Scriptum for SPHN Data Privacy and IT Security Training

Version 3.0, 16 November 2020

Editors: Heinz Stockinger (SIB Swiss Institute of Bioinformatics), Diana Coman Schmid (ETH Zurich & SIB), Sofia Georgakopoulou (University of Basel & SIB)

Overview

Within the Swiss Personalized Health Network (SPHN) and related national initiatives researchers use patient data (i.e., confidential human data) in their research projects. Dealing with confidential human data requires awareness of data privacy, respective laws and information security. The SPHN course **Data Privacy and IT Security Training** explains what should be done in practice to protect the patients' privacy when performing biomedical research on human data. This scriptum provides a written summary of the course.

Applicability and target audience

This training is designed for SPHN project members who (plan to) use the BioMedIT infrastructure in a current or future research project using sensitive human data. In particular:

- Researchers ("Users" of the BioMedIT Infrastructure)
- Project leaders
- IT Personnel who operates the BioMedIT Node infrastructure

Objectives

- Train researchers, project leaders and IT personnel who use or operate confidential human data for research projects (focus on SPHN)
- Provide overview of obligations and practical aspects when using human data on IT infrastructures (e.g., BioMedIT Nodes)

Goals for participants

- Have a good understanding of data protection and privacy issues when dealing with confidential patient data in research projects
- Have the necessary knowledge to conduct research respecting both legislation as well as information security requirements
- Apply specific procedures and follow certain guidelines

Table of Contents

Version history

Version 1.0	4 Sep. 2019	First version
Version 2.0	24 July 2020	Updated URL of ISP
Version 3.0	16 Nov. 2020	Updated URL of BioMedIT to sphn.ch/biomedit

1 Introduction

When using patient data in research, special care needs to be taken in order to protect the data was well as the patients themselves. This is illustrated by the following provocative and fictive example:

Professor Mustermann of University XZY tells the following to a PhD Student: "Yesterday, I met my friend François from the university hospital. They are doing some research on patients having ovarian tumor. They've recently got really interesting data on this type of tumor. You know Teresa, a PhD student of François. Can you please contact her and get all the exome sequencing data from the cancer patients? We should really analyse this data since I have an idea how we can potentially help these patients."

In this example, we already see a few issues:

- Did the patients actually agree that their data would be reused in research? i.e., did they give **consent**?
- One can't just get patient data from a friend without any written permission.
- One can't just transfer data in clear text and use it on any personal computer or laptop without following certain IT security standards and requirements.

The goals of this course is: to **make you aware of possible issues and know how to handle them in a correct way.**

In summary: when dealing with human data (also referred to as "personal data") specific care needs to be taken when using such data in a research project. For instance, you have a research project on cancer data of patients from Swiss hospitals. Data related to humans are often considered to be "sensitive", i.e., need to be protected specifically, and researchers need to know how to deal with that. Later, we will have a more careful look at what it means that data are "sensitive" – for now, we just assume that **we need to be careful when we use and/or share such data**. Often, we are not allowed to use and/or share them like we would do with 'open access data', for example the mice genome files. This could cause conflicts or other issues.

2 Data Privacy and Protection

2.1 Open Access and Open Research Data

In the last decade, a new trend was established with respect to publishing of research results: **open access** to publications (journals, conference papers, etc.). This is in contrast to previous practice where many of the publications were only available when paying a fee. Several scientists still have to get used to the open access method, and it is still a learning process for some scientists. With open access, anybody can freely access data or publications without access restrictions. In addition, journals often allow to publish supplementary material that should also be accessible for free. This is supported by the concept of **open research data** where scientific results or data should also be freely accessible. Internationally, this is known via the **FAIR principles** where data should be: finable, accessible, interoperable and reusable. The FAIR principles enable that data can be usable by others but there might be access restrictions, i.e. not freely accessible to just anyone.

In Switzerland the Swiss National Science Foundation (SNSF) also supports the concept of open science and open research data[1].

2.2 Data Privacy, Protection and Restricted Access

Even in times of open access and open research data, there are many data items that are **personal** and cannot be shared via open access (unless explicit permission for open access has been given). **Data privacy** must be respected – particularly, when dealing with **human data.**

Currently, for patient data to be used in research, an agreement (i.e., **informed consent**) either **generic** (example) or **specific** to a study (example) is required from every research subject - see Section 5.1. In the European Union (GDPR), an explicit consent is one of the requirements for processing personal data (https://gdpr-info.eu/issues/consent/). Research projects running in Switzerland may follow the *swissethics* guidelines to align with the EU GDPR (https://www.swissethics.ch/doc/ab2014/Addendum_GDPR_v1.0_d.pdf).

In the media, we often hear that hackers or even internal IT staff members steal data (physically or over the internet) and sell the data. We have seen that in the context of banking and tax fraud. However, similar security incidents can involve health-related human data. There might be people/organisations that are interested in personal data and prepared to commit data breaches.

A **data breach** is a security incident in which sensitive, protected or confidential data is copied, transmitted, viewed, stolen or used by an individual unauthorized to do so. Unintentional release of private/confidential information to an untrusted environment is also a data breach.

In the event of a data breach, data are in the hands of **unauthorized people.** We need to prevent this from happening, particularly for sensitive personal data that we use in research projects. A breach might have serious consequences for you, your institution or an entire user/research community.

In summary, we need to protect sensitive personal data: physically and via contractual measures.

2.3 Data and Security Awareness

We currently produce huge amounts of data: 2.5 quintillion bytes of data every day. In the last two years alone we produced globally 90% of all data. By 2020, for every person 1.7 MB of data will be created every second, amounting up to 50 TB per year per person[2,3]. Parts of this data is **personal data** produced by people as part of our social digital life. For example, data collected while using

[1] http://www.snf.ch/en/theSNSF/research-policies/open_research_data/

[2] https://www.domo.com/learn/data-never-sleeps-6

[3] https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#51b8b4cf60ba

the Internet, data posted on social media platforms, etc. Furthermore, we generate data related to our own health (i.e., **sensitive personal data**) when opting for having our own genome profiled (direct-to-consumer DNA kits), our physical activity tracked by wearable devices (step counter, heat rates, stress levels, etc.), participating in clinical studies and in research projects. Distinct from the personal data generated by the individuals themselves, large volumes of non-personal data are collected and used, for example the data for airline tracking or for weather predictions.

Compliant with current legislation, services collecting data request an explicit agreement from each individual via the “Terms of Service” forms. However, reading the complete text and properly understanding the specification of such legally binding contracts is an overlooked task. Given that the average person reads at a rate of 200 words per minute, while a standard “Terms of Service” agreement contains about 12’000 words[4], it would take up to one hour for reading one such contract. By simply signing such agreements, without being informed on the full life cycle of the collected data, we may lose control of the data we produce. As a consequence, personal data, both public and sensitive, in practice and from the perspective of the individuals who contributed them, are handled as open access.

For example, data brokers (e.g., acxiom.com, equifax.com) collect personal data from various services to which individuals agreed under Terms of Services, build individual digital profiles (e.g., age, gender, education, job, online shopping purchases, etc.) and then trade personal data for real money. Noteworthy, not only the actual data but also the metadata (describing the data) are a valuable asset because it allows to make connection between raw data files.

It is estimated that currently 2.7 billion personal profiles exist in such broker databases, amounting to almost 35% of all human population. Not surprisingly, such databases are subject to data breaches. For example, Equifax suffered a major breach in 2017[5]. Several terabytes of consumer credit information from 140 million people - about half the population of the USA - was taken. It could have been prevented, had it used basic IT security measures.

2.4 Open Access vs. Restricted Access

Some of the personal information and data we post on social media may be public, i.e., it is open access and, thus, it may be used by anyone, even for profit. Oppositely, restricted access data requires additional security measures to protect the confidentiality of the data and the privacy of the individuals who contributed them.

An illustrative example is the open access research data vs. the restricted access private banking data. Whereas open access is paramount for sustainable research communities, it is instrumental that banking transactions are secure. We therefore accept and adapt to strict digital banking usage rules set according to the relevant legal frameworks. In this context, medical data (e.g., collected routinely in hospitals and approved for re-use in research) would require simultaneously an open and restricted access policy: the privacy and confidentiality should be preserved (similar to banking) and at the same time it should also foster research (similar to the FAIR principle in research).

[4] https://profjourde.wordpress.com/2018/05/12/visualize-our-submission-to-the-tos/
[5] https://techcrunch.com/2018/12/10/equifax-breach-preventable-house-oversight-report/

For example, the following situations require a blend of the security and usability concepts:

- *Modern science - modern questions*: should a microbiome data file be handled securely i.e., as sensitive personal data derived from a human biological sample even though only bacterial species information is contained?. Alternatively, can it be processed like regular bacterial (i.e., non human) data, lowering the usability barrier? Recent research showed that based on the gut bacterial make up alone, one can re-identify patients[6,7]. Was this possible 5 years ago? Given the latest research discoveries, some data types previously considered as non sensitive, may classify as sensitive personal data.
- *Health data persists in time*: if a bank account is hacked, money is lost once. However, if human genome data are stolen, the genetic information will not change, except for the few mutations that might occur during one's life.
- *Health data is not only about "me"*. If human genome data of individuals become public, this implicitly concerns their children and close relatives as their hereditary genetic conditions will be public as well.

Thus, it is up to each one of us to be aware of data security, both of our own personal data and when we do research with data of others.

2.5 Examples of Data Breach

Every year, we can find several examples of data breaches where data are in the hands of people who are not authorised to have access to these data. This can be the case because of criminal acts or simply because of not being careful with data protection. We look at three examples to show you what can go wrong and what are the potential consequences of not handling sensitive data securely. This is to stress that privacy and legal topics need to be taken very seriously:

- "Portuguese Data Protection Authority Imposes EUR 400'000 Fine on Hospital"[8]: Organisations and even hospitals get **fines if data are not correctly protected**. In this case, unauthorized people were able to access patient data.
- "Hospital staff disciplined after Ed Sheeran data breach"[9]: One member of hospital staff has been sacked and another has been given a written warning for accessing Ed Sheeran's personal details without authorisation, it has emerged.
- "More than 200'000 patients' records were exposed on MedEvolve's public FTP server"[10]: Important: even if one was allowed to have access to the data, it must not be made public. Password protection in general is good but not enough in such a case because passwords can be hacked, exposing sensitive personal data to public access.

For more examples, see: https://ocrportal.hhs.gov/ocr/breach/breach_report.jsf

[6] Franzosa et al., 2015, Identifying personal microbiomes using metagenomic codes, PNAS

[7] Wagner et al., 2016, Privacy-preserving microbiome analysis using secure computation, Bioinformatics

[8] https://www.datenschutz-notizen.de/portuguese-data-protection-authority-imposes-400000-e-fine-on-hospital-4821441/

[9] https://www.telegraph.co.uk/news/2018/05/19/hospital-staff-disciplined-ed-sheeran-data-breach/

[10] https://www.databreaches.net/more-than-200000-patients-records-were-exposed-on-medevolves-public-ftp-server-researcher/

3 Laws

What we need to protect and which measures we should implement is defined in various laws, both nationally and internationally. Many of you have heard about the new **European regulation: GDPR (General Data Protection Regulation)** – which gives more rights to the individual such as:

- Article 20: data portability (right to have a copy of personal data)
- Article 17: right to withdraw data ("right to be forgotten").

GDPR is only applicable in specific cases in Switzerland if data from EU residents is involved. Let us first look at the most important Swiss laws:

- The foundation is laid out in the **Federal Act on Data Protection**[11] which defines **"sensitive personal data**" that includes health data.
- The **Human Research Act**[12] goes into details when using human data in *research*, for example: secondary use of collected patient data for research purposes.
- In case things go wrong (i.e., data breach, criminal acts, etc.), the **Swiss penal code** applies. For instance, article 321 covers "Breach of professional confidentiality in research involving human beings".

Note that for clinical treatments and clinical trials, other laws are applicable (Federal Act on Human Genetic Testing, Electronic Patient Record Act) but we will not discuss these further in this training.

3.1 Definition of Personal Data

The Swiss Federal Act on Data Protection (FADP), article 3a defines **personal data** as follows:

> all **information** relating to an **identified or identifiable person**. For example, the name or address of a person.

Specifically**, *sensitive*** personal data (FADP, art. 3c) is defined as:

> (i) religious, ideological, political or trade union-related views or activities;
> (ii) health, the intimate sphere or the racial origin;
> (iii) social security measures; or
> (iv) administrative or criminal proceedings and sanctions

Sensitive personal data (e.g. health data) need particular security measures to protect the privacy of a person.

[11] https://www.admin.ch/opc/en/classified-compilation/19920153/index.html
[12] https://www.admin.ch/opc/en/classified-compilation/20061313/index.html

Sensitive personal data can directly identify individuals or it can contribute to the identification of persons.

Examples of personal data that can **uniquely identify a person** (**directly identifying data**):

- name
- address, phone, email
- birth date/place
- ID number
- biometric information (incl. finger prints)
- genetic information

In the USA, **HIPPA** and PHI (Protected Health Information) have defined 18 identifiers that can be used as examples (https://en.wikipedia.org/wiki/Protected_health_information) but are not explicitly mentioned in the Swiss legislation:

- Names
- Geographical identifiers
- Dates (other than year) directly related to an individual
- Phone Numbers
- Fax numbers
- Email addresses
- Social Security numbers
- Medical record numbers
- Health insurance beneficiary numbers
- Account numbers
- Certificate/license numbers
- Vehicle identifiers and serial numbers, including license plate numbers;
- Device identifiers and serial numbers;
- Web Uniform Resource Locators (URLs)
- Internet Protocol (IP) address numbers
- Biometric identifiers, including finger, retinal and voice prints
- Full face photographic images and any comparable images

Other data that **can contribute to identity a person** if used in context are called **indirectly identifying data**: gender, job position, IP address, hair colour, blood sugar levels, daily movements, height, etc. Examples:

- *If you only have the information about a gender, hair colour or a blood sugar level, you cannot identify a person. However, if used in context or related to a name, address, this data need to be treated with care because it may lead to revealing the identity of persons.*
- *If you know that the data set contains 10 people from a small village with 200 people and you find a person that has a certain job position (e.g. CEO of a large, multi-national company), it might be very easy to guess who the person is without having the name explicitly mentioned in the dataset.*

4 Data Classification

In this chapter, we provide guidelines for data classification with a specific focus on research data types like genomics, proteomics or pathology images. Additionally, we look at terminologies and definitions and explain with concrete examples the differences between anonymized, pseudonymized, coded and encrypted data.

4.1 Use Case: Data Flow From Hospital to Researcher

Before we go into details of data classification, let us look at a very common use case of how researchers get access to patient data from hospitals, and which issues need to be considered:

1. In hospitals, data are routinely collected from patients. If they give their consent, such data may be used for other purposes than the ones it was originally collected for: for example, such data can be re-used in biomedical research.
2. A copy of the data is transferred to the Researcher. To the Researcher, these data are "anonymous" – the names of patients are not revealed. This means that the data are de-identified i.e., the information that identifies patients is masked or deleted.
3. In specific cases, research findings might be reported back to the patient by the medical staff in hospitals. For example, in case medically relevant information is detected, doctors could contact the patient if re-identification is possible.

Overall, there are three important aspects:

1. A patient ("data subject") needs to give informed **consent (in writing with signature)** that the data collected from her/him may be used for research.
2. A Research Project must have **permission to reuse the data for research:** an ethical approval needs to be provided by an authority such as an Ethical Committee.
3. A **contract** is needed to **transfe**r data from the hospital to a Researcher. Example: DTUA (Data Transfer and Use Agreement) that specifies, for example, the access restriction (which researchers can work with the data; no sharing of data with other research projects without explicit permission).

Restricted access comes with more security measures. Once transferred to the Researchers, these **data have to be handled securely**: this is the main objective of this training and the related technical and organisational measures.

4.2 Data Classification in SPHN

In movies we often see that a certain project or information is classified as "top secret". Such a classification (or categorisation) is very useful since based on the classification we can decide which access restrictions and related security measures must be applied.

In SPHN, three classes are used:

1. **Confidential** – highest classification with respect to **access restriction and risk** in case of data leaks or data get in the hands of non-authorized people. Medical records, for example,

classify as confidential data. Other examples of confidential data are: financial information or human resource data.

2. **Public** – opposite of confidential – data can be seen by anybody. For example, press releases, scientific papers with open access, published research data from repositories.
3. **Internal** – data shared only with specific people, e.g. the project budget, names and e-mails of project members.

In summary, in SPHN the term “confidential data” is equivalent to “sensitive personal data” as defined in FADP. Note that this classification is *not* used as such in the Swiss law but it is commonly used **in information security contexts** in several parts of the world. Usually, a project specific **policy defines the details of these classifications.** In the case of SPHN, the **SPHN Information Security Policy (ISP)** defines this classification.

Examples of confidential data related to patients:

- **Clinical data**: focus on what is used in SPHN (e.g., routinely collected patient data in hospitals: blood samples, diagnosis reports, family history, microbiome, DNA sequence etc.)
- **Health data**: already mentioned by the law as “sensitive human data”; for example, the health data recorded by individuals with the help of wearable devices or phone apps.

Research-related data that are not confidential may classify as internal (e.g. project budget) or public (e.g., open access research repositories).

In general, SPHN classifies data as follows:

> **All personal data (either identifying data or pseudonymized) are confidential unless explicitly classified differently.**

4.3 Examples

Confidential data:

Genotyping human data (DNA sequencing)

- WGS (Whole Genome Sequencing) data
- WES (Whole Exome Sequencing) data
- Specialized genomic panels (e.g. cancer panels)
- Single cell sequencing, CHiPseq, ATAC-seq
- Information in some types of Quality Control files (e.g., FASTQC files may contain human sequences (“overrepresented sequences”) in plain text), and this makes such QC reports confidential

Examples: the content of

- raw sequence reads in **FASTA** or similar formats
- **VCF** (Variant Calling Format)
- **SAM** (Sequence Alignment/Map) and **BAM** (binary version of SAM)

Here are some more examples, without being an exhaustive list. (Note that the Swiss legislation is currently not very specific about which biological data type is confidential data.):

- Transcriptome
- Patient derived human cell lines
- Proteomics data
- Metabolomics
- Biobanks and cohorts
- Imaging data
- Immunological screening
- Controlled access data in TCGA, dbGAP

In the examples above it is good to assume that data are confidential (conservative approach) but to be sure, one needs to check with a legal department on a case by case basis. Currently, there is no official list that would say with 100% certainty a certain data type is confidential! There might never be such a list but these are **guidelines for now**.

Some data types are not identifying (e.g., proteomics, metabolomics or pathology images). However, the **combination or aggregation with other identifying data** (e.g., genomics) increases the risk of re-identification and, thus, such data are classified as confidential. Data aggregation is common in the SPHN, PHRT or related biomedical projects, and various types of data (e.g., genomics and clinical images) are usually stored in the same data center.

Example:
Let's assume we have various data types available for several patients: genomics, transcriptomics, clinical, imaging. There is no identity of patients in individual files or data sets irrespective if it is genomics or a pathology image. Individual data sets are aggregated for advanced analyses. By doing so, links between separate information are made, and we can for example find out, legitimately, that one of the female patients was diagnosed with ovarian cancer: we know now her genomic profile, her clinical data, proteomics etc., all based on her unique research patient ID. Illegitimate re-identification can reveal the real name of our female patient. This is not the purpose of research. It requires some effort and additional non-medical data (e.g., social media, election, local news, etc.). Such re-identification is not the purpose of research, and it is illegal. Having various and many sensitive personal data available (as effect of biomedical research in need of aggregating data sets, usually in the same data centers) increases the risk of illegitimate re-identification.

Public data:
Data in various public knowledge bases and archives:

- 1000 Genomes, ExAC, PDB, UniProtKB, etc.
- open access data in: TCGA, ICGC, SRA, genome-phenome archives, etc.

Internal data:
Examples: name of all users involved in SPHN projects; financial information, etc.

4.4 Steps to make data available (Data Provider)

The following are the main tasks for a Data Provider in a hospital to prepare data for later re-use in research:

- Collect consent from data subjects
- "Filter data". This also includes the de-identification step to make sure that patients' directly identifying information is not available to researchers
- Package data in a way it can be reused later. One might also apply FAIR principles when making data re-usable
- Classify data as "confidential" to be sure external users in research protect the data accordingly

4.5 Decision tree for data re-use in research (Users)

We now take the point of view of the Users and show a possible scenario for handling securely confidential patient data in a biomedical research project. We need to be aware if the data can be re-used for research (see Section 4.4), and we assume that the data was already prepared by e.g., at a hospital, and the data set is available. A researcher wants to use that data. In this case, the following guidelines should help Users:

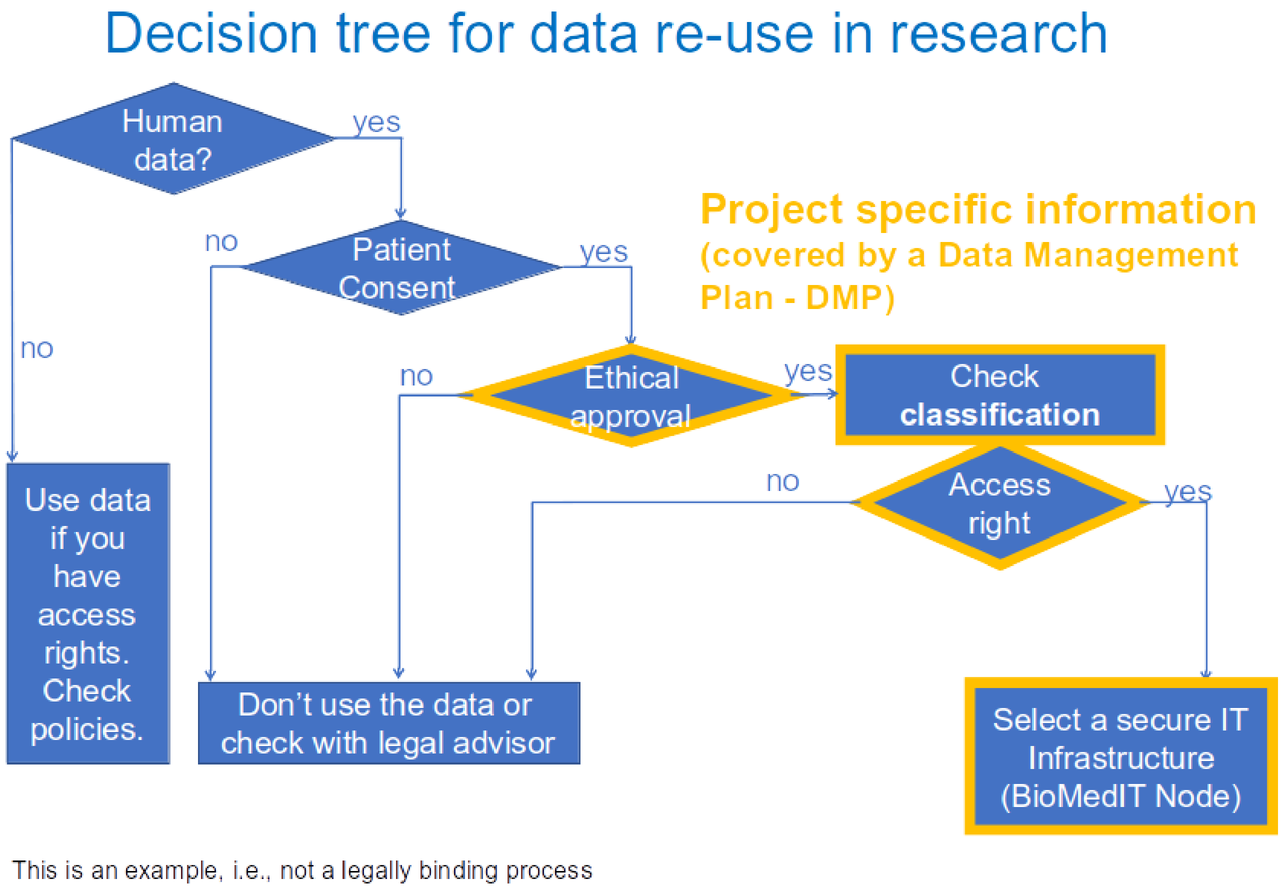

Decision tree for data re-use in research
Human data?
yes
no
Patient Consent
yes
no
Project specific information (covered by a Data Management Plan - DMP)
Ethical approval
yes
no
Check classification
Access right
no
yes
Use data if you have access rights. Check policies.
Don't use the data or check with legal advisor
Select a secure IT Infrastructure (BioMedIT Node)
This is an example, i.e., not a legally binding process

Figure 1. Data re-use decision tree

- If they are **humans data** or **data related to humans** (e.g., bacteria found in human gut), be careful and alarmed that these data might need special protection.

    - Even if data are not related to humans, you might still not have the explicit access/usage rights because data could be protected (e.g., until published by the data producer).
    - Data could also have an open access license: then you can use the data in research but not necessarily in industrial settings.
- Data Provider needs to check the **consent** for data re-use in research. Users should check if the data was packaged/prepared correctly (i.e., subjects have given their **consent**).
- **Ethical approval** is needed for your research project (in some specific cases, this may not be required).
- Check **data classification** – if available: public – internal – confidential. In some cases, data might be public even if they are sensitive personal data: might have been explicitly classified as "public" (e.g., open access in dbGAP)
- Carefully check which **access rights** you have (e.g., via a **Data Transfer and Use Agreement**). If you do not have them yet, your Project Leader might request them.
- Finally, if data are classified as "confidential", use a secure IT infrastructure (e.g., BioMedIT Node) to process your data.

Note that project related information needs to be stored in a Data Management Plan.

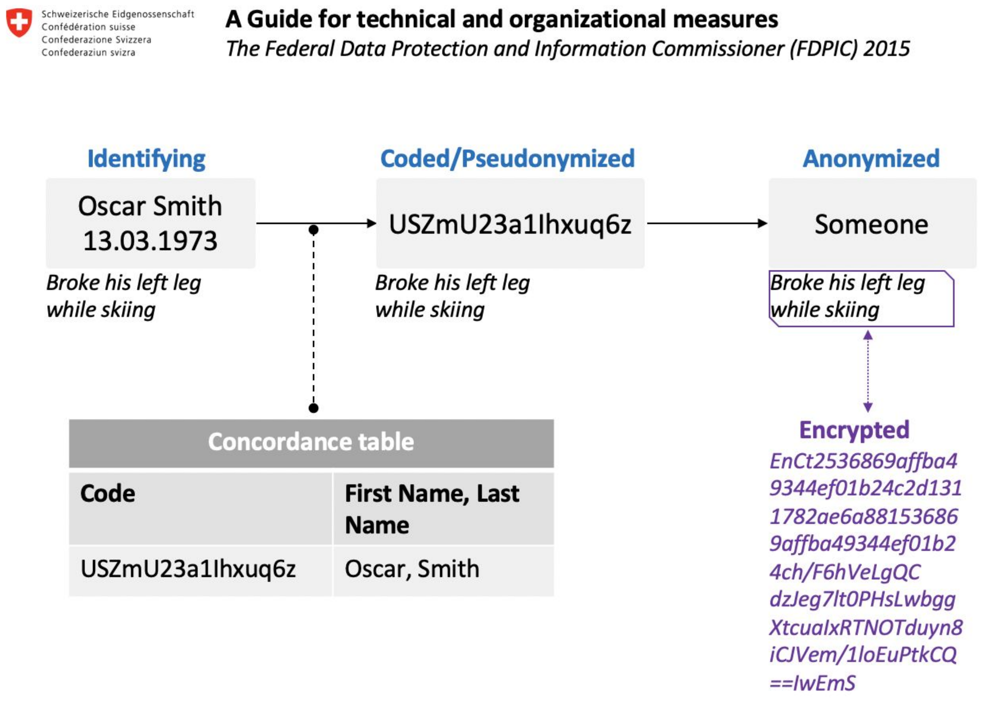
4.6 Representation of Data

Terminology:

- **clear text:** data are readable by anyone, and all identifying information is visible.
- **de-identification:** process used to prevent a person's identity from being connected with information, i.e., the identity of a person cannot be obtained anymore:
    - **pseudonymization** (used in context of GDPR): substitutes the identity of a data subject in such a way that additional information is required to *re-identify* the data subject
    - **coded:** personal data and human biological material linked to a specific person via a code (cf. SPHN Glossary[13])
    - **anonymization:** irreversibly destroys any way of identifying a data subject. *Note that anonymization must not be confused with pseudonymization!*
- **re-identification**: process of matching anonymous data (also known as de-identified data) with publicly available information, or auxiliary data, in order to discover the individual to which the data belongs to.
- **encryption:** processes of encoding a message (cf. SPHN Glossary)

Examples:

[13] https://www.sphn.ch/en/news-events-publications/publications.html



Code	First Name, Last Name
USZmU23a1lhxuq6z	Oscar, Smith

Figure 2. Examples for different data representations

Coding: Identifying information is **replaced by a code** (e.g. concordance table) and only accessible with a **"key"** under strict security regulations.

Anonymisation, HRO[14] (art. 25): "For the anonymisation of biological material and health-related personal data, **all items which, when combined, would enable the data subject to be identified without disproportionate effort, must be irreversibly masked or deleted**.In particular, the name, address, date of birth and unique identification numbers must be masked or deleted."

Note that research with anonymized data is *not* subject to HRA nor GDPR.

One of the following conditions needs to be satisfied in order to break the code, i.e., **reveal the identity of a person:**

- breaking the code is necessary to avert an immediate risk to the health of the person concerned;
  - Example: a very severe and imminent disease was detected and the person might die if she does not yet immediate care
- legal basis exists for breaking the code;
  - Example: There is a law suite that requires information about the person, e.g., a court decision was taken to break the code
- breaking the code is necessary to guarantee the rights of the person concerned, and in particular the right to revoke consent.
  - Example: a person decided not to give consent anymore – need to find her data and remove it

Note that it is ***up to legal institutions/bodies*** to decide in practice if the code can be broken. There should a particular procedure with responsibilities in place.

[14] https://www.admin.ch/opc/en/classified-compilation/20121177/index.html

In summary, de-identification is a process that prevents that the identity of a person is revealed. It can be implemented by pseudonymizing, coding or truly anonymizing the data. Data is anonymized if it is impossible to reveal the identity of the person. In contrast to anonymized data, with coded or pseudonymized data, the concordance table linking the code to an identity does exist (i.e., safely stored at hospitals), and thus there are ways to uniquely identify the data (useful when revealing the patient identity is lawfully required).

5 IT Security

Up to this point, we have discussed a lot about laws and regulations, as well as data classification. In this chapter, we go into the more practical aspects of how IT security can help to protect data.

5.1 Data: consent and ethical approval

We now assume that the data was consented by data subjects, and ethical approval was granted. We also assume that you or your Project Leader have access rights to use confidential human data for research. What are the factors that you should keep in mind regarding your data when starting your project?

First of all, it is important to remember the type of consent you have acquired: data subjects may sign a **general consent** (e.g., not knowing which type of research project may request to use the data) or a **specific consent** (e.g., being explicitly informed that the data will only be used to study the sleep patterns in babies associated with their gut microbiome). Moreover, the **ethical approval** may impose restrictions, e.g., primary research data must be published under strict controlled access. Every project member needs to be aware of these issues! It is in the Project Leader's responsibility to make sure the project members are informed, are behaving appropriately and know where the limits are.

5.2 Two Factor Authentication (2FA)

Let us look at some examples from our everyday life that we are all using. Think about your bank account. Nowadays, you can access it online to check your balance and make transfers and payments. How do you do that? To get online access you need 2 Factor Authentication (2FA). But what is this exactly? A factor in this sense can be one of three things: It can be something you *know* (such as a password), something you *have* (such as a smart card) or something you *are* (such as a fingerprint or other biometric method). 2FA is a combination of one item from each of two separate categories.

Having 2 factors (e.g., a password and a second piece of information you own) makes access restrictions more secure. It is already in place for accessing e.g., your bank account and credit card information, and it is now becoming a standard when dealing with sensitive human data as well.

There are some potential issues with 2FA you should keep in mind: if the second factor is something you have, it ideally should not be present on the device you are using to access the infrastructure. This is a problem with using e.g., a phone to access e-banking. A code sent to the phone (by SMS or App) is no longer a security enhancer when logging in on the phone. However, it is fine to use a portable computer to login and a mobile phone to access a second factor.

5.3 The BioMedIT tools and support - what you need to know

In order for researchers to work with sensitive data within the SPHN projects, they need specific and secure tools that help them protect the data: BioMedIT provides this specific, secure IT infrastructure that allows you to do research on human data.

Users need to be aware of that within this infrastructure one cannot behave in the same way as one would do within a local computer. Use of this infrastructure requires security awareness and specific practice/actions – remember that a BioMedIT Node is shared with several users! You can only use it for specific projects where you have rights to access the data.

The BioMedIT personnel will help you with data handling and guide you through a secure environment.

5.4 A use case from a researcher's point of view

Now, let us look into what a researcher within SPHN needs and what are the steps she needs to take to start working (see also Figure 3).

1. This is Sarah, a Cancer Researcher. She is studying lung cancer. In order to build her project she channels a data query via the Data Coordination Centre (DCC). She learns that three Swiss hospitals have patient data relevant to the project.
2. Sarah submits her project proposal, detailing the type and proposed usage of patient data, to an ethics committee.
3. Sarah signs a "Data Transfer and Use Agreement" (DTUA) with the hospitals to formalize the data use rights and possible Intellectual Property derived from the research.
4. Sarah registers the project and members of the research team with the DCC and obtains access rights.
5. The hospitals encrypt and copy the relevant, de-identified and consented data from their lung cancer patients to a BioMedIT node.
6. Sarah can now access the protected data stored in the nodes, run her analysis and proceed with her research project.

Once the project is completed, data are either securely archived or deleted.

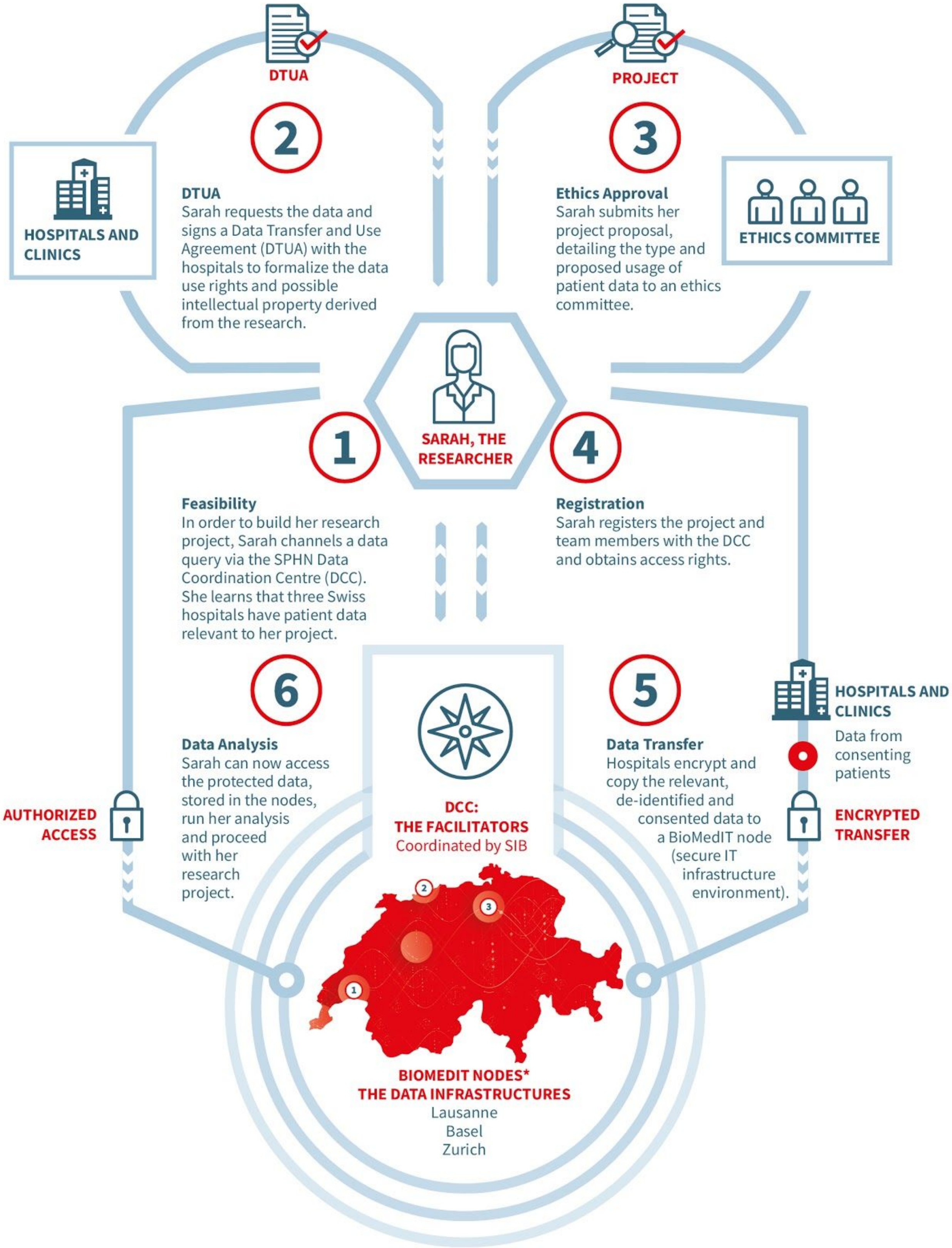

DTUA
PROJECT
HOSPITALS AND CLINICS
2
DTUA
Sarah requests the data and signs a Data Transfer and Use Agreement (DTUA) with the hospitals to formalize the data use rights and possible intellectual property derived from the research.
3
Ethics Approval
Sarah submits her project proposal, detailing the type and proposed usage of patient data to an ethics committee.
ETHICS COMMITTEE
SARAH, THE RESEARCHER
1
Feasibility
In order to build her research project, Sarah channels a data query via the SPHN Data Coordination Centre (DCC). She learns that three Swiss hospitals have patient data relevant to her project.
4
Registration
Sarah registers the project and team members with the DCC and obtains access rights.
6
Data Analysis
Sarah can now access the protected data, stored in the nodes, run her analysis and proceed with her research project.
5
Data Transfer
Hospitals encrypt and copy the relevant, de-identified and consented data to a BioMedIT node (secure IT infrastructure environment).
HOSPITALS AND CLINICS
Data from consenting patients
AUTHORIZED ACCESS
ENCRYPTED TRANSFER
DCC: THE FACILITATORS
Coordinated by SIB
2
3
1
BIOMEDIT NODES*
THE DATA INFRASTRUCTURES
Lausanne
Basel
Zurich

Figure 3. A typical analysis workflow

Currently, many of the steps above are in place in SPHN. However, data management aspects such as "query at DCC" and a "data catalogue" are not yet available - they are currently work in progress.

6 SPHN - BioMedIT Infrastructure

We have talked about SPHN and the requirements for an IT infrastructure. As mentioned earlier, this infrastructure within SPHN is called **BioMedIT** (sphn.ch/biomedit), and we will look into it in more detail in this chapter.

6.1 BioMedIT network in Switzerland

As we have discussed, SPHN.ch is a Swiss national initiative. It encompasses research projects, offers grants, etc. On the other hand, **BioMedIT is the IT infrastructure** in SPHN. It is not just a single computer or single infrastructure but a *network* of local infrastructures that is restricted to the trusted, non-public networks of the participating institutions and can be otherwise accessed by VPN. The Infrastructure is necessary so that we can move and use health-related, personal data. Coordinating the data is the task of the Data Coordination Centre (DCC, sphn.ch).

One of the main tasks is to ensure **interoperability**: the goal is that researchers, who access various types of data from different nodes, are able to work with this data seamlessly without even noticing any variability: essentially, the federation of sources does not pose a disruption in the data analysis process.

As an example of interoperability for health related data, let us imagine you are working in the university hospital of Zurich: all data are in German (incl. attributes such as name, patient history, etc.). If you then move to Geneva or you want to combine data from a lab in Geneva, it becomes difficult to complement and/or exchange data if there are different attributes, data formats, etc. The aim is to have data that are interoperable and can be used across different locations, hospitals, etc.

Interoperability is achieved through the usage of standards for data formats, semantics, governance, and exchange mechanisms among others.

DCC additionally coordinates **user access rights**, project registry, monitoring of data usage, etc.

BioMedIT and DCC are coordinated by the SIB Swiss Institute of Bioinformatics (sib.swiss) and in particular by the Personalised Health Informatics (PHI) group.

So let us look at the big picture, at the actual SPHN and BioMedIT network. SPHN includes university partners and hospitals from many different cantons, and all are tasked to work together under the umbrella of SPHN (see Figure 4).

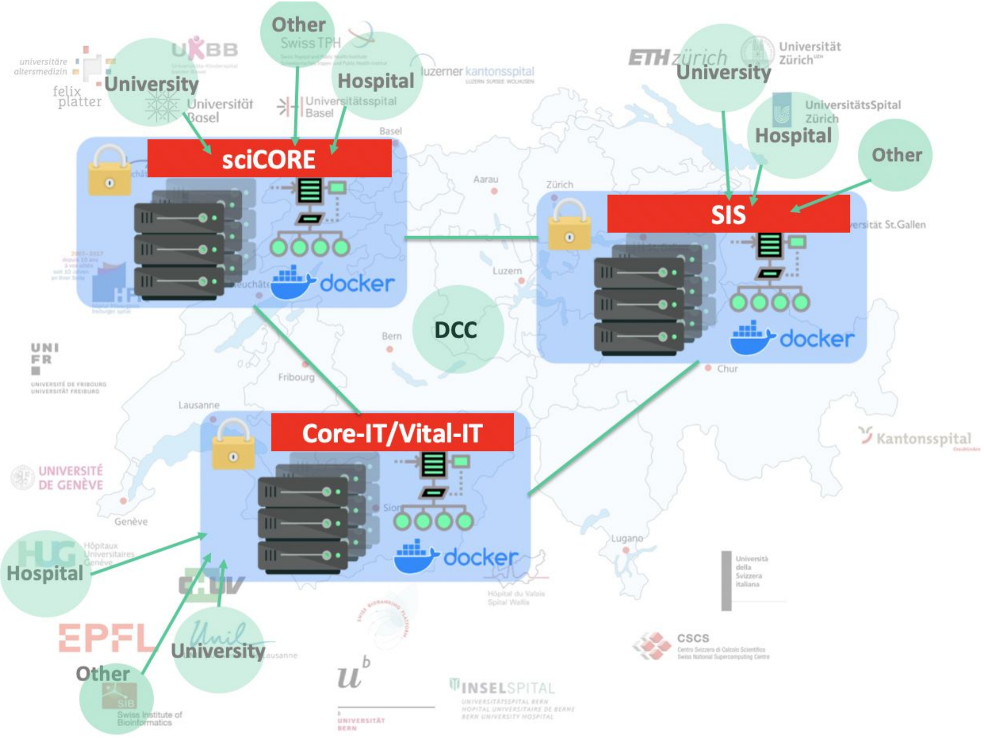


Figure 4. Overview of 3 BioMedIT Nodes in Basel, Lausanne and Zurich

We have currently 3 BioMedIT nodes:

1. Zurich (ETH Zurich) – managed by the Scientific IT Services group (SIS)
2. Basel (University of Basel) – managed by the sciCORE group
3. Lausanne (SIB) – managed by the Core-IT group (scientific and data management support in collaboration with the Vital-IT group)

This is where your projects are running, and you are a researcher e.g., from ETH Zurich, UniSpital Basel or HUG Geneva, etc.

There is the flexibility in the system to accommodate additional nodes in the future.

Now let us zoom into one of the nodes …

6.2 What does a BioMedIT node look like?

A BioMedIT Node consists of hardware/software as well as scientific/technical experts (people) to support scientific projects. All this is encompassed here in the blue box. In the red box you see the zone for confidential data:

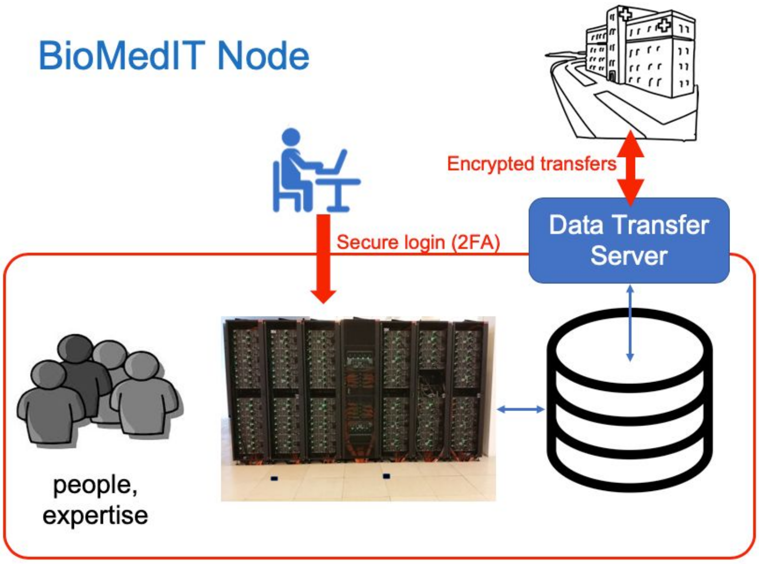


Figure 5. Overview of a typical BioMedIT Node

Data Transfer Servers are employed to get encrypted data into and out of the node. Authorized users can log into the IT infrastructure securely with 2FA and access the data.

6.3 Secure data transfers

At the time being (summer 2019), some of the work in SPHN and BioMedIT is still a work in progress. We have established some pilot protocols which are being put under the test within the running driver projects. At the moment, secure test data transfers have been performed within the PSSS driver project[15]. The pilot protocol for secure data transfer from hospital to BioMedIT is currently as follows:

- The Hospital fetches the project encryption key from an identified Data Manager and verifies its validity with the DCC.
- The Hospital prepares the data and metadata according to the DCC registry. The routing of the data is implicitly defined by the project ID.
- The Hospital encrypts the data and sends it to a local BioMedIT node.
- The next step is the data routing by the BioMedIT nodes.
- The Data Manager gets notified of the delivery of the data to the partner and final BioMedIT node. This is done currently by email by the DCC.
- The Data Manager decrypts the data at the BioMedIT node.
- Transfer is recorded in DCC logs

Additionally, a pilot protocol for the transferred data package structure was created:

- The data package should be consistent across hospitals
- The naming convention is following the ISO 8601 datetime format and is:
  YYYYMMDDhhmmss
- The data is tarballed and zipped as a tar.gz

[15] https://www.sphn.ch/dam/jcr:fe6597e5-78ea-4dd2-bd09-83d94c530e88/2017DRI20_Egli_Lay_summaries_20180220.pdf

- The data is then encrypted
- The metadata file is created
- The metadata file and the encrypted data are then “tarballed” for transfer.

7 Rights and Obligations

7.1 SPHN Information Security Policy

Data protection requires the implementation of **organisational and technical measures**, including the creation of **rules to follow the legislation**. The rules are defined in the SPHN **Information Security Policy (ISP)** which can be found on the SPHN web site at:

https://sphn.ch/document/information-security-policy/

The first version was established in summer 2018. The policy is applicable to data re-use in **research projects** and applies to Project Leaders, Users and BioMedIT Node IT personnel. Note that it does *not* apply to hospital infrastructures! The ELSI “SPHN Ethical Framework for Responsible Data Processing” can be found on the same URL and complements the ISP.

For Project Leaders and Users, the following sections are the most important ones which we expect you to read in details and comply with: 4 Asset Management, 5.2 Users and Responsibilities, 8 Cryptography and 12 Awareness Training (the latter is established with this training). BioMedIT Node personnel needs to be aware of all sections of the policy.

Obligations of a Project Leader:

- Justify data access for projects and respective team members. Access for additional members can be added/removed any time later. In general, she has the responsibility of her team members and their behaviours.
- Data life cycle of all data transferred to the BioMedIT Node (incl. deletion of data)
- Report of all incidents and issues (incl. data breaches)
- Sign Data Transfer and Use Agreement where appropriate

7.2 Data Transfer and Use Agreement (DTUA)

The “Data Transfer and Use Agreement” is a contract between Data Provider (Controller) and Project Leader (Controller) which defines which data are transferred and used on a BioMedIT Node within a specific project. It is a legally binding contract that is signed by the Data Provider(s) (hospitals) and the Project Leader(s).

7.3 Acceptable Use Policy

Whereas a DTUA is signed by a Project Leader, each BioMedIT User needs to sign an Acceptable Use Policy where she confirms compliance with this a related policy.

7.4 Obligations of Users

How to *protect a Linux user account* on BioMedIT? (see ISP, Section 5.2)

- The account must be personnel and must not be shared with other Users.
- Set a good password with many letters (10+) and do not share the password with anybody else.
- Set up a 2 Factor Authentication application e.g., on your mobile phone.
- Confirm your account once/twice a year. This can also be done via your Project Leader.

Data access and sharing

Typically, the BioMedIT Node does *not* (easily) enable you to share data with a different project (technical restrictions). Indeed, you are *not* allowed to share data or parts of data with other people who are not part of the project. **You are not event even allowed to use the data for a different research project!** To re-use data from *existing* projects into a *new* project, you must have the appropriate approvals, and then the technical boundaries will be adjusted so that data can be aggregated without the need of copying between project spaces. If a person is part of several research projects, usually there will be different project spaces which do not permit data to be transferred from one space to another one.

In order to avoid that confidential data is copied without control, only **dedicated transfer tools** can be used: they **monitor who transferred what data** and allows to have an overview of which data is located in which BioMedIT node. Note that the list of dedicated transfer tools is made available by the BioMedIT Nodes.

Usage of personal computer

You might want to use your **personal computer or an institutional laptop computer** to log into a BioMedIT node. Please follow these rules:

- Install the latest security patches and supported software versions.
- Use malware protection (i.e., an antivirus program).
- In case your institute allows you to store confidential data on your personal computer, the disk must be encrypted.
- Confidential data must be encrypted in the following cases:
  - Data is stored on portable media (removable disk drive, tape, USB key, etc.).
  - Transferred to or from a BioMedIT Node.
- Explicitly avoid publishing confidential data in public repositories such as GitHub, dropbox, etc.

Finally, on a BioMedIT node an intrusion detection system (ID) should flag and block any suspicious activity, e.g., a machine talking to an IP address which is flagged as a known spamming address. The ID would report such activity to the node system administrators and potentially block access until the intervention has finished.

7.5 Non-compliance

Non-compliance with this the Information Security Policy can result in immediate removal of access rights to the BioMedIT Node. Additionally, there might be sanctions in your home institute: if a criminal act was detected, the respective person might be fired and/or Swiss penal law might be applied, too. Finally, any bad behaviour might bring BioMedIT in danger - in terms of reputation but also in terms of not being able to fulfil its role.

Finally, we have a moral obligation towards data subjects, and we must protect their privacy.

7.6 Security incidents: data breach

Things might go wrong and data might be breached, i.e., accessible to people who do not have the right to access the data. If that is the case, quick actions are required! Please provide as much information as possible to the breached BioMedIT Node such as

- **date and time of detection and description**
- **potential risk** of this data breach and expect impact for people
- **Who** is affected and what are the potential **consequences**?

The Project Leader is responsible for her team and must report any issues/breaches with her project’s data. She must contact the respective BioMedIT Node (see “contact information” at: https://sphn.ch/biomedit). Next, the respective BioMedIT Node will trigger the necessary steps listed in the Incident Response Plan (internal procedure at the BioMedIT Node – not discussed here).

Here is an example of email reporting a data breach. Such an email must be sent of a BioMedIT Node Manager:

```
Dear BioMedIT Node,

I am the project leader of <PROJECT NAME> where we detected a data breach 1 hour
ago.

Three patient records leaked out and were found on the public web site
people.myuniversity.edu on 15 Oct 2018. The data are not online anymore.
There is a high risk that the breach will lead to a disclosure of medical
information about patients.

Please let me know if you need additional information.

Best regards, Frau Mustermann
```

8 BioMedIT Node: an access example

Now that you have learned what you need to know about security within the SPHN projects and BioMedIT, you have completed the first step towards getting access to the BioMedIT node. What do you have to do next?

The first condition is that you are part of an approved SPHN project. Next, check with your Project Leader to make you part of the project team. You will then need to be registered as a team member and sign a usage agreement to adhere to the regulations. For more information and contacts, see sphn.ch/biomedit.

Note that we are currently working on online project registration forms for both Project Leaders and Users. Access requests will soon be possible for authorized users through the SPHN portal. Until then you can contact your responsible BioMedIT node. Current node status (July 2019):

- Leonhard Med is fully operational
- sciCORE BioMedIT node test space was set up in May 2019 and is being thoroughly tested. Full operational is expected very soon.
- Core-IT/Vital-IT: is currently in development phase – planned to be fully operational in autumn 2019

All three BioMedIT Nodes (commonly called the “BioMedIT network”) have a common way to provide access to users. There are two possibilities:

- Via a web browser: one can connect to the infrastructure through what we call a “remote desktop”. That is the recommended access method for users.
- Advanced users may also be able to connect via a terminal session (SSH). By “advanced users” we refer to users who are very familiar with SSH terminal sessions and are comfortable to work with command line tools.

You will soon be able to login via the SPHN portal, with specified authentication method(s). The portal can be accessed through the address: portal.dcc.sib.swiss. SWITCH edu-ID is being tested and is likely to prevail as the preferred method.

Following this step, you will be required to enter a second authentication factor, possibly using one of your other devices. Only after that is verified, you will be granted access to the portal, and through this to the projects you have been authorized for.

Finally, before a project can be added to the BioMedIT space, the following steps have to be fulfilled:

- There should be a formal appointment of a Data Manager representing the project (appointed by the Project Leader). The Data Manager is the responsible contact point for all transfers and data.
- A formal request to open a working space for the project on the responsible BioMedIT Node should be made by the Project Leader.
- A list of collaborators (researchers/Users) who should be allowed to work with project data in this space must be provided by the Project Leader. This list will be maintained by DCC and the Project Leader.

9 Conclusions

The most important thing that you need to keep in mind is that **you need to respect the privacy of others** (of patients) if you use their data. This data are often confidential (!) so you should be aware of access restrictions and use specifically dedicated IT infrastructures such as the BioMedIT nodes.

You might not remember everything we discussed today. Please go over the scriptum again. When using a BioMedIT Node, keep in mind that the system was designed with certain technical features and “restrictions” which will guide you to fulfill many of the requirements with respect to confidentiality and access restrictions to data. However, just because you are maybe able to do something, it does not mean you should! So if you are unsure, please contact the responsible BioMedIT Node for support!

Acknowledgements

This work is based on the input and participation of the following people:

Acknowledgements

UniversityHospital Zurich

- Karin Edler
- Francisca Jörger
- Cornelia Kruschel
- Michael Weisskopf

ETH Zurich – SIB

- Christian Bolliger
- Diana Coman Schmid
- André Kahles
- Simona Morello
- Gunnar Rätsch
- Bernd Rinn
- Thomas Wüst

University of Zurich

- Isabel Baur
- Julian Mausbach

University of Basel – SIB

- Sofia Georgakopoulou
- Thierry Sengstag

SPHN/DCC – SIB

- Leila Alexander
- Katrin Crameri
- Martin Fox
- Kevin Sayers
- Silvia Schaub
- Torsten Schwede

SIB Swiss Institute of Bioinformatics

- Séverine Duvaud
- Roberto Fabbretti
- Marc Filliettaz
- Vassilios Ioannidis
- Diana Marek
- Warren Paulus
- Heinz Stockinger

SIB Data Protection & IT Security Board

External Reviewer

And all people who participated in the SPHN Information Security Policy