will ethical issues be addressed and handled?

- *What is the relevant protection standard for your data? Are you bound by a confidentiality agreement?*
- *Do you have the necessary permission to obtain, process, preserve and share the data? Have the people whose data you are using been informed or did they give their consent?*
- *What method will you use to ensure the protection of personal or other sensitive data?*

Ethical issues in research projects demand for an adaptation of research data management practices, e.g. how data is stored, who can access/reuse the data and how long the data is stored. Methods to manage ethical concerns may include: anonymization of data; gain approval by ethics committees; formal consent agreements. You should outline that all ethical issues in your project have been identified, including the corresponding measures in data management.

SPHN Recommendations:

Within this section, consider all possible ethical and legal issues you might face in the context of your project and how you plan to address these issues. The following points can help to frame the ethical and legal implications of your project:

a. **Patients' consent:** Describe the type of consent you will use for reusing and/or collecting the data (general consent or informed consent) or if you are planning to request ethics approval for using data without consent (according to Art. 34 of the Human Research Act[3]).
b. **De-identification of the data**: Describe if you will use anonymized, pseudonymized (coded) or identifying personal data for your project. Describe who is responsible for the de-identification and, in case of pseudonymization, who is the key-keeper for the re-identification. If internal guidelines/specifications on how data need to be de-identified exist, provide a reference or link to the corresponding document. If there are no internal specifications available, list at least the rules which apply for de-identifying direct identifiers (e.g., date of birth is replaced by the corresponding year of birth).
c. **Ethics approval:** Describe, if you have submitted a request for approval to the ethics committee or plan to do so, or attach a statement of the ethics committee. Otherwise, describe why ethics approval is not foreseen.

[3] https://www.fedlex.admin.ch/eli/cc/2013/617/en

Please note that legal agreements (Collaboration Agreement, Data Transfer and Use Agreements and Data Processing Agreements), which provide important regulations to protect personal and sensitive information, are addressed in Section 2.3.

SPHN Specifics

All SPHN projects need to follow the SPHN Ethical Framework. The Framework provides ethical guidance to the partners of the Network regarding the collection, storage, analysis and sharing of personal data for research purposes. Health-related personal data are often derived from human biological material. Both data and human biological material are addressed in the second version of the Framework, which is endorsed by the Swiss Biobanking Platform (SBP) and the ETH Domain Strategic Focus Area on Personalized Health and Related Technologies (PHRT).

De-identification of project data in SPHN

Data providing institutions (e.g., hospitals) have to de-identify personal health-related data before sharing data with research projects. Data might be pseudonymized or anonymized depending on the project specifications. Data providers need to apply consolidated de-identification rules, mitigating the risk of re-identification without losing data value. SPHN aims to provide harmonized recommendations on how to de-identify health-related data in compliance with Swiss law requirements. The de-identification process follows a risk-based approach at which a use case evaluation and risk assessment is conducted based on the selected de-identification rules and project controls.

SPHN Examples:

Example 4: “The project will follow the SPHN Ethical Framework. This project will use routinely collected data from patients who signed the general consent at the University Hospital Basel (USB) and the University Hospital Lausanne (CHUV). Project data will be pseudonymized following consolidated de-identification rules and institutional processes, which are described in detail in [Insert appendix]. As example, direct identifiers such as patient names will be replaced by pseudo names. The re-identification key will be kept at a safe place at the respective hospital. The project will be submitted to the ethics committee (Ethikkommission Nordwest- und Zentralschweiz, EKNZ) to obtain the necessary ethical authorizations for the further use of the consented pseudonymized data relevant to this project.”

Example 5: “The project will follow the SPHN Ethical Framework. This project will use anonymized CT images of lung cancer from patients who signed the general consent at the University Hospital Basel (USB) and the University Hospital Zurich (USZ). The anonymization follows the internal guidelines of USB and USZ, which are described in detail [Insert appendix]. All direct identifiers are deleted/suppressed. The project will submit an inquire to the ethics committee (Ethikkommission Nordwest- und Zentralschweiz, EKNZ) to confirm that this project does not fall under the Human Research Act (HRA).”

2.2 How will data access and security be managed?

- *What are the main concerns regarding data security, what are the levels of risk and what measures are in place to handle security risks?*
- *How will you regulate data access rights/permissions to ensure the security of the data?*
- *How will personal or other sensitive data be handled to ensure safe data storage and -transfer?*

If you work with personal or other sensitive data, you should outline the security measures in order to protect the data. Please list formal standards which will be adopted in your study. An example is ISO 27001-Information security management. Furthermore, describe the main processes or facilities for storage and processing of personal or other sensitive data.

SPHN/BioMedIT Recommendations:

Within this section, outline all security measures implemented to protect your data. These might include:

- **Security standards**: Indicate any formal security standards (e.g., ISO27001, NIST) and policies (e.g., SPHN/BioMedIT Information Security Policy) you comply with, including a brief description. Such standards shall encompass prevention of unauthorized access, logging of user access, organizational measures to protect data or reporting in case of any suspected data security breach.
- **Data encryption:** Indicate which type of encryption is used to protect your data, and where (e.g., data transfer, data storage). Note that de-identification procedures are discussed in Section 2.1.
- **Data transfer:** Describe how data transfer will be handled to ensure the safety of the data (e.g., method of encryption, process of authorization to send data, logging of data transfers). If a Data Transfer and Use Agreement (DTUA) is set up for the project, make sure that the information is complementary and in accordance with the DTUA.
- **Data access:** Describe how data access is handled. This should include authentication and authorization of the users, and details about access rights/permissions management, including a traceable logging of data access.

Please note that detailed questions regarding data storage and backup are addressed in Section 3.1

SPHN Specifics

The BioMedIT Network provides the access to a service infrastructure that allows the secure transfer, storage, management, and processing of sensitive research data to all researchers in Switzerland in accordance with the Swiss Federal Act on Data Protection. The network operates under a common SPHN/BioMedIT Information Security Policy using state-of-the-art security techniques, and includes:

BioMedIT secure project environment: Data security in the BioMedIT nodes is principally based on allocation of project-specific IT resources within an access-controlled, private, virtual environment offering network isolation, data isolation and computational resources isolation (private tenant). Shared tenants are only permitted in those cases where there is a specific authorization. A private tenant ensures that data stored in one project space cannot be shared – intentionally or by accident – with another project. Access to the internet from the BioMedIT node is strictly controlled, limited to trusted and explicitly whitelisted web resources. Users can then connect to project spaces for which they are specifically authorized via a virtual desktop with a graphical user interface or a virtual terminal session. Contractual and technical measures prevent that data is shared and/or combined without the appropriate authorization. Transfer, access and processing operations are logged. In addition, physical access to the BioMedIT node server rooms is tightly controlled. The BioMedIT network encompasses three BioMedIT nodes (Leonhard Med at SIS/ETHZ, sciCOREmed at UNIBAS, and SENSA at UNIL/SIB.

Data access rights and permission: BioMedIT users are trained in “Data privacy and IT security”. The network offers the “Data Privacy and IT Security Training” as on-line training or class-room course hosted in different cities in Switzerland. It is the responsibility of the Project Leader to grant access to the project space. An authorized user can then access the project space via the BioMedIT Portal using a SWITCH edu-ID account with two-factor authentication. Additionally, the BioMedIT network can only be accessed from within trusted IT environments (e.g. from within a Swiss university or university hospital network or via VPN). All access to the system is traced. Logs recording user activities, exceptions, faults and information security events are kept and regularly reviewed.

Encrypted data transfer: After a registration process in the BioMedIT portal, data transfer requests are authorized by both the sender and the BioMedIT nodes involved. After encryption and signing by the sender, the data can be sent via a secure channel using Secure File Transfer Protocol (SFTP) or Liquid Files to a data provider-specific landing zone on BioMedIT, specifying the approved data transfer ID. To send or receive data using this method, your PGP key must be authorized by the SPHN DCC. To streamline this process, BioMedIT provides the “secure encryption and transfer tool” (sett). All data transfers are logged centrally. For more information, please consult the sett info sheet.

SPHN/BioMedIT Examples:

Example 6: “The data will be processed and stored on a secure project space in the BioMedIT node sciCORE at the University of Basel following the SPHN/BioMedIT Information Security Policy. Data security in the BioMedIT nodes is principally based on allocation of project-specific IT resources within an access-controlled, private, and virtual environment offering network isolation, data isolation, and computational resources isolation (private tenant). Access to the Internet from the BioMedIT node is strictly controlled, limited to trusted and explicitly whitelisted web resources. Users can then connect to project spaces for which they are specifically authorized via a virtual desktop with a graphical interface or a virtual terminal session. Legal agreements set up for the project provide contractual and technical measures preventing that data is shared and/or combined without the appropriate authorization. Transfer, access and processing operations are logged. In addition, physical access to the BioMedIT node’s server room is tightly controlled.

Following the BioMedIT guidelines, project partners will be authorized by the Project Leader to get access to the data via a remote desktop environment. All data will be securely transferred using end-to-end encryption based on public-key cryptography via the sett tool provided by BioMedIT. Data from USB will be sent directly to sciCORE, while data from USZ will be sent to the BioMedIT nodes SIS and transferred through the BioMedIT network to an isolated project space on sciCORE. Only the authorized data manager of the project has the key to decrypt the data within the isolated project space.”

2.3 How will you handle copyright and Intellectual Property Rights issues?

- *Who will be the owner of the data?*
- *Which licenses will be applied to the data?*
- *What restrictions will apply to the reuse of third-party data?*

Outline the owners of the copyright and Intellectual Property Right (IPR) of all data that will be collected and generated, including the license(s). For consortia, an IPR ownership agreement might be necessary. You should comply with relevant funder, institutional, departmental or group policies on copyright or IPR. Furthermore, clarify what permissions are required should third-party data be reused.

SPHN Recommendations:

This section will encompass all legal matters of project such as rights, obligations and responsibilities of the parties, data ownership, licensing and reuse conditions.

- **Legal agreements and governance:** Briefly describe the legal agreements that regulate the usage and sharing of project-related data between the project partners. If you have to follow internal governance processes to get project relevant data released, mention the respective application for approval. The documents should encompass:
    a. **Data ownership:** Data ownership goes along with the responsibility of the institution/hospital that shared its patients' data. In most cases of (re)using health-related data patient's consent is required. Describe how your project ensures patients' data privacy and compliance with the respective legal regulations.
    b. **Restrictions for reuse of data:** Describe how the reuse of data is regulated respectively restricted by the data providing institution. Verify if there is a restriction for reusing data by third parties.
    c. **Licensing and Intellectual Property (IP) rights:** Indicate how background and foreground IP rights on the results are regulated. "Results" are defined here as any results generated by a party from its participation in the project – such as invention, data, software, algorithms, knowledge, know-how or information, as well as any rights attached to it, including intellectual property rights. Describe how you handle licensing within the project consortia and for the reuse by external parties.
    d. **Publications and authorship:** Describe which guideline and recommendation you will follow for authorships for publication e.g., those of the Swiss Academies of Arts and Sciences. If you do not follow any existing guideline, please describe your guidelines regarding authorship and publication.

SPHN Specifics

SPHN provides different templates to define the conditions for using and transferring data:

- Consortium Agreement (CA),
- Data Transfer and Use Agreement (DTUA), and
- Data Transfer and Processing Agreement (DTPA), when using the BioMedIT infrastructure.

Besides the defined rights, obligations and responsibilities of all parties involved in the multi-center research project, the contractual architecture defined in the legal agreement settles additional important issues that need to be legally addressed, including: permitted use and ownership of data, publications, intellectual property and liability. More information is available here.

Data ownership and data sharing: Typically, SPHN project partners are joint controller for project data defining together the means of processing. However, ownership of original data and its responsibility remains at each institution/hospital. Project data must be de-identified (pseudonymized or anonymized) prior to sharing and transferring to the respective BioMedIT node. Moreover, an approval for sharing from each institution/hospital and ethics committee might be needed. Reuse of original data typically underlies a new data request at the original institution/hospital. You may use the SPHN legal templates and adapt them to your needs, if desired.

Foreground Intellectual Property (IP): The foreground IP means all intellectual property rights made in the performance of work under an agreement. You may choose some of the following options and specify in your DTUA that:

a. The IP is owned and vest solely by the party generating it (Sole Foreground IP)
b. The IP is jointly owned by the project partners. (Joint Foreground IP)

The respective formulations can be found in the legal agreement templates for the Consortium Agreement or Data Transfer and Use Agreement https://sphn.ch/services/dtua/.

Handling of licensing of results: You may choose some of the following options and specify in your DTUA:

a. Each party generating foreground IP by using data, human samples, confidential information or background IP of another party hereby grant to that party a royalty-free, worldwide, non-transferrable, non-exclusive, irrevocable license, with the right to grant sublicenses, to access and use that Foreground IP for purposes of internal scientific research.
b. Special licenses: Upon agreement between the parties and must be specified in an agreement set up between the parties (e.g., CA/DTUA).

The respective formulations can be found in the legal agreement templates for the Consortium Agreement or Data Transfer and Use Agreement https://sphn.ch/services/dtua/.

SPHN Examples:

Example 7: “A Consortium Agreement with an included Data Transfer and Use Agreement and a Data Transfer and Processing Agreement will be set up among the partners following the templates of SPHN (link). The consortium consists of HUG, USZ and ETH Zurich, while HUG and USZ provide the data. All parties act as data recipient and determine together the means of processing, acting as joint controllers. Data will be transferred via the BioMedIT nodes SENSA (University of Lausanne/SIB) and SIS (ETH Zurich) (data processors), to the respective secure project space at SIS.

With respect to the exploitation of Foreground IP generated by the recipient (ETH) using the joined dataset, the following revenue-share mechanism shall apply: One half of net revenues is received by the recipient (ETH) and the other half of net revenues is received by the other parties, who shall, by separate mutual agreement, agree on the individual distribution among each other. For the purposes of this agreement, 'net revenues' means income generated by either party in exchange for the licensing or sale of intellectual property rights to third parties, less demonstrable effective patenting costs. The BioMedIT nodes act as data processors and will not have any IP rights on the results.

With respect to publication, we will follow the guidelines of the Swiss Academies of Arts and Sciences.”

3. Data storage and preservation

3.1 How will your data be stored and backed-up during the research?

- *What is your storage capacity and where will the data be stored?*
- *What are the back-up procedures?*

Please mention what the needs are in terms of data storage and where the data will be stored. Please consider that data storage on laptops or hard drives, for example, is risky. Storage through IT teams is safer. If external services are asked for, it is important that this does not conflict with the policy of each entity involved in the project, especially concerning the issue of sensitive data. Please specify your back-up procedure (frequency of updates, responsibilities, automatic/manual process, security measures, etc.).

SPHN Recommendations:

Within this section, summarize all information related to the storage and back-up procedures in place for the duration of the project. It is important to carefully assess the level of security required by your data in order to select the appropriate data storage and back-up procedures.

- **Data storage:** Describe your needs in terms of storage and your plans in that regard. This should include a description of the system used, its capacity, the security standards, and their physical location.
- **Data back-up:** Describe the back-up system used, the frequency and the related procedures.

For the projects not using the BioMedIT Network, please consult your institutions to obtain the necessary details about data storage and back-up procedures.

SPHN/BioMedIT Specifics

SPHN projects running on BioMedIT will have their data stored and backed-up by their BioMedIT node. All BioMedIT nodes are compliant with the framework established in the SPHN/BioMedIT Information Security Policy and are responsible for its technical implementation, in particular for:

- **Data storage:** After transfer from the data providers, data is stored in the isolated project spaces. Encryption at rest is available on request. Storage capacity is allocated on a need basis and needs to be discussed with the BioMedIT node. The data will be stored on an isolated project space in the BioMedIT nodes until the end of the project. Storage capacity allocation needs to be discussed in advance with the BioMedIT node. The physical access to the BioMedIT node's server room is tightly controlled.
- **Data backup:** Backups are performed regularly through a fully automated process (the frequency is determined on an individual project basis). The nodes are considered responsible for the backup of the data. Backup procedures and controls protect the confidentiality of data on backup media such as employing encryption for media holding confidential data.

Important: Projects willing to be hosted on BioMedIT need to discuss the projects needs and requirements with the main hosting BioMedIT node. To initiate the procedure, please contact the Personalized Health Informatics here: dcc@sib.swiss.

SPHN/BioMedIT Examples:

Example 8: "The original data used in the context of this project are retrieved directly from the clinical data warehouse of the university hospitals storing data according to the local institutional policies. Once transferred via the BioMedIT Network, the data will be stored in the isolated project space on SIS, ETHZ, which is the main BioMedIT node in this project. Project data will be backed up regularly. It is possible to consult the storage capability, security policies and backup procedures of SIS here [Insert the appropriate link]."

3.2 What is your data preservation plan?

- *What procedures would be used to select data to be preserved?*
- *What file formats will be used for preservation?*

Please specify which data will be retained, shared and archived after the completion of the project and the corresponding data selection procedure (e.g. long-term value, potential value for reuse, obligations to destroy some data, etc.). Please outline a long-term preservation plan for the datasets beyond the lifetime of the project. In particular, comment on the choice of file formats and the use of community standards.

SPHN/BioMedIT Recommendations:

Within this section, describe which part of the data will be preserved for archiving or sharing purposes, beyond the duration of the project. Please note that data preservation also concerns data that is not intended for publication. The section shall describe:

- **Data selection procedure**: Describe the procedure which data will be retained, shared and archived after the completion of the project.
    - Potential for reuse (e.g., high quality (meta)data, high research value, or open license),
    - Ethical and legal considerations,
    - Limitations and conditions for reuse, and
    - Data preservation costs.
- **Long term preservation:** Please outline a long-term preservation plan for the datasets beyond the lifetime of the project. Comment about the choice of file format and eventual community standards.

In case part of data will not be preserved and/or made available for reuse, please justify your choice.

SPHN/BioMedIT Examples:

Example 9: "For this project, all processed data will be preserved for 10 years on secure storage servers of our institution, following the internal archiving process. Selected WGS will be deposited in the European Genome-phenome Archive (EGA). The sequences have been selected for their relevance in the field of [Insert field of research] and [Insert specific reuse potential associated to your data]. More details on the EGA repository are covered in Section 4.1."

4. Data sharing and Reuse

4.1 How and where will the data be shared?

- *On which repository do you plan to share your data?*
- *How will potential users find out about your data?*

Consider how and on which repository the data will be made available. The methods applied to data sharing will depend on several factors such as the type, size, complexity and sensitivity of data. Please also consider how the reuse of your data will be valued and acknowledged by other researchers.

SPHN/BioMedIT Recommendations:

Within this section, indicate how your data will be made available to other researchers. In order to select the most appropriate data repository, consider the following points:

- **Before releasing your dataset(s), verify that:**
  a. The data is well described: Ensure that the metadata accompanying your data is rich and exhaustive. Your data needs to be interpretable by humans and machines alike.
  b. Reuse conditions are clearly indicated: It is not uncommon to release a dataset upon certain specific reuse conditions. This might include publications rights, use only in a specific research field, use for not-for-profit research etc.
  c. The acknowledgement procedure is indicated: Define how the data controllers will be acknowledged by other researchers upon the reuse of the data. [This is usually done through an institutional data governance process, such as a Data Access Committee (DAC).]

- **Choice of a FAIR repository, please assess:**
  a. Repository maturity: Favor domain-specific, well-established or certified repositories.
  b. Repository FAIRness (see Section 4.3): Favor repositories that fulfill the FAIR criteria.
  c. Host of the repository (see Section 4.4): Favor repositories hosted by non-for-profit organizations.
  d. Security level of the repository: Choose a repository with the necessary security level for your data (e.g., encrypted data transfer, controlled access).

SPHN/BioMedIT Examples:

Example 10: “Datasets from this project will be submitted to the European Genome-phenome Archive (EGA). EGA is a well-established archive for sensitive data with controlled access. Storage within the EGA enables permanent archiving of sensitive data with no submission or storage fees. The data will be encrypted prior to submission and stored using the interoperable and open-source CRAM format recommended by the Global Alliance for Genomic Health (GA4GH). A Data Access Committee (DAC) will be established to regulate the access to the genetic data as required by EGA. Whole Genome Sequences (WGS) and Whole Exome Sequences (WES) will be submitted in bam format and will include the following metadata (EGA standard):

- Study: information about the sequencing study
- Samples: Information about the sequencing samples
- Experiments: Information about the sequencing methods, protocols and machines. Experiments generate the linkage between samples and study. This is only necessary for FASTQ and BAM/CRAM submissions.
- Runs: Samples, experiments and files are linked through runs - appropriate objects for FASTQ and BAM/CRAM submissions
- Analysis: References the analysis (BAM) files; associated with samples and study.
- DAC: contains information about the Data Access Committee (DAC)
- Policy: contains the Data Access Agreement (DAA); associated with DAC
- Dataset: contains the collection of runs/analysis data files to be subject to controlled access; associated with policy.

The EGA Accession ID of the datasets will be included in the publication.

Metadata about datasets will be publicly searchable on the central EGA web portal where details about the dataset accessibility are readily available.”

4.2 Are there any necessary limitations to protect sensitive data?

- *Under which conditions will the data be made available (timing of data release, reason for delay if applicable)?*

Data have to be shared as soon as possible, but at the latest at the time of publication of the respective scientific output. Restrictions may be only due to legal, ethical, copyright, confidentiality or other clauses. Consider whether a non-disclosure agreement would give sufficient protection for confidential data.

SPHN Recommendations:

Within this section, describe when and under which conditions the data will be made available. If limitations are foreseen, include a brief description of their nature (e.g., legal, ethical, copyright, confidentiality or other) and how this will impact data sharing and reuse.

Keep in mind that data should be shared at the time of publication at the latest. SPHN projects are encouraged to make all suitable data available for sharing and reuse within their ethico-legal boundaries (within the National Data Stream funding of SPHN, projects are obliged to properly plan for third-party use of the data).

Limitations applied to data sharing and reuse might include:

- **Legal and ethics restrictions**: If not already addressed in Sections 2.1 and 2.3, describe ethical and legal limitations associated to your data. This may include the provision of amendments to the legal agreements and document approved by the ethics committee.
- **Publication embargo:** An embargo may be requested to delay the distribution of a dataset until an indicated time. Embargoes are most commonly used when dealing with sensitive data, to ensure publication exclusiveness and upon previous agreements with funders/industry partners.
- **Revocation of consent**: The revocation of consent from the participants of the study might affect the availability of the data. Describe how a resulting reduction in available data will be handled.
- **Sensitive data anonymization:** Projects dealing with personal identifying information need to undergo all necessary de-identification procedures before diffusion (please check with your institutions for details).

SPHN Examples:

Example 11: “All sensitive data will undergo an internal data anonymization process following the guidelines of the University Hospital of Basel (USB) [Insert link here]. All data associated to a publication will be made available at the time of publication. Unpublished data will be made available [Indicate embargo duration] after publication. The reasons for the embargo are [Provide a case or justification to for the application of the embargo].”

4.3 Do all digital repositories I chose conform to the FAIR Data Principles?

SPHN projects should follow the latest SNSF directives regarding the choice of repository. The requirements state that all data generated during a research project will be archived into open repositories adhering to the FAIR principles. In that regard, the SNSF defined a set of minimum criteria that repositories have to fulfil to conform with the FAIR Data Principles:

- *Are datasets (or ideally single files in a dataset) given globally unique and persistent identifiers (e.g., DOI)?*
- *Does the repository allow the upload of intrinsic (e.g., author's name, content of dataset, associated publication, etc.) and submitter-defined (e.g., definition of variable names, etc.) metadata?*
- *Is it clear under which license (e.g., CC0, CC BY, etc.) the data will be available, or can the user upload/choose a license?*
- *Are the citation information and metadata always (even in the case of datasets with restricted access) publicly accessible?*
- *Does the repository provide a submission form requesting intrinsic metadata in a specific format (to ensure machine readability/interoperability)?*
- *Does the repository have a long-term preservation plan for the archived data?*

4.4 Are the digital repositories I chose maintained by a non-profit organization?

Although not mandatory, SNSF strongly encourages the usage of repositories held by a non-profit organization for data storage and archive. To assess if the desired solution is considered non-profit by SFNS standards, please refers to the following checklist:

1. *The first step is to consult www.re3data.org, where most repositories are listed.*
2. *Under the tab "Institutions", check if a commercial entity is involved in 'general' or 'technical' responsibility (categories "Type of institution" and "Type(s) of responsibility")*
    a. *If not the case, SNSF considers the repository to be non-commercial (even if 'funding' or 'sponsoring' is provided by a commercial entity).*
    b. *If a commercial entity is listed, the SNSF considers the solution to be a commercial repository.*
3. *If the repository is not listed on www.re3data.org, the repository should be contacted to clarify this point. Researchers should also suggest that the repository be included in www.re3data.org*

In case of usage of a commercial repository, please explain your choice.

4 Useful links

Data Management Plans reference templates:

- SNSF Data Management Plan - Guidelines for researchers – Link
- EPFL/ETH Data Management Plan template – Link

SPHN BioMedIT:

- The BioMedIT network – Link
- SPHN BioMedIT security policy – Link
- BioMedIT security concept – Link
- Secure transfer tool (sett) – Link

SPHN ethical and legal framework:

- SPHN ethical framework – Link
- SPHN legal agreement templates – Link

SPHN semantic framework:

- SPHN semantic interoperability framework – Link
- SPHN RDF Schema – Link
- Full documentation – Link

Other:

- Swiss Federal Act for Data Protection (FADP) – Link
- Human Research Act (HRA) – Link
- FAIR principles – Link

5 Annex: SPHN data types

Data streams	Concepts/data types	Examples/remarks
Clinical routine data		
	Demographics	Age, gender
	Diagnoses	
	Problems	
	Allergies	
	Immunization	
	Medication	Dose, administration route, galenic
	Procedures	
	Lab orders and values	Clinical chemistry, hematology, microbiology
	Vital signs	Heart rate, body temperature
	Biosample data	
	Clinical reports	Discharge letter
	Pathology/Histology data	
	Other data types	Free text, surveys, etc.
	Genomic data	WGS, WES, gene panels
	Imaging data	MRI, CT,
	Multimedia data	Echoes,
	Waveform data	ICU monitoring
	Patient generated data	Health / treatment history, symptoms, bio-metric data, PREMS, PROMS
Cost and utilization data		
	Administrative data	
	Claims data	
Family history		
	Pedigree information	
Clinical research data		
	Cohort data	
	Registry data	Cancer registry data
	Clinical study data	Collected in eCRFs
	Public Health data	environmental data, surveillance data
	Genomics	WGS, WES, gene panels
	Transcriptomics	RNA-seq
	Proteomics	
	Metabolomics	
	Epigenomics	Methyl-seq
Healthy citizen data		
	Citizen/Consumer health data	
	Lifestyle data	
	Social media data	
	Mobile health	Wearable data, fitness tracker, sensor data
	Tracking apps	