ADOPT BBMRI-ERIC
Grant Agreement no. 676550

DELIVERABLE REPORT

Deliverable no	D3.2
Deliverable Title	Security and Privacy Architecture
Contractual delivery month	M12 (September 2016)
Responsible Partner	BBMRI-ERIC
Author(s)	Petr Holub and Common Service IT contributors
Actual delivery date	2016–09–30

Security and Privacy Architecture

Executive Summary

BBMRI-ERIC IT ecosystem deals with human material and data as the principal component; therefore the privacy by design paradigm is very important. Privacy protection comprises analysis of risks and design of countermeasures, such as appropriate use of privacy-enhancing technologies and security measures.

This document has been written as a Deliverable of the ADOPT BBMRI-ERIC project, describing risk analysis, architecture of BBMRI-ERIC IT services and how the security and privacy protection is built into those. It also reuses previous work on Security & Privacy Requirements published as a milestone of the BBMRI Competence Centre in the EGI-Engage project (EGI-Engage) project, which have been updated and included as an appendix to this deliverable.

ADOPT BBMRI-ERIC is funded by the European Union (EU)
Horizon 2020 under Grant number 676550.

Horizon 2020

Document Overview

ADOPT BBMRI-ERIC is funded by the European Union (EU)
Horizon 2020 under Grant number 676550.

Horizon 2020

ADOPT BBMRI-ERIC is funded by the European Union (EU)
Horizon 2020 under Grant number 676550.

Horizon 2020

Revision history of this document

Type of revision	*Author (Institution)*	*Date*
Initial version 0.9	Petr Holub (BBMRI-ERIC)	September 15, 2016
Review	Florian Kohlmayer (TUM-MED)	September 29, 2016
Updated version 1.0	Petr Holub (BBMRI-ERIC)	September 30, 2016

Glossary

AAI	Authentication and Authorization Infrastructure. 25, 28, 29, 37, 60
AARC	Authentication and Authorisation for Research and Collaboration. See `https://aarc-project.eu/` and GÉANT Association (GÉANT), 60
AC	(Data|Samples) Access Committee. 63
(de facto) anonymized data	Anonymous data is such data, that is is no longer identifiable. See appendix A.5 for definition and appendix A.5.1 for practical recommendations on anonymization procedures. 11, 12, 19, 22–24, 29–32, 34, 36, 38, 66, 67, 74–77, 79
BIMS	Biobank Information Management System. 20
CA	Certification Authority. 24, 32, 71
CO	Control Objective (ISO 27001). 11
coded data	Pseudonymous data is such data for which identifiers of persons have been replaced by a code (pseudonym) [1]. See DT-1b in appendix A.5. 12, 29, 34, 36, 66, 67, 75–77, 80
Common Service	A formal way of organizing full member countries of BBMRI-ERIC to provide services of common interest. 5
CS ELSI	Common Service ELSI. See Common Service and ELSI, 53, *see* ELSI
CS IT	Common Service IT. See Common Service, 16–18, 24, 32, 33, 42, 78
DAC	Discretionary Access Control. 3, 64
DDoS	Distributed Denial of Service. 23
DFD	Data Flow Diagram. [2], 2, 15, 20, 21, 26, 28, 36, 37, 49, 51
directly identifying data	Raw data with original direct identifiers of persons, to which none of the privacy-enhancing technologies has been applied. Complement to privacy-enhanced data when dealing with human data. See DT-1a in appendix A.5. 12, 74–77, 80
BBMRI-ERIC Directory	Information service by BBMRI-ERIC, providing highly aggregated data about the biobanks and their collections of biological material and data. During BBMRI Preparatory Phase also known as Catalogue. 11, 15, 19, 21, 22, 25, 28, 29, 37
DoS	Denial of Service. 23, 50
DS	Discovery Service. See Shibboleth, 54, 55, 60
DTA	Data Transfer Agreement. 2, 11, 12, 26, 29, 30, 36, 38, 53, 74–76, 79

eduID	Research and educational identity federations, represented by national federations such as eduID.se, eduID.hu, eduID.cz, etc. 54, 59
EGI	`http://www.egi.eu/`. 60
EGI-Engage	EGI-Engage project. `https://www.egi.eu/about/egi-engage/`, 9, 74, 80
ELSI	Ethical, Legal, and Social Issues. 5
EoP	Elevation of Privilege. 50
EU	European Union. 62
FIPS	Federal Information Processing Standard. 58, 59
GA4GH	Global Alliance for Genomics & Health. 2, 9, 11, 23, 31, 39, 42
GDPR	General Data Protection Regulation. 7, 9, 29, 62, 67
GEDE	Group of European Data Experts in RDA. See `https://rd-alliance.org/groups/gede-group-european-data-experts-rda`
GÉANT	GÉANT Association. `http://www.geant.net/`, 5, 8, 54, 60, 75
HTTP	Hypertext Transfer Protocol. 57
IaaS	Cloud service providing direct access to the virtualized infrastructure. See [3]. 80
ICD-10	International Classification of Diseases, 10th revision, provided by World Health Organization (WHO). See `http://www.who.int/classifications/icd/en/`. 15
IdP	Identity Provider. See Shibboleth, 54–57, 59–61, 64, 65
IoI	Item of Interest. 50, 51
ISMS	Information Security Management System. 72
LINDDUN	Linkability, Identifiability, Non-repudiation, Detectability, Disclosure of information, Content Unawareness, Policy and consent non-compliance. [4], 2, 9, 11, 15, 40, 41, 49–51, 63, 69
LoA	Level of Assurance. 2, 16, 23, 30, 31, 36, 38, 56–59, 71, 75
MAC	Mandatory Access Control. 3, 63, 64
MOSLER	Secure platform for processing sensitive data. See `https://bils.se/resources/mosler.html`. 15, 65, *see* TSD
MSC	Message Sequence Chart(s), standard schema for defining communication among components in distributed systems. Defined in ITU-T Z.120 [5]. See also [6] and `https://en.wikipedia.org/wiki/Message_sequence_chart`. 19, 25–27, 35
MTA	Material Transfer Agreement. 2, 11, 12, 26, 29, 30, 36, 38, 53, 74–76, 79
N/A	not applicable. 23, 24

ADOPT BBMRI-ERIC is funded by the European Union (EU)
Horizon 2020 under Grant number 676550.

Horizon 2020

National or Organizational Node
National Nodes are entities designated to represent member countries in BBMRI-ERIC. 10, 16

non-human data
Type of data that does not contain any trace of personal/human data and thus is not privacy sensitive. See DT-1b in appendix A.5. 22–24, 29–32, 38, 66, 75

ODbL
Open Data Commons Open Database License. `http://opendatacommons.org/licenses/odbl/`, 52

OpenID
standard decentralized protocol for authentication with substantial support in commercial environments. See `http://openid.net/`, 54, 59

OPM
Open Provenance Model. `http://openprovenance.org/`, 70

Perun
Virtual group management system with support for virtual identity consolidation [7]. 28, 29, 37, 60, 61

PET
Privacy-Enhancing Technologies. 3, 65, 67

PII
Personally Identifiable Information. 15, 72

privacy-enhanced data
Data on which some of the privacy-enhancing technologies has been applied, e.g., identifiers have been removed or replaced (coded data) or anonymized data. See appendix A.5 page 66 for more detailed discussion.. 5, 66

PROV-DM
PROV Data Model. `http://www.w3.org/TR/prov-dm/`, 70

pseudonymized data
Based on strict General Data Protection Regulation (GDPR) wording, pseudonymous data is such data for which if the key is not known, it can be considered anonymous. See DT-3 in appendix A.5. As discussed in appendix A.5, this definition substantially differs from previous definition, where pseudonymous data has been equivalent to coded data.. 29–32, 34, 36, 38, 66, 67, 75

RBAC
Role-Based Access Control. 3, 61, 63–65, 74

RDA
Research Data Alliance. See `https://rd-alliance.org/`

REMS
Resource Entitlement Management System. `http://www.csc.fi/rems` and [8], 63

S/MIME
Secure/Multipurpose Internet Mail Extensions is a standard for public key encryption and signing email data in MIME format, defined in RFCs 3369, 3370, 3850, and 3851.. 23

SAML V2.0
Security Assertion Markup Language, Version 2.0. See `https://www.oasis-open.org/committees/security/`, 54, 59

Shibboleth
Federated identity system [9, 10], `https://shibboleth.net/`. 5, 6, 8, 54, 61

SNOMED CT
Clinical health terminology by The International Health Terminology Standards Development Organisation (IHTSDO). See `http://www.ihtsdo.org/snomed-ct`. 15

SOP
Standard Operating Procedure. 16, *see*

SP	Service Provider. See Shibboleth, 54, 55, 59–61, 64
SSH	Cryptographic network protocol for operating network services securely over an unsecured network. See RFC 4251. 33
SSL	Secure Socket Layer. 71, 77
SSO	Single Sign On. 56
STORK	Secure idenTity acrOss boRders linked. `https://www.eid-stork.eu/`, 59
STORK 2.0	Secure idenTity acrOss boRders linked 2.0. `https://www.eid-stork2.eu/`, 59
STRIDE	Spoofing, Tampering, Repudiation, Information Disclosure, Denial of service, Elevation of privilege. [2], 2, 9, 11, 15, 40, 49–51, 63
TCS	Service to provide variety of trusted digital certificates to research and educational institutions. See `https://www.terena.org/activities/tcs/`. 24, 32
TLS	Transport Level Security. 22–24, 30, 32, 33, 38, 71
TSD	Secure platform for processing sensitive data. See `https://www.norstore.no/services/TSD` and for TSD 2.0 `https://www.usit.uio.no/prosjekter/tsd20/`. 15, 65
VOPaaS	VO Platform as a Service provided by GÉANT. GÉANT and [11, 12], 60
WAYF	Where Are You From service. See Shibboleth, 54, 55, 60
WHO	World Health Organization. 6

1. Introduction

BBMRI-ERIC IT ecosystem deals with human material and data as the principal component and therefore the privacy by design paradigm is very important. Privacy protection is comprised of analysis of risks and design of countermeasures, such as appropriate use of privacy-enhancing technologies and security measures.

This deliverable of the ADOPT BBMRI-ERIC project summarizes: architectures of main tools currently being implemented or anticipated to be implemented, the risk analyses and how the security & privacy protection is incorporated into these. Because of pan-European and the possibly global impact of BBMRI-ERIC, we are also exploring compliance to the recommendations of the Global Alliance for Genomics & Health (GA4GH), which focuses on rules for providing and sharing genomics and clinical data worldwide. The main part of this deliverable is organized as follows: Section 2 provides basic overview of overall IT architecture of BBMRI-ERIC and data management strategy. It discusses the basic types of data BBMRI-ERIC deals with, as well as their life cycle and sharing. The main part of the deliverable is section 3, which describes architecture of each system (following from use case), analyses data storage and data flows and discusses risks associated with these, using Spoofing, Tampering, Repudiation, Information Disclosure, Denial of service, Elevation of privilege (STRIDE) and Linkability, Identifiability, Non-repudiation, Detectability, Disclosure of information, Content Unawareness, Policy and consent non-compliance (LINDDUN) methodologies, discusses types of data processed, defines privacy and security measures and maps the result to the GA4GH Security Infrastructure requirements.

As security and privacy protection are one of the cornerstones of BBMRI-ERIC, this document naturally builds on previous developments in the BBMRI-ERIC IT community, namely on the Security & Privacy Requirements document delivered by the BBMRI Competence Centre of EGI-Engage (EGI-Engage Milestone document M6.2), reusing material into appendices A and B. These sections have been also updated in order to make the terminology compliant with the latest interpretations of the upcoming GDPR, namely with respect to different semantics of the word "pseudonymized data" and "pseudonymization". These sections can be understood as background information for the readers not familiar with some of the important privacy and security concepts as well as with previous developments in the BBMRI-related community.

It is important to understand that while the ADOPT BBMRI-ERIC Deliverable D3.2 is a static snapshot of the *Security & Privacy Architecture* at the time of contractual delivery, this document *will be continuously updated* after releasing the deliverable. This is common procedure for all security & privacy policies, as these must reflect latest developments of tools as well as latest advances in privacy protection and computer security.

2. IT Architecture and Data Management Strategy of BBMRI-ERIC

BBMRI-ERIC relies on a component-based software stack with well-defined components of reasonable size (preferably not excessively large), interconnected using well-defined and well-documented APIs. The component diagram is shown in figure 1 and relevant components (in production or under development) are described in further detail in section 3. Architecture of the system is fully distributed, following the distributed architecture of BBMRI-ERIC itself, where it is called "hub and spokes" with central level, level of National or Organizational Nodes, and individual biobanks level. This architecture is applied to all the aspects including the long-term data storage and curation, querying data, and migration of computations to data, etc. The architecture is, however, not only forwarding all the queries to the destination layers (from central BBMRI-ERIC via National or Organizational Nodes to biobanks) and retrieving results from there, but it must support temporary data caching for those services that prioritize performance. From this perspective, BBMRI-ERIC has no ambition to setup large central storage facilities, although some members or specific BBMRI-ERIC-related projects may opt for aggregation of data into highly secure storage systems.

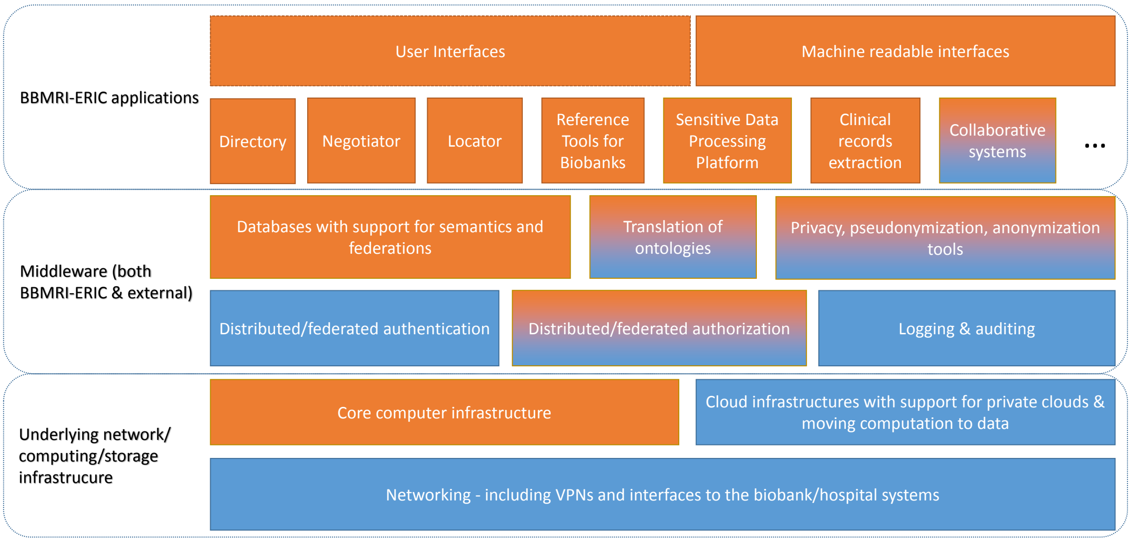
BBMRI-ERIC applications
User Interfaces
Machine readable interfaces
Directory
Negotiator
Locator
Reference Tools for Biobanks
Sensitive Data Processing Platform
Clinical records extraction
Collaborative systems
...
Middleware (both BBMRI-ERIC & external)
Databases with support for semantics and federations
Translation of ontologies
Privacy, pseudonymization, anonymization tools
Distributed/federated authentication
Distributed/federated authorization
Logging & auditing
Underlying network/ computing/storage infrastrucure
Core computer infrastructure
Cloud infrastructures with support for private clouds & moving computation to data
Networking - including VPNs and interfaces to the biobank/hospital systems

Figure 1: Software stack of BBMRI-ERIC IT system. Orange components are assumed to be built by BBMRI-ERIC, blue components are expected from other e-Infrastructures. Orange-blue components are assumed to be developed jointly with other e-Infrastructures.

From the data exchange perspective, BBMRI-ERIC is committed to FAIR principles[1] (Findable, Accessible, Interoperable, Reusable), extended by additional principles on quality and privacy protection.[2] This implies that access is only provided to authorized users, i.e., typically researchers who work on research projects that have been reviewed by a competent ethical review board.

[1] Data FAIRport, `http://datafairport.org/`

[2] This relates to a yet unpublished paper by BBMRI-ERIC contributors on extending the FAIR principles to FAIR QIP.

ADOPT BBMRI-ERIC is funded by the European Union (EU)
Horizon 2020 under Grant number 676550.
Horizon 2020

Furthermore, BBMRI-ERIC is committed to comply with Security Infrastructure guidelines provided by GA4GH.[3] The main risks identified by GA4GH are subset of the risks taken into account in this document using STRIDE and LINDDUN methodologies, with the mapping as shown in table 2. Compliance to individual subsections 4.x of Security Infrastructure will be discussed for each use case later in section 3.

GA4GH risk	*STRIDE*	*LINDDUN*
breach of confidentiality (CO-1)	information disclosure, elevation of privilege	
breach of individual privacy and autonomy (CO-2)	information disclosure	all
corruption/destruction of data (CO-3)	tampering, denial of service	
disruption of availability (CO-4)	denial of service	
adverse publicity due to unethical/illegal/ inappropriate actions (CO-5)	all	all

Table 2: Mapping of the risks identified in the GA4GH Security Infrastructure to STRIDE and LINDDUN risks. All risks for STRIDE are identified as: spoofing, tampering, repudiation, information disclosure, denial of service, elevation of privilege. All risks for LINDDUN are identified as: linkability, identifiability, content unawareness, policy/-consent non-compliance. Lastly, CO-5 is focused on GA4GH, but can be equally applied also to BBMRI-ERIC or any other medical research infrastructure.
Note that Section 4.2 in the Security Infrastructure document by GA4GH, which defines the CO-* labels, is to be replaced by Security & Privacy Policy once finalized by the GA4GH Regulatory and Ethics Working Group.

A typical workflow starts with an authenticated user searching for samples and/or data, or trying to identify biobanks to start collaboration with (see the BBMRI-ERIC Directory, Sample/Data Negotiator, and Sample/Data Locator components described in section 3). Before accessing samples and/or actual privacy-sensitive data (data that is personal and not anonymous – see requirement Req-4 on page 74 for definition and discussion of (de facto) anonymized data) – the user must submit a project which typically undergoes ethical evaluation, and only users with approved projects may be allowed any further. The users then request the samples and/or data and negotiates with biobankers. At this step, the user's request may still be rejected for several reasons: the samples or data may not be adequate for the intended purposes or the sample may be reserved for another project with higher priority or for another purpose (e.g., biobanks make certain samples reserved for quality management purposes including verification of previous experiments in case of dispute). Once the user's request is approved, the user signs Material Transfer Agreement (MTA) and/or Data Transfer Agreement (DTA) and the sample/data is transferred to the user.

[3] https://genomicsandhealth.org/category/search-topics/policy or https://genomicsandhealth.org/files/public/SecurityFramework-v1.1-2015-03-12-FINAL.pdf

ADOPT BBMRI-ERIC is funded by the European Union (EU)
Horizon 2020 under Grant number 676550.
Horizon 2020

When processing privacy-sensitive data, it is typically required that directly identifying data never leaves the biobank (or if the biobank is outside of the clinical facility, this data may not even reach the biobank). Depending on the type of the request, the biobank can transfer either (de facto) anonymized data or coded data with strong-enough MTA/DTA that prevents recipients from any re-identification attempts. Alternatively, the federated approach to the analysis can be used, which means that the processing of coded data or even directly identifying data takes place inside the biobank and only the aggregate anonymized data is sent out to the researcher; this has been previously described and demonstrated, e.g., using DataSHIELD[4] [13–15].

Because of size of the data and its nature, the paradigm of moving computations to data, can substantially improve the computational applications. This has been promoted in last 10 years and has become practically available with the advent of cloud technologies that can be deployed also within the perimeter of a biobank; use of private clouds for processing of biobank data has been developed and demonstrated by the BiobankCloud project.[5] An extended version of this scenario is envisioned by the Sensitive Data Processing Platform component in the software stack diagram.

Another specific aspect of the BBMRI-ERIC infrastructure is the heterogeneity of data that is collected in the biobanks and that needs to be mapped into consistent integrated data sets. Therefore BBMRI-ERIC works with federated databases with semantic data support (triple store systems) and translation of ontologies, which have been worked upon, e.g., in the BioMedBridges project.[6] Specific issues for the clinical biobanks arise from unstructured parts of clinical records that are on one hand one of the most valuable sources of information, but on the other hand require reliable extraction including natural language processing, which is still a research challenge.

2.1. Data Organization Description

The schema below tries to provide an overview of data storage locations. Please note there are two major types of biobanks that differ in how they store and access data in most cases: (a) population biobanks, which typically store all the relevant data inside the biobank together with the biosamples, (b) clinical biobanks, which rely on their connection to the clinical source of biosamples/data (hospital or other healthcare provider) and which typically need to query that source for more detailed data beyond the very basic data structure that is transferred initially together with the biosample.

(1) **Data stored inside a biobank.**

This is data that is stored within physical or at least logical perimeter of the biobank. Typically comprises several subtypes:

(1a) **Data generated inside a biobank.**

[4] http://www.p3g.org/biobank-toolkit/datashaper

[5] http://www.biobankcloud.com/

[6] http://www.biomedbridges.eu/

ADOPT BBMRI-ERIC is funded by the European Union (EU)
Horizon 2020 under Grant number 676550.

Horizon 2020

Typically operational data related to the biosamples, such as information about storage systems where the samples are located. In some cases, biobanks also perform further biosample analysis on their own, such as sequencing.

Example data: location of biosamples (in storage system).

(1b) **Data received together with the biosample and stored in a biobank.**

This is the data the comes into the biobank as a part of ingestion of the biosample into the biobank storage system. For clinical biobanks, it may consist of a subset of structured clinical data, while for population biobanks it may contain complete data set collected in the research/study about the donor.

Example data: (a) description of the sample (information on how and when the sample was taken and processed), (b) excerpt of structured patient's clinical data (pre-approved structure – typical for the clinical biobanks), (c) donor-related information related to the purpose of the research or biobank, such as life-style data, phenotype data, etc. (typical for the population biobanks).

(1c) **Data generated outside biobank and stored in a biobank.**

Example data: omics data generated by a user of a biobank, which is returned back to the biobank.

(2) **Data used by biobanks but stored outside the biobank.**

This category is typical for clinical biobanks detached from the hospital on a technical or administrative basis.[7] For any data access that is not part of the initial data transfer with the biosample (Item (1b)), the biobank needs to apply for the data from the hospital information system managers.

Example data: clinical records of patients.

(3) **Data stored at national level.**

Amounts and types of the data stored on this level varies largely based on the type of the national node. Typically consists of administrative/operational data of the national node itself and data linking to the biobanks. For some (typically smaller) national nodes, it may also store some data on behalf of the biobanks.

Example data: (a) Lists of interfaces to the biobanks, (b) authorization data for the services on the national level, (c) access/usage logs, (d) data query caches, (e) registry data on behalf of biobanks (if there is no on-line interface for the biobank), (f) terminology mappings.

(4) **Data stored at central BBMRI-ERIC level.**

This typically consists of administrative/operational data and data linking national nodes to the central BBMRI-ERIC level. BBMRI-ERIC intentionally avoids storing any privacy-sensitive data on the central level.

Example data: (a) Lists of interfaces to the national node services and service discovery, (b) terminology mappings, (c) authorization data for the services on the central BBMRI-ERIC level, (d) access/usage logs, (e) data query caches.

[7] This happens often that biobanks are considered research infrastructures and as a part of their institutionalization, they become detached from the clinical network in the hospital and from the hospital information systems, even though they may still reside in the same hospital premise.

ADOPT BBMRI-ERIC is funded by the European Union (EU)
Horizon 2020 under Grant number 676550.

Horizon 2020

(5) **Data stored outside of EU.**

This data may consist of any of the previously described data types (Items (1)–(4)), but regulations of other countries as well as European Union apply, if integrated into BBMRI-ERIC.

As one can see from the list above, BBMRI-ERIC features a fully federated and distributed architecture with distributed databases in autonomous organizations and organizational units (working under same umbrella of BBMRI-ERIC allowing for the federated operations) and distributed querying.

Data life cycle and traceability. An important aspect for traceability, are data modifications/updates, which are an inherent part of the data life cycle in the BBMRI-ERIC ecosystem. This aspect is particularly critical for the clinical biobanks, where the data coming from the clinical practice may come in largely varying quality and may require several rounds of refinement before they become usable for further research. The issue of data improvements and fixes should not be underestimated, however, even for other types of biobanks. The primary data can be only edited on the level where they are stored, see the Items (1)–(5). All the changes must result in a traceable and identifiable changes that can be used, e.g., in the provenance graphs [16, 17].

3. Architecture

This section describes the security architecture using the basic BBMRI-ERIC use cases [18], which are the core of IT development, as a part of Common Service IT of BBMRI-ERIC and supported by WP3 of ADOPT BBMRI-ERIC project:

- S+UCs-1: biobank browsing/lookup – implemented by BBMRI-ERIC Directory;
- S+UCs-2: negotiation of access to samples – implemented by Sample/Data Negotiator;
- S+UCs-{5,6}: lookup of samples – implemented by Sample/Data Locator;
- S+UCs-15: secure scalable data processing – to be implemented outside of the scope of ADOPT BBMRI-ERIC project.

The additional use case of secure scalable data processing, which is not subject to the ADOPT BBMRI-ERIC project, is only briefly mentioned in this document, as it is also part of the overall architecture. This is expected to be implemented by the trusted computing platforms such as MOSLER[8] and TSD[9] and possibly also by utilizing cloud service providers compliant with the standards generally accepted for processing of Personally Identifiable Information (PII) in medicine and medical research (appendix A.8).

Data Flow Diagrams (DFDs) are used to model use cases of BBMRI-ERIC [19], in order to evaluate them using STRIDE and LINDDUN (appendix A.1), as described in the previous section. This analysis results in the definition of requirements for implementation of those services.

Beyond the main components implementing the use cases discussed in this section, there is also an Ontology Translation Service. With the distributed nature of BBMRI-ERIC, the data comes in many different ontologies even in a single domain.[10] As data harmonization and ontology translation is an extremely important service for other tools (such as BBMRI-ERIC Directory, Sample/Data Negotiator and Locator), we define it as a separate component with well-defined interfaces to be incorporated into other applications. This service will be discussed as a part of each of the use cases where appropriate.

Measures to mitigate security risks proposed in this document are denoted using global numbering such as measure Me-1 on the following page. They are discussed first as general measures for the whole BBMRI-ERIC IT infrastructure and its operations, and then specifically for each modelled use case.

[8] https://bils.se/resources/mosler.html

[9] https://www.norstore.no/services/TSD

[10] A nice illustration is simple diagnosis coding, where not all the European countries use standard ICD-10 system and some use nationally customized variants of it of or customized variants of SNOMED CT.

ADOPT BBMRI-ERIC is funded by the European Union (EU)
Horizon 2020 under Grant number 676550.

Horizon 2020

3.1. Security Architecture of BBMRI-ERIC IT Infrastructure and its Operations

This section defines basic operational principles, that are common for all the systems and use cases. Also note that the systems operated by BBMRI-ERIC do not permanently store personal data (data type DT-1 on page 66 – including coded data), unless explicitly stated otherwise. This architecture and security measure applies to the whole BBMRI-ERIC IT infrastructure, including all the services operated by BBMRI-ERIC Common Service IT (CS IT) contributors.

Note that these principles are also to be applied on the level of biobanks and possibly other organizations operating their own infrastructure (e.g., National or Organizational Node) as minimum requirements, where measures can be hardened as appropriate; biobanks retain their own responsibility for operating their systems.

User Management for Operations. User managements for operation focus on the staff providing services, not on the users of these services, which is handled per use case below.

Me-1 User accounts are strictly individual and must not be shared.

Me-2 Account generation shall follow a well-documented standard operating procedure (SOP) and shall be documented.

Me-3 Both identity verification and authentication instances must be LoA ≥ 2.

Me-4 If only passwords are used for authentication (compared to using hardware tokens and/or multi-factor authentication), the passwords must be 12 characters long at minimum and must follow common guidelines [20, 21].

Me-5 Failed logins must be logged and system administrators must be notified about more than 3 consecutive login attempts. More than 3 consecutive logins shall result in time delay before additional login attempt is allowed.

Me-6 Inactivity logout or screen lock should be employed.

Me-7 User groups or roles are used for access control to resources and these groups/roles should be documented. Least privilege principle should be applied and privileges reviewed periodically to avoid collection of access over the time.

Physical Security.

Me-8 Server infrastructure must be physically accessible only to the designated IT personnel. This includes access to server rooms or their specific compartments. Physical access rights to the servers must be documented and for systems storing personal data (data type DT-1 on page 66 – including coded data) also individual accesses must be logged for minimum of 24 months (cf. measures Me-38-1 and Me-44-1 on page 31 and on page 38).

ADOPT BBMRI-ERIC is funded by the European Union (EU)
Horizon 2020 under Grant number 676550.

Horizon 2020

System Protection, System Separation, and Network Protection.

Me-9 Server systems should be clearly purposed, documented and separated from other servers at least on the level of virtual machines.

Me-10 Any system (server or client) connected to the network must be protected on its own, including applying automated security updates, network connection protection (local network traffic filtering), virus/malware detection software, and intrusion detection software.

Me-11 Systems should be hardened before putting them into production. This includes not running any excessive services and should not have unnecessary applications installed (i.e., applications not needed for operations and support). Vendor recommendations on hardening shall be applied as available and appropriate.

Me-12 Default passwords must not be used by any system connecting to the network or by systems they deliver network functionality.

Me-13 Networks must be protected by network traffic filtering with clearly documented (but not necessarily published) rules. Least privilege principles shall be applied when constructing firewall rules, i.e., only legitimate and documented services will be allowed and only the minimum necessary traffic will be enabled for them to operate.

Me-14 Application of network-wide virus/malware detection and intrusion detection is highly recommended.

Me-15 Only authorized systems may connect to the server segments of CS IT network infrastructure. Server segments must be clearly isolated from any networks that allow for connecting computers of common users (i.e., not CS IT staff on duty).

Software Development & Deployment.

Me-16 Any BBMRI-ERIC software must be tested by automated integration testing (including unit testing) before it is deployed into the production.

Me-17 BBMRI-ERIC Common Service IT mandates a clear hand over from the development to the operation. This includes transfer of knowledge (training, documentation) necessary for operating and supporting services.

Me-18 Any software installed must come from trusted installation sources (original media, signed software packages, etc.).

Me-19 When any security defect or vulnerability is found in any of BBMRI-ERIC software, it has to be corrected as soon as possible and the respective software release must clearly mark that the defect has been resolved. Internal documentation of the software development team must also document how the problem was resolved.

Me-20 BBMRI-ERIC CS IT development teams are responsible for monitoring software on which their systems are dependent. When the dependent software is packaged as a part of their distribution package for operational deployment, the complete package has to be updated as soon as possible.

ADOPT BBMRI-ERIC is funded by the European Union (EU)
Horizon 2020 under Grant number 676550.

Horizon 2020

Me-21 BBMRI-ERIC CS IT operations team is responsible for monitoring application of security updates to the deployed systems, including BBMRI-ERIC own software and third-party software.

Me-22 As additional measure to decrease chance of intrusion, the development teams are responsible for notifying operations teams about required security updates for both their own software packages as well as required third-party software.

Security Incident Handling.

Me-23 Any security incident must be properly investigated and documented. This documentation must involve identification of the source of the incident, consequences of the incident and corrective actions taken.

Me-24 BBMRI-ERIC IT and Data Protection Manager must be informed about any security incidents concerning BBMRI-ERIC IT infrastructure. Any affected third parties shall be notified, too.

Me-25 BBMRI-ERIC and its CS IT contributors are obliged to take over and handle any incidents reported by respective Computer Security Incident Response Teams (CSIRTs).

User Training.

Me-26 CS IT will organize a yearly webinar-based security and data protection oriented training for its staff.

Me-27 Data protection and privacy aspects shall be included in all the relevant training curricula produced by or supported by BBMRI-ERIC.

ADOPT BBMRI-ERIC is funded by the European Union (EU)
Horizon 2020 under Grant number 676550.

Horizon 2020

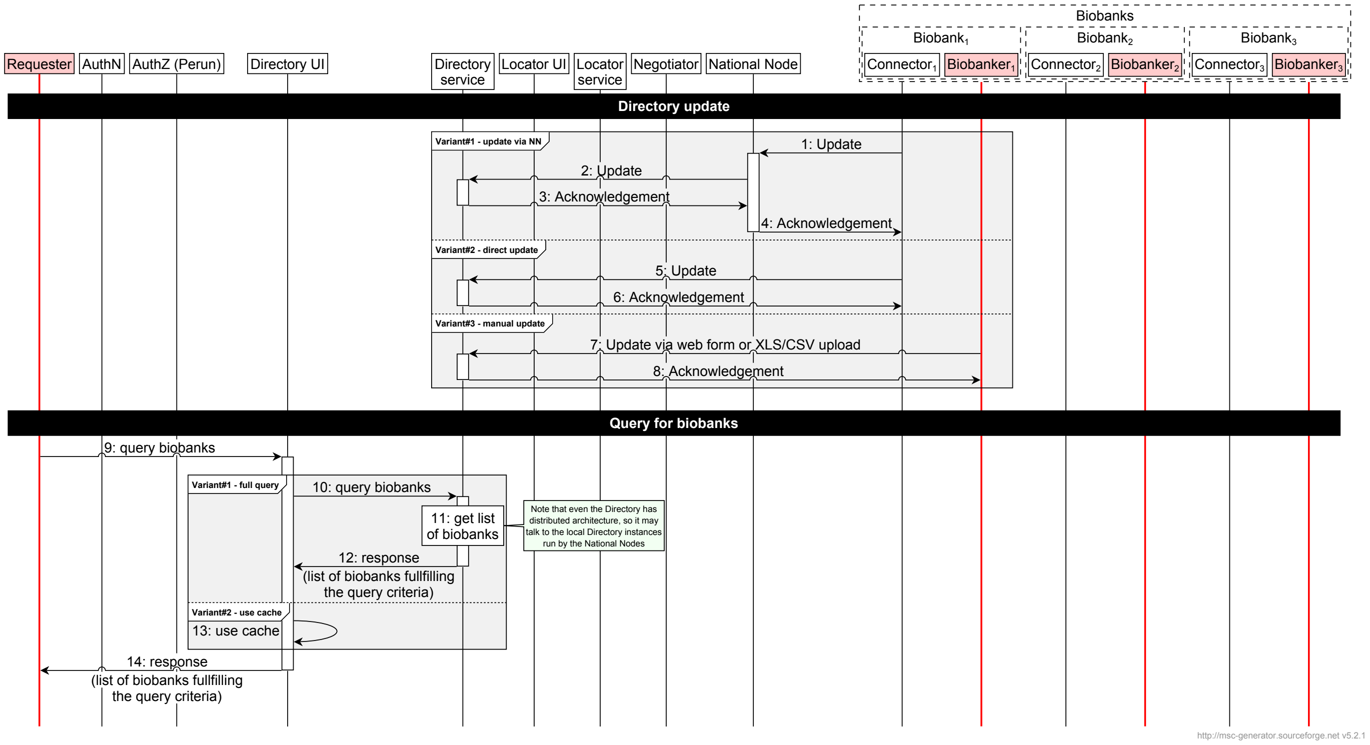

Figure 2: Detailed overview of interaction of components of BBMRI-ERIC Directory, modeled using MSC.

3.2. S+UCs-1: Biobank browsing/lookup

This use case deals with publishing highly aggregated information about biobanks, collection, biobank networks, other possible entities in the future (e.g., datasets without samples) and with various users accessing this information. In the future, it can be extended to publishing more detailed information, but only such that is considered (de facto) anonymized data (see appendix A.5.1 on page 69 and requirement Req-4 on page 74 for discussion). In practice, this use case is implemented by the BBMRI-ERIC Directory.[11]

BBMRI-ERIC Directory A distributed tool to provide highly aggregated information about biobanks, biobank networks, sample and data collections, and studies. This tool is primarily intended for the researchers to identify biobanks that may potentially have samples/data of their interest. The data is typically collected from the local biobanks via national nodes to the central level of BBMRI-ERIC, while national nodes utilize this structure to also run their national directories. This tool is used to assign identifiers to all the entities (biobanks, biobank networks, sample and data collections, studies), which can be further used not only for reproducibility and traceability, but also to assess their impact.[12] A detailed view of BBMRI-ERIC Directory modeled using Message Sequence Chart (MSC) is shown in figure 2.

[11] http://bbmri-eric.eu/bbmri-eric-directory
[12] See, e.g., BioResource Impact Factor (BRIF)[13] [22, 23].

3.2.1. DFD-Based Modeling

As shown in a DFD in figure 3, the system comprises three levels: (a) biobanks, (b) BBMRI-ERIC national nodes, and (c) BBMRI-ERIC central level. BBMRI-ERIC biobanks generate the metadata from their primary databases, usually a Biobank Information Management System (BIMS), and send it to the national node. The national node typically provides both a web interface presenting their national data and a machine readable interface (online query interface) to be used by internal and with some restrictions also external tools. The national nodes publish the data to the central level of BBMRI-ERIC, which again provides web interface as well as programmatic interface. Optionally the national nodes can also get information from the central level, so that their users may see similar results on the European level in addition to information from their national node.

Because data may come with different ontologies, the biobank metadata generator may also obtain data harmonization recipes from either BBMRI-ERIC ontology translation databases, or from external databases. This process does not involve sending the data out of the biobank, as only *recipes* (algorithms) are received and thus no privacy-sensitive data is transmitted. The same process may also occur on the national node level or central level, but it is omitted for the sake of simplicity from the diagram, since no privacy-sensitive data is involved.

BBMRI-ERIC infrastructure is also capable of dealing with non-BBMRI-ERIC biobanks or whole biobank networks, which are shown as “external biobank” in the figure 3. Information from these can be ingested either on the national level and republished into central BBMRI-ERIC level by the national node. Alternatively the external biobanks and biobank networks can be ingested directly into the central BBMRI-ERIC level; this mechanism is primarily intended for international biobank networks.

3.2.2. Data Types Employed

In this scenario, any data that gets out of the biobank (BBMRI-ERIC biobank or external biobank) is highly aggregated metadata (or anonymous data) about biobanks, their capabilities and their sample and data collections. The metadata typically includes:

- *biobank level:* information about the institutional aspect of the biobank, such as IDs of the biobank, juridical person (hosting and legally responsible institution), contact information, capabilities of the biobanks (what services it can offer, such as hosting various material types, processing data, etc.);

- *collection level:* type of the collection, amount of samples/data sets, types of the material stored, age ranges and sex of participants (patients/donors), available diagnoses, and collection-specific contact information. Collection-level data is expected to become more granular in the future (creating finer-grained sub-collections, e.g., reflecting standard operating procedures for retrieval, processing, and storing samples), resulting in

ADOPT BBMRI-ERIC is funded by the European Union (EU)
Horizon 2020 under Grant number 676550.
Horizon 2020

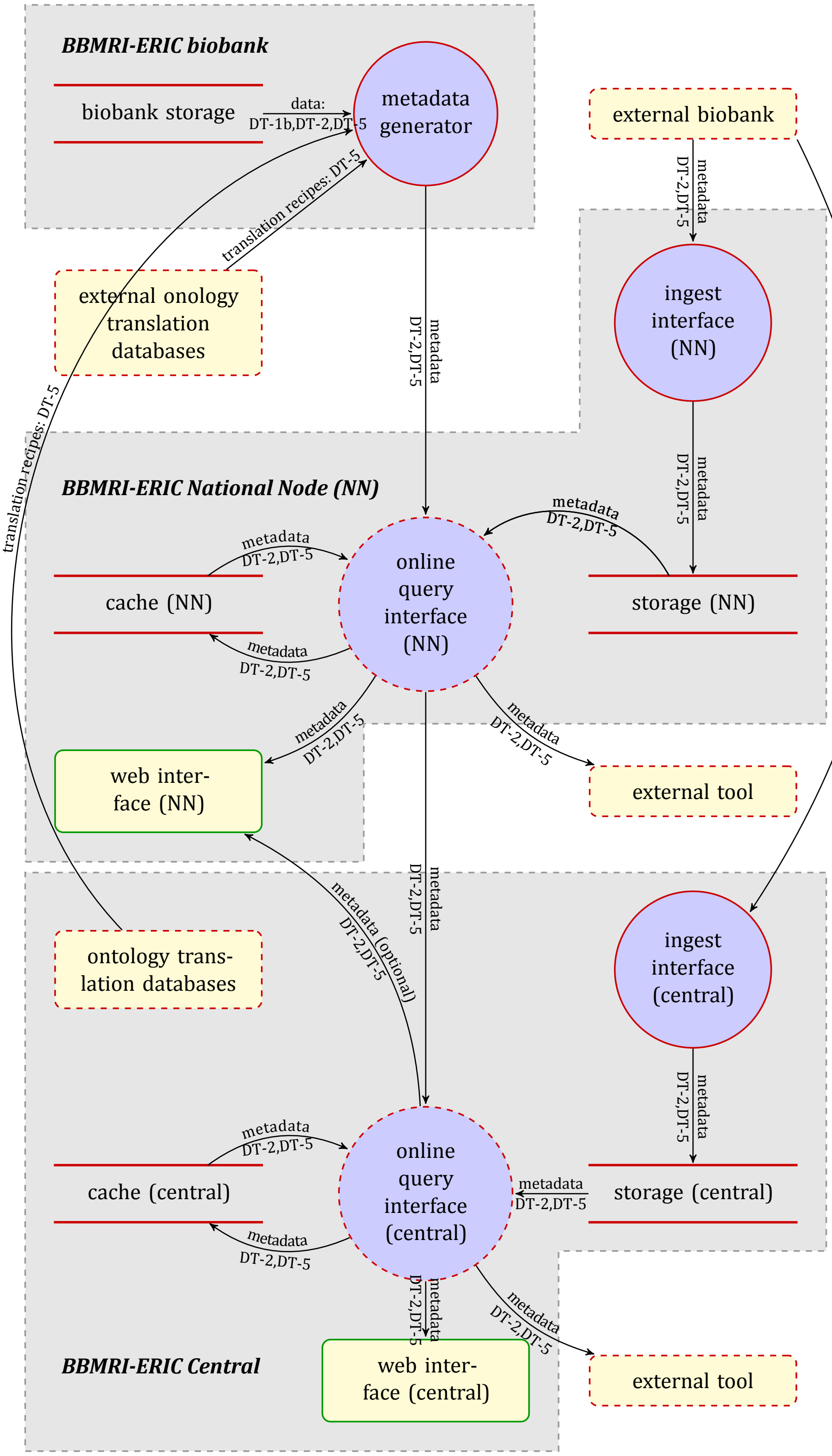

Figure 3: S+UCs-1: Biobank browsing/lookup using the BBMRI-ERIC Directory.
Note that data harmonization may also occur on the national node level or central level, but this is omitted from the diagram for simplicity reasons, as no privacy-sensitive data is transmitted during this process. Semantics of DFD is described in appendix A.1 on page 49, datatypes DT-*n* are used based on appendix A.5 on page 65.

Horizon 2020
ADOPT BBMRI-ERIC is funded by the European Union (EU) Horizon 2020 under Grant number 676550.

number of samples for each combination of parameters, while ensuring the data is still (de facto) anonymized data – see requirement Req-29 on page 79.

Overall, the data can be considered either non-human data (data type DT-5 on page 66) or (de facto) anonymized data (data type DT-2 on page 66) due to very high level of aggregation.

Note on contact information: For the purposes of this document, the contact information of a collection or a biobank or a biobank network (including phone number and email) is not considered personal information, i.e., it is data type DT-5. Such information is the official institutional contact and as such it does not fall under the protection of personal information. In many cases, such contact information also points into helpdesk or request tracking systems.

3.2.3. Security & Privacy Protection Measures

Me-28 Data protection

Me-28-1 Privacy-sensitive data (data type DT-1) stays in a biobank as only metadata leave the biobank. Metadata may include only highly aggregated (de facto) anonymized data DT-2) complying with requirement Req-4 and non-human data DT-5), thus complying with requirement Req-1 and minimum access control requirement Req-6.

Me-28-2 Biobanks are responsible for protecting against unauthorized access to their systems *including metadata generator service*, thus fulfilling requirement Req-1. Furthermore biobanks are obliged to comply with requirement on accountability and archiving described in appendix B.2 on page 75.

Me-29 Data anonymity

Me-29-1 biobanks are responsible for ensuring that the collection-level information is anonymous to the national/European standards, according to requirements Req-4 and Req-29,

Me-29-2 national nodes are responsible for verifying data anonymity status required by measure Me-29-1 if data flows to the central service via national node,

Me-29-3 central BBMRI-ERIC is responsible for verifying data anonymity status required by measure Me-29-1 in exceptional cases where data flows does not go via national node.

Me-30 Data integrity and authenticity

Me-30-1 when data is transmitted into the BBMRI-ERIC Directory service component using machine-to-machine communication, channels are encrypted using Transport Level Security (TLS) 1.1 or higher (for HTTPS/JSON and for LDAP – see appendix A.7) and the originating server is authenticated using a server certificate confirmed by the national node or a biobank using an independent channel (signed email or telephone),

ADOPT BBMRI-ERIC is funded by the European Union (EU)
Horizon 2020 under Grant number 676550.

Horizon 2020

Me-30-2 when data is entered manually using web-based user interface, channels are encrypted using TLS 1.1 or higher (for HTTPS) and the person is authenticated either using federated authentication or using local account,

Me-30-3 when data is sent by email: email must be signed by a S/MIME using a trusted certificate,

Me-30-4 access must be available via secure communication channel using TLS 1.1 or higher, although insecure channels may be provided in addition.

Me-31 In order to mitigate Denial of Service (DoS) attacks, per-client request throttling should be enabled for anonymous users. (Note that this does not prevent sophisticated large-scale Distributed Denial of Service (DDoS) attacks.)

Me-32 Data recovery/disaster plan

Me-32-1 all the primary sources of information is regularly backed up,

Me-32-2 data cached centrally on the BBMRI-ERIC Directory server is backed up on daily basis with minimum of 30 days backup availability.

Me-33 Logging and auditing is done based on policies of participating institutions, except for biobanks, which must also comply with requirement on accountability and archiving described in appendix B.2 on page 75 (see also measure Me-28-2).

3.2.4. Mapping to GA4GH Security Infrastructure

- 4.1 – Information Security Responsibilities:
  - Individuals = research participants,
  - Data Stewards = BBMRI-ERIC + national nodes + biobanks
  - Data Service Providers = BBMRI-ERIC (+ national nodes)
  - Application Service Providers = BBMRI-ERIC (+ national nodes + biobanks and their software vendors + third party software vendors)
  - Infrastructure Service Providers = BBMRI-ERIC (+ national nodes)
  - Service Consumers = researchers, biobankers, BBMRI-ERIC, national nodes, research participants, policy makers
- 4.3 – Identity Management:
  - N/A – this use case deals only with highly-aggregate (de facto) anonymized data (DT-2) and non-human data (DT-5),
  - optional authentication (LoA ≥ 0) may be provided for storing user preferences.
- 4.4 – Authorization Management:
  4.5.1 – Access Control:
  4.5.2 – Privacy Management:
  4.5.3 – Audit Log Recording and Review:

ADOPT BBMRI-ERIC is funded by the European Union (EU)
Horizon 2020 under Grant number 676550.
Horizon 2020

  - N/A – this use case deals only with highly-aggregate (de facto) anonymized data (DT-2) and non-human data (DT-5).

- 4.5.4 – Data Integrity:
  4.5.5 – Non-repudiation:

  - service-to-service authentication and communication channel encryption (TLS 1.1 or newer) for reception/aggregation of data (biobanks → national nodes → BBMRI-ERIC Directory),
  - service authentication and communication channel encryption (TLS 1.1 or newer) for retrieval of the data by users,
  - server certificates issued by one of the commonly accepted Certification Authorities (CAs) (e.g., server certificates provided via TERENA Trusted Certificate Service (TCS) will be sufficient for this purpose).

- 4.6 – Cryptographic Controls: 4.7 – Physical and Environmental Security:

  - N/A – this use case deals only with highly-aggregate (de facto) anonymized data (DT-2) and non-human data (DT-5).

- 4.7 – Physical and Environmental Security:
  4.8 – Operations Security:
  4.9 – Communications Security:

  - not required – this use case deals only with highly-aggregate (de facto) anonymized data (DT-2) and non-human data (DT-5),
  - it will be implemented to minimum extent (for cost reasons) to prevent tampering with the data.

- 4.10 – Service Supplier Assurances:

  - BBMRI-ERIC does not use external service suppliers (contributors to CS IT are considered part of BBMRI-ERIC and contractually bound to act as such).

ADOPT BBMRI-ERIC is funded by the European Union (EU)
Horizon 2020 under Grant number 676550.

Horizon 2020

3.3. S+UCs-2: Sample/Data Negotiator

This use case is about simplifying negotiation of access to samples and data between the sample/data custodians (biobankers and managers/operators of other bioresources) and requesters. A typical problem in this scenario, as it is implemented manually now, is that *(a)* the requesters often provide insufficiently specified requests that need to be refined with each biobank that might potentially have samples, *(b)* the requester needs to communicate with multiple (potentially tens or hundreds) of candidate biobanks at the same time. As a part of this process, biobankers also need to assess suitability of their samples/data for intended analytical methods. Such an approach creates tremendous overhead on both requester and participating biobanks, as it results in communication in the order of $N * M$ steps for each request, where N is the number of requesters and M is number of biobanks. With the Sample/Data Negotiator in place, it is sufficient if a single biobank helps to refine the request or if multiple biobanks refine different aspects of the request. Hence the communication complexity is lowered to approximately $N + M$. In the future the workflow may also support optional sample reservations and access to other services offered by the biobanks (such as sample/-data hosting).

For *requesting human samples or privacy-sensitive data*, this use case presumes the requester has a *project that has been approved by an ethical committee*. This is particularly important since as a part of the negotiation, the custodian (biobanker) needs to assess compliance of the project for that samples/data are requested with the informed consent for the candidate samples/data – see requirement Req-5 on page 75 and requirement Req-32 on page 79.

The *sample reservations* are intended for situations when a *project application* is only submitted for evaluation (incl. evaluation by ethical committee) and the user needs a time-limited guarantee, that if the project is accepted, they can have access to the samples necessary for conducting the research. From the data flow perspective, this follows the same two-step process as with the sample access (i.e., querying for the samples/data as the first step and access to the samples/data as second step), except that the actual sample access is replaced by time-limited sample reservation. Sample reservations can either expire after predefined time or can be deleted explicitly the project proposal is known to be rejected.

Sample/Data Negotiator is the web-based tool intended to implement this use case. Both requesters and biobankers interact using web-based forms, creating an environment similar to well known discussion forums with specific visibility properties: refinement communication within each request is visible to all the candidate biobanks (hence no need to ask and answer identical questions), while the offers of samples/data set from biobanks to the requester are treated as confidential. The requester provides structured and unstructured data and project description as a part of the request. The Sample/Data Negotiator interacts with the BBMRI-ERIC Directory to query candidate collections based on structured data in each request, and with the group management system in Authentication and Authorization Infrastructure (AAI) to retrieve contact information for each relevant collection (or biobank or biobank network, depending on communication preferences of the specific collection). The communication schema using MSC is shown in figures 4 and 5.

ADOPT BBMRI-ERIC is funded by the European Union (EU)
Horizon 2020 under Grant number 676550.
Horizon 2020

In the future it is expected that the Sample/Data Negotiator will also interface to Sample/Data Locator to achieve higher specificity when identifying candidate biobanks – this interface is yet to be specified.

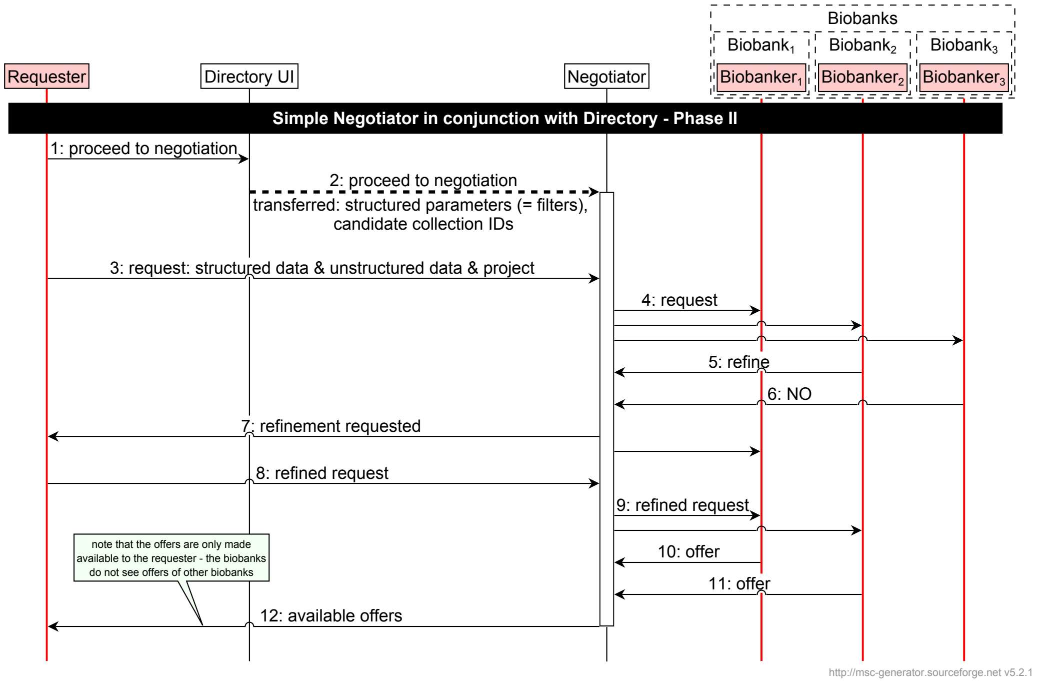
Biobanks
Biobank$_1$
Biobank$_2$
Biobank$_3$
Requester
Directory UI
Negotiator
Biobanker$_1$
Biobanker$_2$
Biobanker$_3$
Simple Negotiator in conjunction with Directory - Phase II
1: proceed to negotiation
2: proceed to negotiation
transferred: structured parameters (= filters), candidate collection IDs
3: request: structured data & unstructured data & project
4: request
5: refine
6: NO
7: refinement requested
8: refined request
9: refined request
note that the offers are only made available to the requester - the biobanks do not see offers of other biobanks
10: offer
11: offer
12: available offers
http://msc-generator.sourceforge.net v5.2.1

Figure 4: High-level overview of interaction between the Sample/Data Negotiator and its users modeled using MSC.

3.3.1. DFD-Based Modeling

As shown in figure 6 on page 28, the whole process starts with the requester communicating via the BBMRI-ERIC web interface with the request tracker process. The request is persistently stored in the request tracking database in the BBMRI-ERIC storage. The requests and their updates are then propagated to BBMRI-ERIC biobanks, which can either refine them (requesting further input from the users), or respond by contributing available samples/data sets.

As can be seen from the DFD, during the sample/data negotiation, no sample-level or individual-level data leaves the biobank. The restricted access to the services is in place for the following reasons: (a) to protect biobankers from communication with counterfeit identities, (b) to assert affiliation of users to the projects, and (c) to assert affiliation of persons to institutions that are juridical persons for the projects for liability reasons.

As a part of the sample/data release to the requester, the MTA and/or DTA must be signed – this process is not covered by the figure 3, as no relevant data flow is involved there. However, both MTA and DTA create a contractual binding for the requester, limiting how the samples and the data can be used.

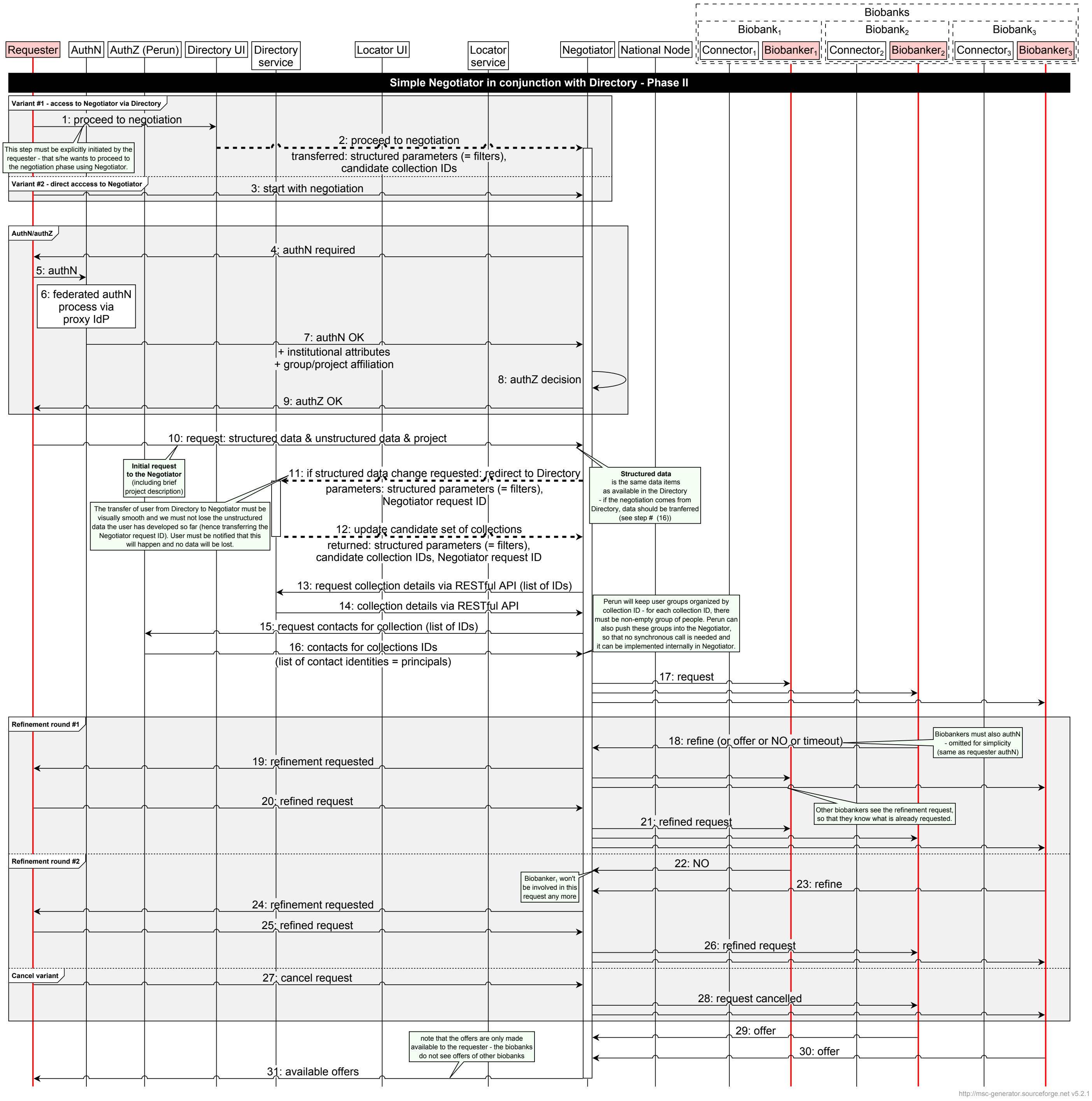
Biobanks
Biobank$_1$
Biobank$_2$
Biobank$_3$
Requester
AuthN
AuthZ (Perun)
Directory UI
Directory service
Locator UI
Locator service
Negotiator
National Node
Connector$_1$
Biobanker$_1$
Connector$_2$
Biobanker$_2$
Connector$_3$
Biobanker$_3$
Simple Negotiator in conjunction with Directory - Phase II
Variant #1 - access to Negotiator via Directory
1: proceed to negotiation
This step must be explicitly initiated by the requester - that s/he wants to proceed to the negotiation phase using Negotiator.
2: proceed to negotiation
transferred: structured parameters (= filters), candidate collection IDs
Variant #2 - direct acccess to Negotiator
3: start with negotiation
AuthN/authZ
4: authN required
5: authN
6: federated authN process via proxy IdP
7: authN OK
+ institutional attributes
+ group/project affiliation
8: authZ decision
9: authZ OK
10: request: structured data & unstructured data & project
Initial request to the Negotiator (including brief project description)
Structured data is the same data items as available in the Directory - if the negotiation comes from Directory, data should be tranferred (see step # (16))
11: if structured data change requested: redirect to Directory
parameters: structured parameters (= filters), Negotiator request ID
The transfer of user from Directory to Negotiator must be visually smooth and we must not lose the unstructured data the user has developed so far (hence transferring the Negotiator request ID). User must be notified that this will happen and no data will be lost.
12: update candidate set of collections
returned: structured parameters (= filters), candidate collection IDs, Negotiator request ID
13: request collection details via RESTful API (list of IDs)
14: collection details via RESTful API
15: request contacts for collection (list of IDs)
16: contacts for collections IDs
(list of contact identities = principals)
Perun will keep user groups organized by collection ID - for each collection ID, there must be non-empty group of people. Perun can also push these groups into the Negotiator, so that no synchronous call is needed and it can be implemented internally in Negotiator.
17: request
Refinement round #1
18: refine (or offer or NO or timeout)
Biobankers must also authN - omitted for simplicity (same as requester authN)
19: refinement requested
20: refined request
Other biobankers see the refinement request, so that they know what is already requested.
21: refined request
Refinement round #2
22: NO
Biobanker$_1$ won't be involved in this request any more
23: refine
24: refinement requested
25: refined request
26: refined request
Cancel variant
27: cancel request
28: request cancelled
29: offer
30: offer
note that the offers are only made available to the requester - the biobanks do not see offers of other biobanks
31: available offers
http://msc-generator.sourceforge.net v5.2.1

Figure 5: Detailed overview of interaction between the Sample/Data Negotiator, users, and various other components of the BBMRI-ERIC IT ecosystem, modeled using MSC.

ADOPT BBMRI-ERIC is funded by the European Union (EU)
Horizon 2020 under Grant number 676550.

Horizon 2020

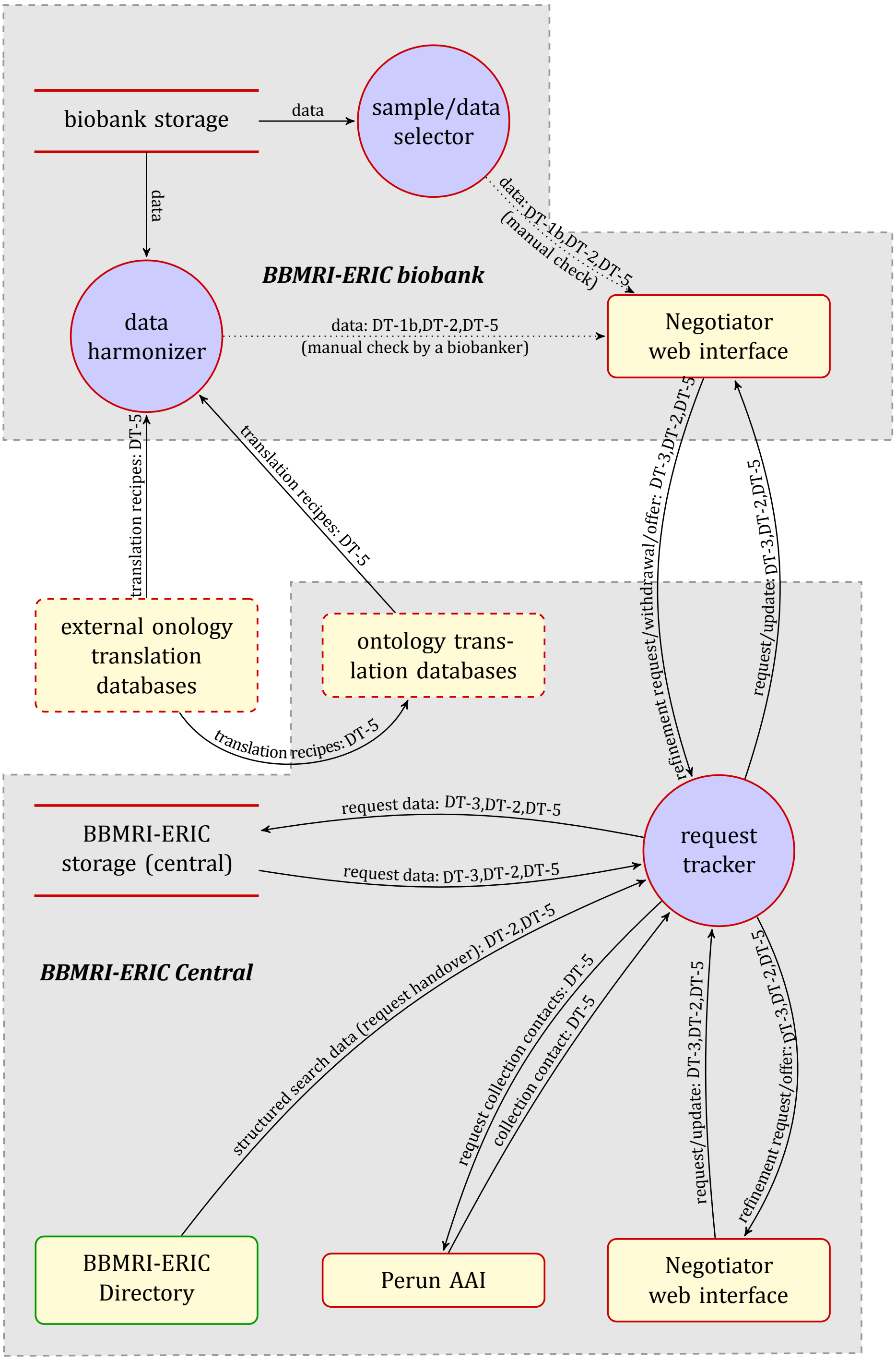
biobank storage
data
sample/data selector
data
data: DT-1b,DT-2,DT-5 (manual check)
BBMRI-ERIC biobank
data harmonizer
data: DT-1b,DT-2,DT-5 (manual check by a biobanker)
Negotiator web interface
translation recipes: DT-5
translation recipes: DT-5
refinement request/withdrawal/offer: DT-3,DT-2,DT-5
request/update: DT-3,DT-2,DT-5
external onology translation databases
ontology translation databases
translation recipes: DT-5
request data: DT-3,DT-2,DT-5
BBMRI-ERIC storage (central)
request data: DT-3,DT-2,DT-5
request tracker
BBMRI-ERIC Central
structured search data (request handover): DT-2,DT-5
request collection contacts: DT-5
collection contact: DT-5
request/update: DT-3,DT-2,DT-5
refinement request/offer: DT-3,DT-2,DT-5
BBMRI-ERIC Directory
Perun AAI
Negotiator web interface

Figure 6: S+UCs-{2,3}: Request refinement and access negotiation using the Sample/Data Negotiator.
Dotted lines denote manual check of data by a system operator (biobanker) in a disconnected system. Note the comment on contact information in section 3.2.2 on page 22. Semantics of DFD is described in appendix A.1 on page 49, datatypes DT-*n* are used based on appendix A.5 on page 65.

From the risk analysis perspective, an important aspect is that the requesters cannot browse/ search through informations about individual samples, which is functionality reserved for the biobankers. The sample/data selector module is detached/disconnected from the request processor, and even if there might be an online connection in the future as a part of interface to the Sample/Data Locator, the transfer of the data from the selector to the request processor is a manually controlled step, subject to approval by the biobanker (practically equivalent to committee-controlled access).

As a part of the use of the Sample/Data Negotiator, the biobankers get access to information that can be considered confidential: *projects* as a part of sample/data requests and even more importantly *project proposals* as a part of the sample reservations. This information needs to be treated as confidential, i.e., these will not be released beyond the biobank, nor will they be used by the biobank as their own novel research ideas.

3.3.2. Data Types Employed

This scenario involves the following data types:

- *Information about projects and project proposals:* which typically contains some level of intellectual property of the requester. Therefore as a part of terms & conditions of using Sample/Data Negotiator (and also as a part of general Acces Policy of BBMRI-ERIC), the contact persons of collections (which may be contact persons from the collection itself, the biobank hosting the collection or biobank network to which the collection belongs – depends on contact priority settings in BBMRI-ERIC Directory and group population in Perun AAI) must consent to treat this information confidential – thus complying with requirement Req-33

- *Structured (BBMRI-ERIC Directory search) request data:* contains query on subset of data found in the BBMRI-ERIC Directory, and therefore this part of the query can be considered highly aggregated (de facto) anonymized data (DT-2) or non-human data (DT-5)—see section 3.2.2 on page 20 for further discussion.

- *Unstructured request data:* contains additional requirements of the project on the samples and/or data, which cannot be expressed using structured query, as well as expected processing of the samples to assess their fitness for the given purpose. Unstructured data may evolve as a part of the refinement process and based on the communication with the collection contacts, some pseudonymized data (DT-3 in the GDPR sense, see 66) may appear as the mapping might be known to the collection contact (but not to the requester). Coded data or more sensitive data must not be used as a part of this communication, since MTA/DTA is not signed yet.

- *Contact information:* see comment on contact information in section 3.2.2 on page 22.

ADOPT BBMRI-ERIC is funded by the European Union (EU)
Horizon 2020 under Grant number 676550.

Horizon 2020

3.3.3. Security & Privacy Protection Measures

Me-34 Policy compliance

Me-34-1 Each first-time user must first agree to the terms & conditions of using the Sample/Data Negotiator service before it can proceed any further. The terms & conditions specify: *(a)* confidentiality of project information and project proposal information, *(b)* avoiding any re-identification efforts on any data obtained as a part of negotiation, *(c)* avoiding unethical behavior and complying with "The European Code of Conduct for Research Integrity" [24] and "GÉANT Data Protection Code of Conduct" [25].

Me-35 Data protection

Me-35-1 Privacy-sensitive data (data type DT-1) will not leave biobank. The data used in conversations in the Sample/Data Negotiator may include only non-human data (DT-5), highly aggregated (de facto) anonymized data (DT-2) and pseudonymized data (DT-3) complying with requirement Req-4, thus also complying with requirement Req-1 and minimum access control requirement Req-6.

Me-35-2 Contacts of collections must be authenticated using LoA ≥ 2 to ensure their entitlements to act on behalf of the collection.

Me-35-3 Contacts of collections are responsible for not sharing any personal data (DT-1) as a part of the negotiation process before the MTA/DTA is signed. This is mandated by the terms & conditions of using Sample/Data Negotiator together with authentication measure Me-35-2, hence complying with requirement Req-1 and requirement Req-6.

Me-35-4 Requesters must be authenticated LoA ≥ 1.[14]

Me-35-5 Biobanks are responsible for protecting against unauthorized access to their systems, thus fulfilling requirement Req-1. Furthermore biobanks are obliged to comply with the requirement on accountability and archiving described in appendix B.2 on page 75.

Me-36 Data anonymity

Me-36-1 Biobanks are responsible for ensuring that the collection-level information is anonymous to the national and European standards, in accordance with the requirements Req-4 and Req-29.

Me-36-2 Collections contacts are responsible for ensuring that the information provided as a part of negotiation is anonymous/pseudonymous – see measure Me-35-3.

Me-37 Data integrity and authenticity

Me-37-1 communication between requesters and collection contacts must be protected by a secure communication channel using TLS 1.1 or higher, and users must be authenticated as specified in measures Me-35-2 and Me-35-4.

Me-38 Logging and auditing

[14] Although it would be preferred to require LoA ≥ 2, it may become substantial barrier for access to the Sample/Data Negotiator. Dropping the minimum requirement to LoA ≥ 1 is acceptable because any privacy-sensitive data is only shared after signing MTA/DTA.

ADOPT BBMRI-ERIC is funded by the European Union (EU)
Horizon 2020 under Grant number 676550.
Horizon 2020

- Me-38-1 All accesses to the Sample/Data Negotiator will be logged and logs are stored for a minimum of 24 months.
- Me-38-2 Access logs to the service will be examined on weekly basis for suspicious behavior patterns.
- Me-38-3 Biobanks must comply with the requirement on accountability and archiving described in appendix B.2 on page 75 (see also measure Me-28-2).

Me-39 Data recovery/disaster plan

- Me-39-1 Sample/Data Negotiator database and access logs will be backed up daily in a non-proprietary backup format for minimum of 3 months backward availability,
- Me-39-2 Sample/Data Negotiator access logs will *not* contain details of the requests (structured nor unstructured data nor project details),
- Me-39-3 backups will be monthly tested for their readability.

3.3.4. Mapping to GA4GH Security Infrastructure

- 4.1 – Information Security Responsibilities:
    - Individuals = research participants,
    - Data Stewards = BBMRI-ERIC + biobanks (collections and their contacts)
    - Data Service Providers = BBMRI-ERIC
    - Application Service Providers = BBMRI-ERIC
    - Infrastructure Service Providers = BBMRI-ERIC
    - Service Consumers = researchers, biobankers, BBMRI-ERIC
- 4.3 – Identity Management:
    - authentication is required – this use case deals only with highly-aggregate (de facto) anonymized data (DT-2), pseudonymized data (DT-3), and non-human data (DT-5) (see measure Me-35-1),
    - authentication LoA $\geq$ 2 is required for collection contact persons (see measure Me-35-2),
    - authentication LoA $\geq$ 1 is required for requesters (see measure Me-35-4).
- 4.4 – Authorization Management:
    - all BBMRI-ERIC biobanks and collections are granted access to participate in the Sample/Data Negotiator,
    - each collection is assigned a primary contact person, which may in turn delegate this role to further persons via group management in authorization management system,
    - any researcher with LoA $\geq$ 1 is allowed to use the Sample/Data Negotiator (see measure Me-35-4) and this entitlement may be revoked based on breaching term & conditions of Sample/Data Negotiator service,

ADOPT BBMRI-ERIC is funded by the European Union (EU)
Horizon 2020 under Grant number 676550.

Horizon 2020

  - management of the Sample/Data Negotiator platform is assigned to BBMRI-ERIC CS IT operations group (WP7).

- 4.5.1 – Access Control:

  - collection contact may only see the requests for which the collection has become considered, candidate collection based on structured data information, or which was explicitly allowed by the requester,

  - for each allowed request, the collection contact is authorized to see progress of the request refinement and *not* authorized to see offers from other collections,

  - requester cannot see requests of other requesters,

  - access right of the requester to the Sample/Data Negotiator service *does not imply* access to the samples nor data sets (this is on the discretion of collection/biobank management,

  - IT management of the Sample/Data Negotiator platform may see any of the requests in arbitrary detail and is bound to treat details of these requests confidentially (also as a part of professional secrecy).

- 4.5.2 – Privacy Management:

  - confidentiality of the requests is enforced contractually (see measure Me-34-1) – otherwise this use case only deals with highly-aggregate (de facto) anonymized data (DT-2), pseudonymized data (DT-3), and non-human data (DT-5) (see measures Me-35-1 and Me-35-3),

- 4.5.3 – Audit Log Recording and Review:

  - access logs to the service will be examined on weekly basis for suspicious behavior patterns (see measure Me-38-2),
  - Sample/Data Negotiator runs on a dedicated virtual machine and the access logs to both the virtual machine and to the virtual machine monitor and examined on regular basis with minimum weekly frequency.

- 4.5.4 – Data Integrity:
  4.5.5 – Non-repudiation:
  4.6 – Cryptographic Controls:

  - user authentication and communication channel encryption (TLS 1.1 or newer) for any communication about requests in the Sample/Data Negotiator (see measure Me-37-1),
  - server certificates issued by one of the commonly accepted CAs (e.g., server certificates provided via TCS will be sufficient for this purpose).

ADOPT BBMRI-ERIC is funded by the European Union (EU)
Horizon 2020 under Grant number 676550.

Horizon 2020

- 4.7 – Physical and Environmental Security:
  4.8 – Operations Security:

  - Sample/Data Negotiator is run on a dedicated virtual machine in a physically protected facility operated by BBMRI-ERIC CS IT WP7 – CNR (subject to European, Italian, and Austrian law), operated *(a)* by a documented and trained team of system administrators, *(b)* with written operational procedures, *(c)* written availability commitments, *(d)* written commitments to ensure privacy and integrity of data, *(e)* written procedures for monitoring security including vulnerability of installed software and application of fixes.

  4.9 – Communications Security:

  - communication channels are encrypted (TLS 1.1 or newer) for any communication about requests in the Sample/Data Negotiator (see measure Me-37-1),
  - hosting machine is only accessible via Secure Shell (SSH).

- 4.10 – Service Supplier Assurances:

  - BBMRI-ERIC does not use external service suppliers (contributors to CS IT are considered part of BBMRI-ERIC and contractually bound to act as such).

ADOPT BBMRI-ERIC is funded by the European Union (EU)
Horizon 2020 under Grant number 676550.

Horizon 2020

3.4. S+UCs-{5,6}: Sample Locator

This use case deals with access of requesters to the sample-level data: search through individual samples stored in the biobanks and data sets related to individuals. The source data may be either (de facto) anonymized data or pseudonymized data or even coded data, depending on dimensionality of data (the higher the worse) and acceptable level of data quality loss (the lower the harder). The major difference to the previous use cases S+UCs-{1} and S+UCs-{2,3} is its automated access to statistics of the sample-level data or individual-level data, which may be highly multi-dimensional and thus problematic to achieve practical anonymity without very high data quality loss due to suppression/generalization/perturbation. Automated access to sample-level data is particularly sensitive from the privacy perspective, as it might be relatively easily abused for re-identification or unwanted information disclosure (e.g., using statistical inference). Therefore it must be the subject of high-security restricted access (appendix A.4.1) and acceptance of liability by the user (researcher, possible requester).

Sample/Data Locator If there were no privacy concerns (e.g., in case of non-human biosamples), the researchers could easily look up individual samples of their interest based on parametric search. For many biobank, retaining control about responses to the sample search query is of utmost importance and therefore the Sample/Data Locator implements a federated search paradigm as shown in figure 7. This means that once the search is initiated by the requester, the new request is created in the Locator and the connectors in the biobanks poll for the new requests on periodic basis (this is to ensure that all the communication out of the biobank is initiated by the components from inside of the biobank and no communication is allowed to be initiated from outside). Once new requests are received by a connector, it prepares the response based on the internal databases implemented inside the biobank; this means the connector can either access such warehouse, or it may store a copy of the privacy-enhanced data in its local database (cache). The response contains a number of samples that fulfill the given search criteria. Once the response is generated, the biobanker is notified and s/he may decide to approve/reject the response or to modify it. Once the response is approved, it is sent back to the Locator service. If the response is rejected by the biobanker, an empty response is sent back. Furthermore, if the biobanker does not react within predefined timeout (in order of several days), the Locator service triggers a timeout event. The resulting data is checked for anonymity – particularly small numbers of resulting samples ($k < 5$ in the initial proposal) are supressed.

More complicated approaches based on differential privacy [26–28] will be explored in the future, to minimize “hidden black matter” due to suppression, while also minimizing probability of re-identification and attribute disclosure (inference). Non-trivial amount of “hidden black matter” may occur with the initial approach to descibed above because of the high-dimensionality of data, which is relateively sparse in real world [29].

Despite the fact that only subset of samples and data is assumed to be available through this tool, it will still be part of the overall system because of its unique capability to support generation of novel research ideas.

ADOPT BBMRI-ERIC is funded by the European Union (EU)
Horizon 2020 under Grant number 676550.
Horizon 2020

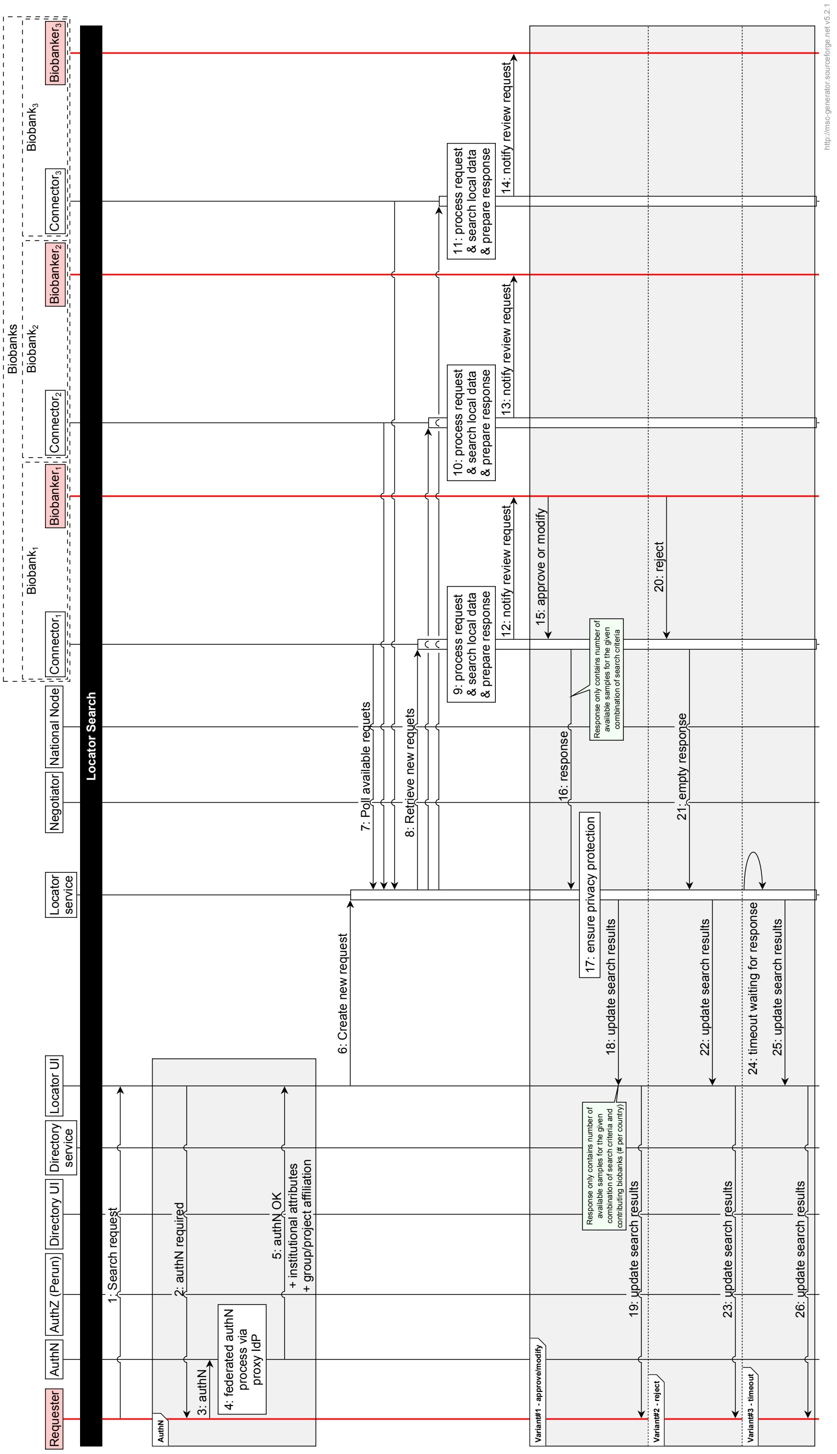
Requester
AuthN
AuthZ (Perun)
Directory UI
Directory service
Locator UI
Locator service
Negotiator
National Node
Biobanks
Biobank$_1$
Connector$_1$
Biobanker$_1$
Biobank$_2$
Connector$_2$
Biobanker$_2$
Biobank$_3$
Connector$_3$
Biobanker$_3$
Locator Search
1: Search request
2: authN required
3: authN
4: federated authN process via proxy IdP
5: authN OK + institutional attributes + group/project affiliation
AuthN
6: Create new request
7: Poll available requets
8: Retrieve new requets
9: process request & search local data & prepare response
10: process request & search local data & prepare response
11: process request & search local data & prepare response
12: notify review request
13: notify review request
14: notify review request
15: approve or modify
16: response
Response only contains number of available samples for the given combination of search criteria
17: ensure privacy protection
18: update search results
19: update search results
Response only contains number of available samples for the given combination of search criteria and contributing biobanks (# per country)
Variant#1 - approve/modify
20: reject
21: empty response
22: update search results
23: update search results
Variant#2 - reject
24: timeout waiting for response
25: update search results
26: update search results
Variant#3 - timeout
http://msc-generator.sourceforge.net v5.2.1

Figure 7: High-level overview of interaction between the Sample/Data Locator and its users (requester and biobankers) modeled using MSC. The Sample/Data Locator is implemented by three components: Locator UI, Locator Service, and Connectors. AuthN stands for "authentication" in the MSC.

ADOPT BBMRI-ERIC is funded by the European Union (EU)
Horizon 2020 under Grant number 676550.
Horizon 2020

3.4.1. DFD-Based Modeling

As shown in figure 8 on the next page, the whole process is initiated by a requester initiating a search request via the Sample/Data Locator web interface. Access to this component is already restricted and requires authentication with LoA ≥ 1 – this is sufficient as only (de facto) anonymized data or pseudonymized data (in GDPR sense) is returned, i.e., number of available samples and contributing biobanks. However, because the system allows for modifying requests and because issuing too many requests to the dataset can still pose a threat (namely compared to the highly aggregated anonymous data considered in section 3.2), LoA ≥ 1 authentication is still required even if the anonymous data could be retrieve without it.

Note hat privacy sensitive data, namely coded data (DT-1b) never leaves the biobank in this type of the search. If such data is released to the requester, it is not in this scenario, but it requires access negotiation (section 3.3) and signing MTA/DTA is required beforehand. Access to the biobank systems is restricted and is within the full responsibility of the biobank.

3.4.2. Data Types Employed

This scenario involves the following data types:

- *Structured (Sample/Data Locator search) request data:* combinations of search parameters, which could be theoretically used for inference of project ideas considered by requesters. Biobankers are obliged to adhere to ethical standards not to abuse knowledge about users' projects and project proposals.

- *Numbers of samples fulfilling given search criteria and a number of contributing biobanks per country:* this is (de facto) anonymized data (DT-2) or pseudonymized data (DT-3) (if the mapping of the original data to (de facto) anonymized data is stored at Locator service). Low numbers of samples will be suppressed initially ($k < 5$) and alternative differential privacy approach will be explored to reduce the suppression rates while also keeping risk of re-identification and attribute disclosure (inference).

3.4.3. Security & Privacy Protection Measures

Me-40 Policy compliance

Me-40-1 Each first-time user must first agree to the terms & conditions of using the Sample/Data Locator service before it can proceed any further. The term & conditions specify: *(a)* confidentiality of project information and project proposal information, *(b)* avoiding any re-identification efforts on any data obtained as a part of negotiation, *(c)* avoiding unethical behavior and complying with “The European Code of Conduct for Research Integrity” [24] and “GÉANT Data Protection Code of Conduct” [25].

Me-41 Data protection

ADOPT BBMRI-ERIC is funded by the European Union (EU)
Horizon 2020 under Grant number 676550.

Horizon 2020

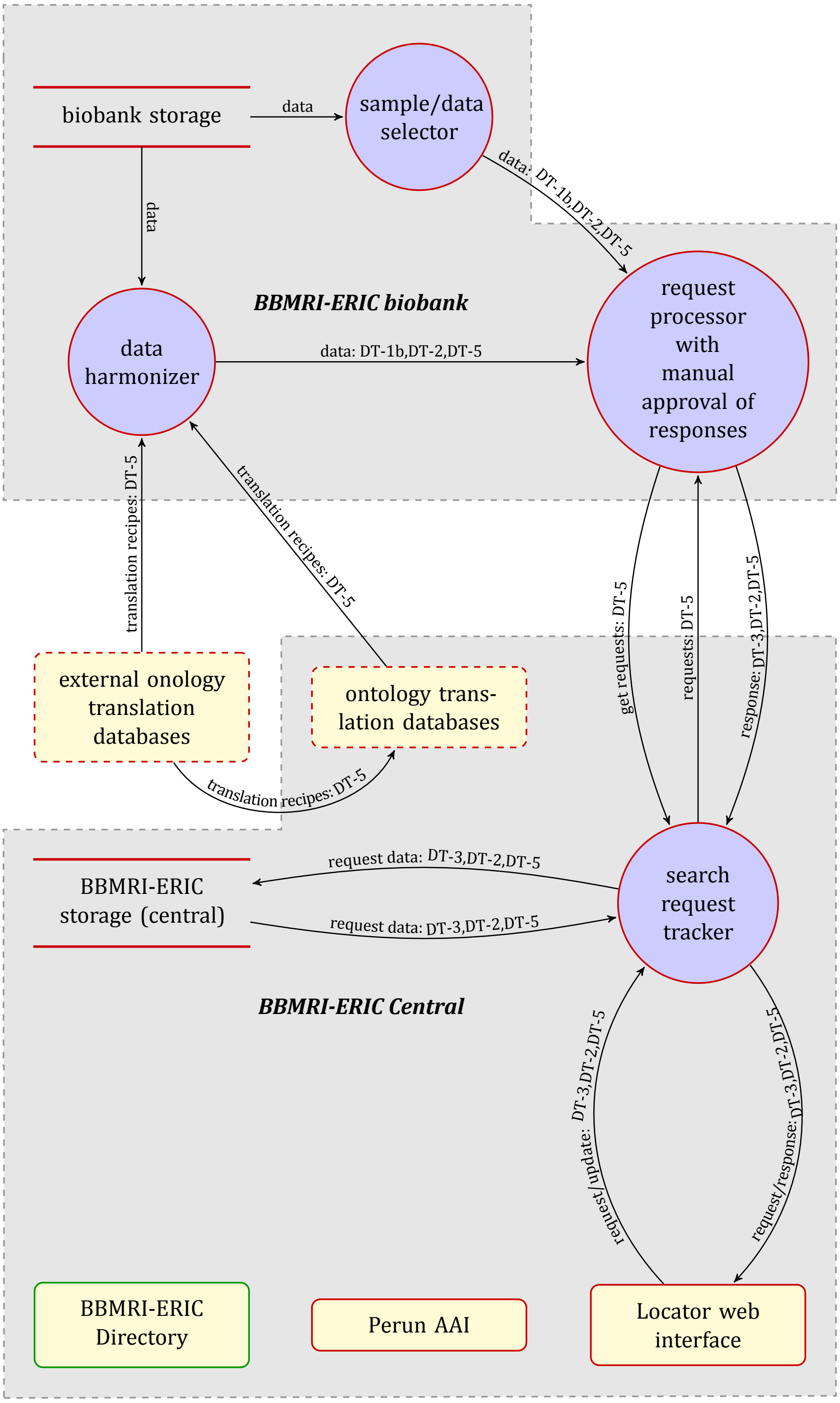
biobank storage
data
sample/data selector
data: DT-1b,DT-2,DT-5
data
BBMRI-ERIC biobank
data harmonizer
data: DT-1b,DT-2,DT-5
request processor with manual approval of responses
translation recipes: DT-5
translation recipes: DT-5
get requests: DT-5
requests: DT-5
response: DT-3,DT-2,DT-5
external onology translation databases
ontology translation databases
translation recipes: DT-5
request data: DT-3,DT-2,DT-5
BBMRI-ERIC storage (central)
request data: DT-3,DT-2,DT-5
search request tracker
BBMRI-ERIC Central
request/update: DT-3,DT-2,DT-5
request/response: DT-3,DT-2,DT-5
BBMRI-ERIC Directory
Perun AAI
Locator web interface

Figure 8: S+UCs-{5,6}: Sample Locator.
Semantics of DFD is described in appendix A.1 on page 49, datatypes DT-*n* are used based on appendix A.5 on page 65.

ADOPT BBMRI-ERIC is funded by the European Union (EU) Horizon 2020 under Grant number 676550.
Horizon 2020

Me-41-1 Privacy-sensitive data (data type DT-1) will not leave biobank. The data transmitted outside of the biobank in the Sample/Data Locator search includes only the number of available samples ((de facto) anonymized data (DT-2 or pseudonymized data (DT-3) and possibly non-human data (DT-5). Therefore it complies with requirement Req-4, thus complying with requirement Req-1 and minimum access control requirement Req-6.

Me-41-2 Biobankers must be authenticated using LoA ≥ 2 to ensure their entitlements to act on behalf of the biobanks/collection.

Me-41-3 Requesters must be authenticated LoA ≥ 1.[15]

Me-41-4 Biobanks are responsible for protecting against unauthorized access to their systems, thus fulfilling requirement Req-1. Furthermore biobanks are obliged to comply with requirement on accountability and archiving described in appendix B.2 on page 75.

Me-42 Data anonymity

Me-42-1 Locator service is responsible for anonymizing the response data at least to the requirements defined in requirements Req-4 and Req-29, i.e., $k < 5$ data will be suppressed.

Differential privacy approach will be designed to provide even better protection while also optimizing for as low suppression rates as possible.

Me-43 Data integrity and authenticity

Me-43-1 communication between requesters and biobankers must be protected by a secure communication channel using TLS 1.1 or higher, and users must be authenticated as specified in measures Me-41-2 and Me-41-3.

Me-44 Logging and auditing

Me-44-1 All accesses to the Sample/Data Locator will be logged and logs stored for minimum of 24 months.

Me-44-2 Access logs to the service will be examined on weekly basis for suspicious behavior patterns.

Me-44-3 Biobanks must comply with requirement on accountability and archiving described in appendix B.2 on page 75 (see also measure Me-28-2).

Me-45 Data recovery/disaster plan

Me-45-1 Sample/Data Locator database and access logs will be backed up daily in a non-proprietary backup format for minimum of 3 months backward availability,

Me-45-2 Sample/Data Locator database will be encrypted before backups using state-of-the-art encryption and the key will be securely stored separately from backups,

Me-45-3 Sample/Data Locator access logs will *not* contain details of the search requests and responses,

[15] Although it would be preferred to require LoA ≥ 2, it may become substantial barrier for access to the Sample/-Data Locator. Dropping the minimum requirement to LoA ≥ 1 is acceptable because any privacy-sensitive data is only shared after signing MTA/DTA.

ADOPT BBMRI-ERIC is funded by the European Union (EU)
Horizon 2020 under Grant number 676550.
Horizon 2020

Me-45-4 backups will be monthly tested for their readability.

3.4.4. Mapping to GA4GH Security Infrastructure

Compliance to the GA4GH Security Infrastructure will be evaluated before the first complete implementation of the Sample/Data Locator. This is due to ongoing minor adjustments that may occur as a part of the development process.

3.5. STRIDE/LINDDUN-Based Risk Analysis of BBMRI-ERIC Use Cases

Table 3: Risk assessment for threats (STRIDE and LINDDUN) to the "Data Flow" element of the DFD.

"Data Flow" threat	*Example*	*Risk*				*Countermeasure*
		S+UCs-1	S+UCs-{2,3}	S+UCs-{5,6}	S+UCs-14	
Tampering	Malicious modification of data or code, e.g., by man-in-the middle attack possible because of weak message or channel integrity checks	++	+++	+++	+++	Secure data communication
Information disclosure	Exposure of data to unauthorized persons, e.g. by man-in-the-middle because of lack of confidentiality for the channel	–	++	+++	+++	
Denial of service	Consumption of large quantities of fundamental resources due to weak message or channel integrity	++	++	++	++	
– (not relevant), + (low), ++ (medium), +++ (high)						

Table 4: Risk assessment for security (STRIDE) threats to the "Data Store", "Process", and "Entity" elements of the DFD associated to the use cases.

Security threat	*Example*	*Risk*				*Countermeasure*
		S+UCs-1	S+UCs-{2,3}	S+UCs-{5,6}	S+UCs-14	
Spoofing	Pose as something or somebody else	–	++	+++	+++	Authentication system, configuration management
Tampering	Malicious modification of data or code	–/+	++	+++	+++	Authorization system
Repudiation	Denial of having received data	–	+++	+++	+++	Auditing and logging
Information disclosure	Exposure of information to unauthorized individuals	–	++	+++	+++	Authorization System, Input Validation
Denial of service	Resources are not available due to overload or attack	++	++	++	+	Configuration management, input validation
– (not relevant), + (low), ++ (medium), +++ (high)						

Continued on next page...

ADOPT BBMRI-ERIC is funded by the European Union (EU)
Horizon 2020 under Grant number 676550.

Horizon 2020

... continued from previous page.

Security threat	*Example*	*Risk*				*Countermeasure*
		S+UCs-1	S+UCs-{2,3}	S+UCs-{5,6}	S+UCs-14	
Elevation of privilege	A user gains unauthorized access to resources	−/+	+++	+++	+++	Authorization system
– (not relevant), + (low), ++ (medium), +++ (high)						

Table 5: Risk assessment for privacy (LINDDUN) threats to the “Data Store”, “Process”, and “Entity” elements of the DFD associated to the use cases.

Privacy threat	*Example*	*Risk*				*Countermeasure*
		S+UCs-1	S+UCs-{2,3}	S+UCs-{5,6}	S+UCs-14	
Linkability	Possibility to detect that different data items are related to the same entity	−/+	+++	+++	+++	Anonymization tool, pseudonymization modules, encryption, access control system.
Identifiability	Possibility to relate a set of data to a specific entity / person; to recognize a person by characteristics	−/+	+++	+++	+++	
Content unawareness	A patient is unaware of the information used/shared by the system	–	+++	+++	+++	Informed consent management
Policy/consent non-compliance	Lack of evidence that data shared by the system meets applicable legal, policy or consent requirements	–	+++	+++	+++	Legal regulations, informed consent mgmt., data provider forms, ethics committee approval, data access comm. approval, DTA/MTA.
– (not relevant), + (low), ++ (medium), +++ (high)						

Note that for S+UCs-1, there is sometimes two values present in the tables above: −/+. This is because S+UCs-1 covers both data that is not considered personal at all (highly aggregate data and operational data of biobanks), for which there is no significant risk, but it may go also for the practically anonymous data, which introduces some low risk related to linking and re-identification.

ADOPT BBMRI-ERIC is funded by the European Union (EU)
Horizon 2020 under Grant number 676550.

Horizon 2020

3.6. Organization Compliance of BBMRI-ERIC to GA4GH Security Infrastructure

- 4.11 – Information Security Aspects of Business Continuity Management:

    - BBMRI-ERIC will respond to the potential security incidents as quickly as possible, typically within 24 hours on business days,
    - BBMRI-ERIC will investigate and resolve security incidents and reported threats as quickly as possible,
    - BBMRI-ERIC will comply with the legal requirements and regulations on reporitng breaches, with jurisdiction typically being Austria (for services located at BBMRI-ERIC headquarters) or Italy (for services hosted by CS IT).

- 4.12 – Compliance:

    - BBMRI-ERIC is committed to comply with the specified controls.

ADOPT BBMRI-ERIC is funded by the European Union (EU)
Horizon 2020 under Grant number 676550.

Horizon 2020

References

1. Health informatics – Pseudonymization. ISO/TS 25237:2008. 2008.
2. Howard, M and Lipner, S. The security development lifecycle: SDL-A process for developing demonstrably more secure software. 2006.
3. Peng, J, Zhang, X, Lei, Z, et al. Comparison of several cloud computing platforms. In: *2009 Second International Symposium on Information Science and Engineering*. IEEE. 2009:23–27.
4. Deng, M, Wuyts, K, Scandariato, R, et al. A privacy threat analysis framework: supporting the elicitation and fulfillment of privacy requirements. Requirements Engineering 2011;16:3–32.
5. ITU Telecommunication Standardization Sector (ITU-T). Z.120: Message Sequence Chart (MSC). Recommendation Z.120. ITU-T, 2011. URL: https://www.itu.int/rec/T-REC-Z.120/en.
6. Jonsson, B and Padilla, G. An execution semantics for MSC-2000. In: *International SDL Forum*. Springer. 2001:365–378.
7. Procházka, M, Licehammer, S, and Matyska, L. Perun—Modern approach for user and service management. In: *IST–Africa Conference Proceedings, 2014*. IEEE. 2014:1–11.
8. Linden, M, Nyrönen, T, and Lappalainen, I. Resource Entitlement Management System. In: *TERENA Networking Conference 2013 (TNC2013)*. 2013. URL: http://tnc2013.terena.org/getfile/870.
9. Cantor, S and SCAVO, T. Shibboleth architecture. Protocols and Profiles 2005;10:16.
10. Barton, T, Basney, J, Freeman, T, et al. Identity federation and attribute-based authorization through the globus toolkit, shibboleth, gridshib, and myproxy. In: *5th Annual PKI R&D Workshop*. Vol. 4. 2006.
11. Dijk, N van. Virtual Organization - as a Service. In: *AARC kickoff*. 2015. URL: https://aarc-project.eu/wp-content/uploads/2015/05/20150603-AARC_KICKOFF-Virtual-Organization-as-a-Service.pdf.
12. Dijk, N van. VOPaaS: Virtual Organisation Platform as a Service. In: *Internet2 2015 Technology Exchange*. 2015. URL: https://aarc-project.eu/wp-content/uploads/2015/05/20150603-AARC_KICKOFF-Virtual-Organization-as-a-Service.pdf.
13. Wolfson, M, Wallace, SE, Masca, N, et al. DataSHIELD: resolving a conflict in contemporary bioscience—performing a pooled analysis of individual-level data without sharing the data. International journal of epidemiology 2010:dyq111.
14. Jones, E, Sheehan, N, Masca, N, et al. DataSHIELD–shared individual-level analysis without sharing the data: a biostatistical perspective. Norsk epidemiologi 2012;21.
15. Gaye, A, Marcon, Y, Isaeva, J, et al. DataSHIELD: taking the analysis to the data, not the data to the analysis. International journal of epidemiology 2014;43:1929–1944.
16. Cuccuru, G, Leo, S, Lianas, L, et al. An automated infrastructure to support high-throughput bioinformatics. In: *High Performance Computing & Simulation (HPCS), 2014 International Conference on*. IEEE. 2014:600–607.

ADOPT BBMRI-ERIC is funded by the European Union (EU)
Horizon 2020 under Grant number 676550.

Horizon 2020

17. Zanetti, G. Data intensive biology and data provenance graphs. BiobankCloud Project. 2015. URL: http://www.biobankcloud.com/sites/default/files/ngshadoop/hadoop_ngs_zanetti.pdf.

18. Holub, P, Quinlan, P, DiezFraile, A, et al. BBMRI-ERIC Use Cases. 2016. DOI: 10.5281/zenodo.159302. URL: https://doi.org/10.5281/zenodo.159302.

19. DiezFraile, A, Holub, P, Hummel, M, et al. BBMRI-ERIC Use Cases. 2015.

20. SANS. Password Construction Guidelines. 2014. URL: https://www.sans.org/security-resources/policies/general/pdf/password-construction-guidelines.

21. US-CERT. Security Tip (ST04-002): Choosing and Protecting Passwords. Original release date: May 21, 2009. Last revised: October 01, 2016. URL: http://www.us-cert.gov/ncas/tips/st04-002.

22. Cambon-Thomsen, A, Thorisson, GA, Mabile, L, et al. The role of a bioresource research impact factor as an incentive to share human bioresources. Nature Genetics 2011;43:503–504.

23. Mabile, L, Dalgleish, R, Thorisson, GA, et al. Quantifying the use of bioresources for promoting their sharing in scientific research. Gigascience 2013;2:7.

24. European Science Foundation and ALLEA. The European Code of Conduct for Research Integrity. 2011. URL: http://www.esf.org/fileadmin/Public_documents/Publications/Code_Conduct_ResearchIntegrity.pdf.

25. GÉANT Data Protection Code of Conduct. Document GN3-12-215. Version 1.0. GÉANT, 2013. URL: http://www.geant.net/uri/dataprotection-code-of-conduct/V1/Documents/GEANT_DP_CoC_ver1.0.pdf.

26. Dwork, C. Differential privacy: A survey of results. In: *Theory and applications of models of computation*. Springer, 2008:1–19.

27. Li, N, Qardaji, WH, and Su, D. Provably private data anonymization: Or, k-anonymity meets differential privacy. CoRR, abs/1101.2604 2011;49:55.

28. Dwork, C and Roth, A. The algorithmic foundations of differential privacy. Theoretical Computer Science 2013;9:211–407.

29. Aggarwal, CC. On k-anonymity and the curse of dimensionality. In: *Proceedings of the 31st international conference on Very large data bases*. VLDB Endowment. 2005:901–909.

30. Bishop, MA. The Art and Science of Computer Security. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2002.

31. Bild, R, Kohlmayer, F, Brunner, S, et al. Report describing the security architecture and framework. BioMedBridges Project, Deliverable 5.3. 2014. URL: http://www.biomedbridges.eu/sites/biomedbridges.eu/files/documents/deliverables/d5-3_report_describing_the_security_architecture_and_framework-formatted_pdf.pdf.

32. Stevens, WP, Myers, GJ, and Constantine, LL. Structured design. IBM Systems Journal 1974;13:115–139.

33. Information technology - Security techniques - Information security management systems - Overview and vocabulary. ISO 27000. 2009.

34. Shirey, RW. Internet Security Glossary. RFC 4949. IETF, 2007:1–365. URL: https://www.rfc-editor.org/rfc/rfc4949.txt.

35. Pfitzmann, A and Hansen, M. A terminology for talking about privacy by data minimization: Anonymity, unlinkability, undetectability, unobservability, pseudonymity, and identity management. Version v0.34. 2010. URL: https://dud.inf.tu-dresden.de/literatur/Anon_Terminology_v0.34.pdf.

36. McCallister, E, Grance, T, and Scarfone, K. NIST Special Publication 800-122. Guide to Protecting the Confidentiality of Personally Identifiable Information (PII). 2010. URL: https://dl.acm.org/citation.cfm?id=2206206.

37. Chadwick, R and Berg, K. Solidarity and equity: new ethical frameworks for genetic databases. Nature Reviews Genetics 2001;2:318–321.

38. Ewing, AT, Erby, LA, Bollinger, J, et al. Demographic Differences in Willingness to Provide Broad and Narrow Consent for Biobank Research. Biopreservation and biobanking 2015;13:98–106.

39. Hansson, MG, Dillner, J, Bartram, CR, et al. Should donors be allowed to give broad consent to future biobank research? The Lancet Oncology 2006;7:266–269.

40. Hoeyer, K. Informed consent: the making of a ubiquitous rule in medical practice. Organization 2009;16:267–288.

41. Williams, H, Spencer, K, Sanders, C, et al. Dynamic consent: a possible solution to improve patient confidence and trust in how electronic patient records are used in medical research. JMIR medical informatics 2015;3.

42. Fletcher, G, Lockhart, H, Anderson, S, et al. Identity Provider Discovery Service Protocol and Profile. Committee Specification 01. 2008. URL: http://docs.oasis-open.org/security/saml/Post2.0/sstc-saml-idp-discovery.pdf.

43. Procházka, M, Kouřil, D, and Matyska, L. User centric authentication for web applications. In: *Collaborative Technologies and Systems (CTS), 2010 International Symposium on*. IEEE. 2010:67–74.

44. Burr, WE, Dodson, DF, Newton, EM, et al. Electronic Authentication Guideline. NIST Special Publication 800-63-2. 2013. DOI: http://dx.doi.org/10.6028/NIST.SP.800-63-2. URL: http://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-63-2.pdf.

45. auEduPerson Definition and Attribute Vocabulary. Version 2.1. 2009. URL: https://aaf.edu.au/wp-content/uploads/2012/05/auEduPerson_attribute_vocabulary_v02-1-0.pdf.

46. European Parliament. Regulation (EU) No 910/2014 of the European Parliament and of the Council of 23 July 2014 on electronic identification and trust services for electronic transactions in the internal market and repealing Directive 1999/93/EC. OJ L 257, 28.8.2014, p. 73–114 2014.

47. Richer, J and Johansson, L. Vectors of Trust. Draft. IETF, 2015. URL: https://tools.ietf.org/html/draft-richer-vectors-of-trust-02.

48. Morgan, RB, Madsen, P, and Cantor, S. SAML V2.0 Identity Assurance Profiles Version 1.0. Committee Specification 01. OASIS, 2010. URL: http://docs.oasis-open.org/security/saml/Post2.0/sstc-saml-assurance-profile.html.

ADOPT BBMRI-ERIC is funded by the European Union (EU)
Horizon 2020 under Grant number 676550.

Horizon 2020

49. Kumar, S, Walker, D, West, A, et al. Assurance Enhancements for the Shibboleth Identity Provider Draft v17. 2013. URL: https://spaces.internet2.edu/download/attachments/9185/AssuranceReqShibIdPv17.pdf.

50. Recordon, D, Jones, M, Bufu, J, et al. OpenID Provider Authentication Policy Extension 1.0. 2008. URL: http://openid.net/specs/openid-provider-authentication-policy-extension-1_0.html.

51. Brosch, F, Koziolek, H, Buhnová, B, et al. Parameterized Reliability Prediction for Component-Based Software Architectures. In: *Research into Practice: Reality and Gaps*. Ed. by Heineman, GT, Kofroň, J, and Plašil, F. Vol. 6093. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2010:36–51. DOI: 10.1007/978-3-642-13821-8_5. URL: http://dx.doi.org/10.1007/978-3-642-13821-8_5.

52. Orawiwattanakul, T, Yamaji, K, Nakamura, M, et al. User-controlled privacy protection with attribute-filter mechanism for a federated sso environment using shibboleth. In: *P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC), 2010 International Conference on*. IEEE. 2010:243–249.

53. EU Directive. 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. Official Journal of the EC 1995;23.

54. EU Directive. Directive 1999/93/EC of the European Parliament and of the Council on a Community framework for electronic signatures. Official Journal L 1999;13.

55. EU Directive. Directive 2006/123/EC of the European Parliament and of the Council of 12 December 2006 on services in the internal market. 2006.

56. EU Directive. Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector. JL 201, 31.7. 2002, at 37.(Directive on Privacy and Electronic Communications) 2002.

57. Bessani, A, Brandt, J, Bux, M, et al. BiobankCloud: a Platform for the Secure Storage, Sharing, and Processing of Large Biomedical Data Sets. In: *the First International Workshop on Data Management and Analytics for Medicine and Healthcare (DMAH 2015)*. 2015.

58. Bux, M, Brandt, J, Lipka, C, et al. SAASFEE: Scalable Scientific Workflow Execution Engine. Proc. VLDB Endow. 2015;8:1892–1895.

59. Information technology – Security techniques – Privacy framework. ISO/TS 29100:2011. 2011.

60. Parliament, TE and Council of the European Union, the. Position of the Council at first reading with a view to the adoption of a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) - Adopted by the Council on 8 April 2016. 2016.

61. Health informatics – Trusted end-to-end information flows. ISO/TR 21089:2004. 2004.

ADOPT BBMRI-ERIC is funded by the European Union (EU)
Horizon 2020 under Grant number 676550.

Horizon 2020

62. Holmes, N and Emam, K el. Big Data Meets Privacy:De-identification Maturity Model for Benchmarking and Improving De-identification Practices. In: *O'Reilly Strata Rx Conference*. Boston, MA, 2013. URL: http://pt.slideshare.net/kelemam/strata-rx-2013big-datafinal24sepkee.

63. Li, N, Li, T, and Venkatasubramanian, S. t-closeness: Privacy beyond k-anonymity and l-diversity. In: *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*. IEEE. 2007:106–115.

64. Li, T, Li, N, Zhang, J, et al. Slicing: A new approach for privacy preserving data publishing. Knowledge and Data Engineering, IEEE Transactions on 2012;24:561–574.

65. Nergiz, ME, Atzori, M, and Clifton, C. Hiding the presence of individuals from shared databases. In: *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*. ACM. 2007:665–676.

66. Sweeney, L. k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 2002;10:557–570.

67. Machanavajjhala, A, Kifer, D, Gehrke, J, et al. l-diversity: Privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data (TKDD) 2007;1:3.

68. Adam, NR and Worthmann, JC. Security-control methods for statistical databases: a comparative study. ACM Computing Surveys (CSUR) 1989;21:515–556.

69. Denning, DE, Denning, PJ, and Schwartz, MD. The tracker: A threat to statistical database security. ACM Transactions on Database Systems (TODS) 1979;4:76–96.

70. Zhou, B, Pei, J, and Luk, W. A brief survey on anonymization techniques for privacy preserving publishing of social network data. ACM SIGKDD Explorations Newsletter 2008;10:12–22.

71. Institute of Medicine (IOM). Sharing clinical trial data: Maximizing benefits, minimizing risk. Washington, DC: The National Academies Press, 2015.

72. CDC (Centers for Disease Control and Prevention) and HRSA (Health Resources and Services Administration). Integrated guidelines for developing epidemiologic profiles: HIV Prevention and Ryan White CARE Act community planning. Atlanta, GA, 2004. URL: http://www.cdph.ca.gov/programs/aids/Documents/GLines-IntegratedEpiProfiles.pdf (visited on 2015).

73. De Waal, A and Willenborg, L. A view on statistical disclosure control for microdata. Survey Methodology 1996;22:95–103.

74. NRC (National Research Council). Private lives and public policies: Confidentiality and accessibility of government statistics. Washington, DC: National Academy Press, 1993.

75. Office of the Privacy Commissioner of Quebec (CAI). Chenard v. Ministere de l'agriculture, des pecheries et de l'alimentation (141). 1997.

76. U.S. Department of Education. NCES statistical standards. NCES 2003–60. 2003. URL: http://nces.ed.gov/pubs2003/2003601.pdf.

77. CMS (Centers for Medicare & Medicaid Services). 2008 basic stand alone Medicare claims public use files. 2008. URL: http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/BSAPUFS/Downloads/2008_BSA_PUF_Disclaimer.pdf (visited on 2015).

ADOPT BBMRI-ERIC is funded by the European Union (EU)
Horizon 2020 under Grant number 676550.

Horizon 2020

78. CMS (Centers for Medicare & Medicaid Services). BSA Inpatient Claims PUF. 2011. URL: http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/BSAPUFS/Inpatient_Claims.html (visited on 2015).

79. Erdem, E and Prada, SI. Creation of public use files: lessons learned from the comparative effectiveness research public use files data pilot project. 2011 JSM Proceedings 2011:4095–4109.

80. Instructions for Completing the Limited Data Set Data Use Agreement (DUA) (CMS-R-0235L). 2008. URL: http://innovation.cms.gov/Files/x/Bundled-Payments-for-Care-Improvement-Data-Use-Agreement.pdf (visited on 2015).

81. El Emam, K, Paton, D, Dankar, F, et al. De-identifying a public use microdata file from the Canadian national discharge abstract database. BMC Med Inform Decis Mak 2011;11:53.

82. El Emam, K, Arbuckle, L, Koru, G, et al. De-identification methods for open health data: the case of the Heritage Health Prize claims dataset. J. Med. Internet Res. 2012;14:e33.

83. Curcin, V, Miles, S, Danger, R, et al. Implementing interoperable provenance in biomedical research. Future Generation Computer Systems 2014;34:1–16.

84. Jajodia, S, Noel, S, and O'Berry, B. Topological Analysis of Network Attack Vulnerability. In: *Managing Cyber Threats*. Ed. by Kumar, V, Srivastava, J, and Lazarevic, A. Vol. 5. Massive Computing. Springer US, 2005:247–266. DOI: 10.1007/0-387-24230-9_9. URL: http://dx.doi.org/10.1007/0-387-24230-9_9.

85. Barnes, R, Thomson, M, Pironti, A, et al. Deprecating Secure Sockets Layer Version 3.0. RFC 7568. IETF, 2015:1–7. URL: https://www.rfc-editor.org/rfc/rfc7568.txt.

86. Polk, T, McKay, K, and Chokhani, S. Guidelines for the Selection, Configuration, and Use of Transport Layer Security (TLS) Implementations. NIST Special Publication 800-52 Revision 1. 2014. DOI: http://dx.doi.org/10.6028/NIST.SP.800-52r1. URL: http://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-52r1.pdf.

87. Bradner, S. Key words for use in RFCs to Indicate Requirement Levels. RFC 2119. IETF, 1997:1–3. URL: https://www.rfc-editor.org/rfc/rfc2119.txt.

88. Graber, TE, Kopelman, J, Watkeys III, EH, et al. Method and apparatus for tracking the navigation path of a user on the world wide web. US Patent 5,717,860. 1998.

89. Damico, T, Kopelman, J, Wamoglu, SF, et al. Apparatus for capturing, storing and processing co-marketing information associated with a user of an on-line computer service using the world-wide-web. US Patent 5,819,285. 1998.

90. Ingrassia Jr, MI, Shelton, JA, and Rowland, TM. Method for monitoring user interactions with web pages from web server using data and command lists for maintaining information visited and issued by participants. US Patent 6,035,332. 2000.

91. Allard, JE, Treadwell III, DR, and Ludeman, JF. Method, system and apparatus for client-side usage tracking of information server systems. US Patent 6,018,619. 2000.

92. Oberle, D, Berendt, B, Hotho, A, et al. Conceptual user tracking. In: *Advances in Web Intelligence*. Springer, 2003:155–164.

93. Atterer, R, Wnuk, M, and Schmidt, A. Knowing the user's every move: user activity tracking for website usability evaluation and implicit interaction. In: *Proceedings of the 15th international conference on World Wide Web*. ACM. 2006:203–212.

ADOPT BBMRI-ERIC is funded by the European Union (EU)
Horizon 2020 under Grant number 676550.

Horizon 2020

A. Relevant Security & Privacy Concepts

This section provides an overview of the most important concepts in privacy and security, with which BBMRI-ERIC infrastructure will need to deal. It is intended as a summary of information to harmonize necessary knowledge among readers coming with diffent IT backgrounds and specializations. Because of the scope of this field, this section is unable to provide equally deep insights into different topics and is by no means meant as a substitute for dedicated literature (e.g., [30] as well as literature referred to throughout this section).

Parts of this section, namely appendices A.1, A.2, and A.5, use excerpts from Deliverable 5.3 [31] of the BioMedBridges project with permission of the original contributor, Raffael Bild. However, note that *there are two substantial differences in concepts compared to the BioMedBridges Deliverable 5.3: (a)* formal mathematical definition of anonymity using anonymity set, which makes *anonymization distinct from pseudonymization* (see appendix A.5 for further discussion, including explicitly stated incompatibility with ISO 25237 [1], which deals with anonymity in a way incompatible with state-of-the-art computer science), *(b)* introduction of high-security restricted access and low/medium-security restricted access, which is due to the different understanding of the purpose of committee controlled access (see appendix A.4.4 for further discussion).

A.1. Risk Analysis and Management

As proposed in BioMedBridges Deliverable 5.3 [31], we will use DFDs [32] for basic modeling of processes and evaluation of risks. The DFD components are: (a) Data stores (DS), (b) Data flows (DF), (c) Processes (P), and (d) External Entities. On top of standard DFD, [31] proposed to use the following color and line coding: green full line to show elements with open access, red full line for restricted access and red color with dashed lines for restricted or open access. Furthermore the labels on data flows (edges) should specify data types transferred with respect to privacy protection, as defined in appendix A.5 on page 65. A sample DFD is shown in figure 9.

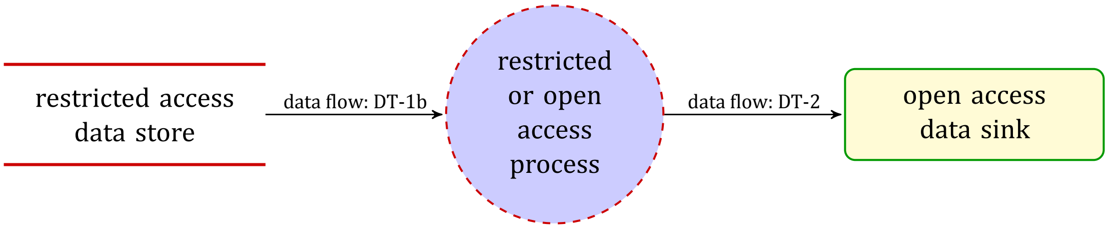
restricted access data store
data flow: DT-1b
restricted or open access process
data flow: DT-2
open access data sink

Figure 9: Sample DFD with color coding proposed in [31]. This DFD is only intended as an example of entities without any real-world meaning.

The risks will be analyzed using STRIDE [2] and LINDDUN [4] methodologies. The STRIDE focuses on security threats, while LINDDUN focuses on privacy threats.

ADOPT BBMRI-ERIC is funded by the European Union (EU)
Horizon 2020 under Grant number 676550.
Horizon 2020

STRIDE [2] identifies the following security risks, connected to the imperiled security properties [33, 34]:

Spoofing threats allow an attacker to pose as something or somebody else. This threatens **authenticity**, which is property that an entity is what it claims to be [33].

Tampering threats involve malicious modification of data or code. This threatens **integrity**, which is property of correctness and completeness of assets [33].

Repudiation An attacker makes a repudiation threat by denying to have performed an action that other parties can neither confirm nor contradict. This threatens **accountability**, which is responsibility of an entity for its actions and decisions [33].

Information disclosure threats involve the exposure of information to individuals who are not supposed to have access to it. This threatens **confidentiality**, which is property that information is not made available or disclosed to unauthorized individuals, entities, or processes [33].

Denial of Service (DoS) attacks deny or degrade service to valid users. This threatens **availability**, which is property of being accessible and usable upon demand by an authorized entity [33].

Elevation of Privilege (EoP) threats often occur when a user gains increased capability. This threatens **authorized access**, which is approval that is granted to a system entity to access a system resource [34].

LINDDUN identifies the identifies the following privacy risks, connected to the imperiled privacy properties:

Linkability of two or more Items of Interest (IoIs), e.g., subjects, messages, actions, allows an attacker to sufficiently distinguish whether these IoIs are related or not within the system. This threatens **unlinkability** of two or more IoIs ... means that within the system ..., the attacker cannot sufficiently distinguish whether these IoIs are related or not [4, 35].

Identifiability of a subject means that the attacker can sufficiently identify the subject associated to an IoI. This threatens **anonymity/pseudonymity**. LINDDUN defines "*anonymity* of a subject ...means that the attacker cannot sufficiently identify the subject within a set of subjects, the anonymity set." LINDDUN defines that "a subject is pseudonymous if a pseudonym is used as identifier instead of one of its real names" [4]. Please note we are using slightly different definition of anonymity as discussed in the appendix A.5.

Non-repudiation allows an attacker to gather evidence to counter the claims of the repudiating party, and to prove that a user knows, has done or has said something. This threatens **plausible deniability**, which means that an attacker cannot prove a user knows, has done or has said something [4, 35].

Detectability of an IoI means that the attacker can sufficiently distinguish whether such an item exists or not. This threatens **undetectability/unobservability**which means that the attacker cannot sufficiently distinguish whether given IoI exists or not [35].

Information disclosure threats expose personal information to individuals who are not supposed to have access to it. This threatens **confidentiality**, which means preserving authorized restrictions on information access and disclosure, including means for protecting personal privacy and proprietary information [36].

Content unawareness indicates that a user is unaware of the information disclosed to the system. This threatens **content awareness**which means the user needs to be aware of the consequences of sharing information [4].

Policy and consent non-compliance means that even though the system shows its privacy policies to its users, there is no guarantee that the system actually complies to the advertised policies. This threatens **policy and consent compliance**, which ensures that the system's (privacy) policy and the user's consent ... are indeed implemented and enforced. [4].

Mapping of risks described by STRIDE and LINDDUN to the DFD entities is shown in tables 6 and 7.

Security property	STRIDE security threats	DF	DS	P	EE
Authentication	Spoofing			X	X
Integrity	Tampering	X	X	X	
Non-repudiation	Repudiation		X	X	X
Confidentiality	Information disclosure	X	X	X	X
Availability	Denial of service	X	X	X	
Authorization	Elevation of Privilege			X	

Table 6: Mapping STRIDE security threats and countermeasures to data flow diagram element types (see Tables 9-5 and 9-8 in Chapter 9 of [2]).

Privacy objective	LINDDUN privacy threats	DF	DS	P	EE
Unlinkability	Linkability	X	X	X	X
Anonymity & Pseudonymity	Identifiability	X	X	X	X
Repudiation	Non-Repudiation	X	X	X	
Undetectability & unobservability	Detectability	X	X	X	
Confidentiality	Information disclosure	X	X	X	
Content awareness	Content unawareness				X
Policy & consent compliance	Policy/consent noncompliance	X	X	X	

Table 7: Mapping LINDDUN privacy threats and objectives to DFD element types (see Tables 4 and 6 in [4])

.

ADOPT BBMRI-ERIC is funded by the European Union (EU)
Horizon 2020 under Grant number 676550.

Horizon 2020

The overall **risk level** is qualitatively assessed using **likelihood of a threat** and **level of impact** as shown table 8.

Likelihood of a threat	Level of impact		
	Low (+)	Medium (++)	High (+++)
Low (+)	+	+	++
Medium (++)	+	++	+++
High (+++)	+	++	+++

Table 8: Qualitative risk assessment.

A.2. Sensitivity of Information and Biological Material (Samples)

A.2.1. Sensitivity of Information

Open/public information Information that is available publicly without any access restrictions. Examples include public domain datasets and information, datasets available under open licenses such as Open Data Commons Open Database License (ODbL).[16]

Information with higher integrity requirements A specific subclass of the previous class, where information is available publicly without any access restrictions, but that is needs to have its integrity preserved and recipient of the information must be able to verify its integrity.

Protected information The information, that requires access restrictions, be it to protect intellectual property, to protect privacy of individuals, or for any other reason. There are various types of access restrictions as further discussed in the next appendix A.4.1.

Protected information with privacy impact. A specific subclass of the previous class, where the reason for protection is to protect privacy of individuals. Examples of this information include any information that may identify an individual, information about sensitive attributes of the individual (e.g., diseases, salary, etc.).

A.2.2. Informed consent

Informed consent is a consent of an individual, typically a patient or a donor, that he/she agrees with the fact that his/her material and/or data is collected for given purpose. When processing any samples/data of patients/donors, the custodian of the material (typically a biobank) has to collect and safely store informed consent, or the this informed consent must be available to the custodian from the originating institution (a healthcare facility from which

[16] http://opendatacommons.org/licenses/odbl/

ADOPT BBMRI-ERIC is funded by the European Union (EU)
Horizon 2020 under Grant number 676550.

Horizon 2020

the biobank receives the samples/data). Before processing any human samples or data, the informed consent must be examined if the intended purpose is compliant with it.

There are ongoing discussions on national and international levels about acceptable forms of informed consent, whether generic consent for all the future research purposes is acceptable or whether specific consent must be given. These discussion are often motivated to prevent commercial use of privacy-sensitive information, but it is not uncommon that results of the discussion have unintended impact into biomedical research [37–41]. This field is the expertise of Common Service ELSI[17] of BBMRI-ERIC and any issues should be consulted with this body.

A.2.3. Material Transfer Agreement (MTA) and Data Transfer Agreement (DTA)

These transfer agreements specify conditions, under which the data or biological material (samples) is handed over from the repository to the user. The transfer agreements for data are commonly called DTAs, while biological material is covered by MTAs.

Both MTAs and DTAs may include statements that the data/samples may be used only for the purpose specified in the access application. This is necessary to ensure that both data and material is used in policy and consent compliant way. MTAs often require that any leftovers of samples must be either demonstrably destroyed or returned to the biobank.

A.3. Authentication

Authentication might be a slightly confusing term, as it needs to comprise two equally important steps, one of which is sometimes also called “authentication”: (a) **registration** process, which binds the virtual identity to the physical identity of the person (e.g., by showing up in registration office with government-issued ID card while creating the virtual identity), and (b) **authentication instance**, which is verification of the persons virtual identity (e.g., a person proves possession of her virtual identity using a password)..

In this section, we will provide a brief overview of authentication architectures (appendix A.3.1), commonly used levels of assurance of persons physical and virtual identities (appendix A.3.2), problems of identity merging for persons possessing multiple virtual identities (appendix A.3.3), as well as aspects related to the robustness of the authentication systems (appendix A.3.4. Since authentication often provides additional means for authorization, we will discuss also attribute issuing as a part of the authentication (appendix A.3.5). Finally, we will conclude with references to the regulations that constitute legal framework to the authentication (appendix A.3.7).

[17] `http://bbmri-eric.eu/common-services`

ADOPT BBMRI-ERIC is funded by the European Union (EU)
Horizon 2020 under Grant number 676550.
Horizon 2020

A.3.1. Architecture of Authentication

Centralized authentication Centralized authentication architecture means that the identity management is implemented by a single organization. On the technology level, it may still be implemented as a distributed system for performance and robustness reasons, but we understand it as a centralized authentication architecture for the purpose of this document if it spans single organization only. Such authentication architecture can be easily implemented when low assurance of user identity (see appendix A.3.2) is sufficient for given application (e.g., such as Google ID or Facebook ID).

Advantages of this approach include (a) adherence to a single set of authentication policies, which result in (b) easily achievable consistence of registration process. Because the organization is typically responsible for both providing user authentication and subsequent services for the users, the other advantage is that (c) the provided services can implement consistent high-level availability for both authentication service as well as for the other services which depend on authentication service.

The main disadvantage of centralized authentication is lack of scalability for infrastructures which have large user base coming from different institutions and countries, especially (a) if registration process includes validation of government-issued ID documents and (b) if authentication system is supposed to provide assertions about user, such as the fact that the user is employed by some institution at the time of authentication.

Federated authentication Federated authentication systems integrate authentication services of multiple institutions. In order to describe such systems consistently and to work with them in the rest of the document, we will introduce Identity Provider (IdP), Service Provider (SP), and Where Are You From service (WAYF)/Discovery Service (DS) terms, which come from Shibboleth identity management system and Security Assertion Markup Language, Version 2.0 (SAML V2.0) [42] respectively, but they are applicable more generally. IdP is the actual authentication service at an institution which verifies a person's virtual identity and Service Provider (SP) is any service provided to the person that consumes the virtual identity and uses it for authorization purposes, as shown in figure 10. Several different IdPs can be integrated together into a federation using component called WAYFs, which allows the person to choose, which institution will be used for authentication (see figure 11 for example of such communication). Inherently, federated authentication also implies separation between IdPs and SPs, each of which may come from a different administrative domain (typically different organization or organization units).

These systems are now becoming widely available in the various flavors: research and educational communities have successfully established identity federations such as eduID[18]; commercial companies having organized themselves in OpenID[19] or at least provid-

[18] eduID activities are organized by GÉANT (formerly by TERENA), see https://wiki.refeds.org/display/GROUPS/EduID+Working+Group, with national nodes being known eduID.yy, where .yy corresponds to the national DNS domain.

[19] http://openid.net/

ADOPT BBMRI-ERIC is funded by the European Union (EU)
Horizon 2020 under Grant number 676550.

Horizon 2020

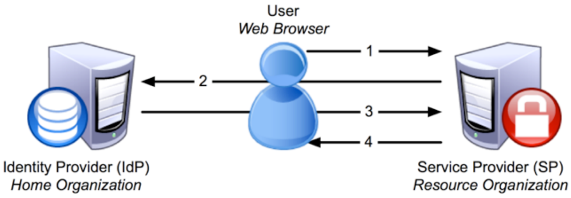

Figure 10: Simple interaction of an IdP and a SP (without WAYF/DS). The diagram starts with user accessing the Resource (1). See `https://wiki.shibboleth.net/confluence/display/CONCEPT/Home` for more details.
Source: `https://wiki.shibboleth.net/confluence/download/attachments/4358538/sso-flow.png?version=2&modificationDate=1249311729063&api=v2`

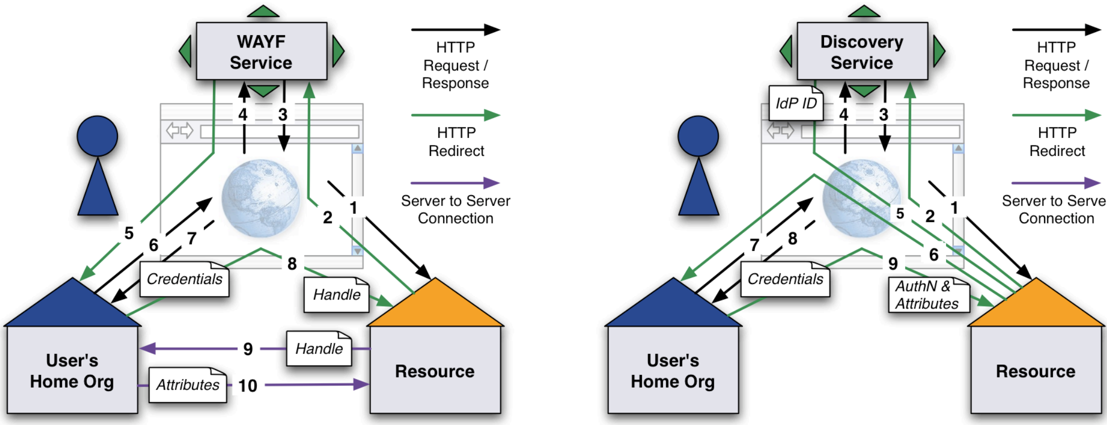

Figure 11: Interaction of an IdP (User's Home Org), a SP (Resource), and a WAYF or DS. The diagram starts with user accessing the Resource (1). See `https://www.switch.ch/aai/support/tools/wayf/` for more details.
Source: `https://www.switch.ch/aai/support/tools/wayf/wayf-vs-ds.png`

ADOPT BBMRI-ERIC is funded by the European Union (EU)
Horizon 2020 under Grant number 676550.
Horizon 2020

ing comparable interfaces such as Facebook Connect[20]; and there are pilot efforts of government-backed identity federations called STORK discussed in appendix A.3.2 on page 59.

The major advantage of this system stems from the fact, that the authentication of a user is implemented by an institution with which the user has a close relation, typically some form of legal contract (e.g., employment contract). Thus the institution can also provide real-time or near real-time assertion on the status of the user. Furthermore, the institution typically validates user identity to the level that is acceptable at least for LoA 2 (see appendix A.3.2 below). Another advantage of the federated authentication system is that they allow for Single Sign On (SSO) even across multiple administrative domains. Thus a user can log in once and have access to multiple resources from the same administrative domain, or even from different administrative domains that enjoy mutual trust.

Disadvantages of federated authentication include (a) online dependence on availability of several components of a distributed system, which naturally threatens availability for users in the real world, (b) problems with consistent implementation of policies in a distributed system spanning multiple administrative domains, (c) need to solve a situation when a user does not have affiliation to any IdP in the given federated authentication infrastructure. This results in the need for some "catch-all" IdPs, which may be hard to implement at the same LoA as "normal" IdPs. Another aspect is that (d) user's home institution releases privacy sensitive attributes into other administrative domains, and thus user must be given an option to control what is released about him/her, as further discussed in appendix A.3.5. Last but not least, (e) if a user has affiliation with multiple institutions, it may be desirable to merge credentials/attributes coming from different institutions in order for the user to obtain the requested service.

User-centric authentication Recognizing problematic scalability of centralized authentication as well as disadvantages associated with commonly used approaches to federated authentication, user-centric authentication is now explored [43]. One of the proposed approaches is to have a "wallet" for each user, where the user stores time-limited "ID cards" provided by the IdPs. This approach addresses both the problem of online availability IdP, as well as allowing user direct control of released attributes. Unfortunately, user-centric authentication systems are not yet available in practice as of time of writing this document, resulting in various "hacks" for federated authentication systems to address the same issues.

A.3.2. LoA

The main purpose of LoA is to allow service providers to assess the trustworthiness of the asserted identity of the user. Generally accepted approach to defining the level of assurance

[20] https://developers.facebook.com/blog/post/2008/05/09/announcing-facebook-connect/, https://developers.facebook.com/docs/facebook-login

ADOPT BBMRI-ERIC is funded by the European Union (EU)
Horizon 2020 under Grant number 676550.
Horizon 2020

comes from NIST SP 800-63-2 [44], while a nice summary of implementation in practical federated authentication systems is available on the Tuakiri Federation website[21] and in [45].

There are two main aspects of level of assurance:

1. the strength of the process of *identity proofing and verification* (see [46, Article 8 and 9(1)]) of the person during registration of the user (we will use **identity verification** in the following text, but sometimes identity vetting is used for the same purpose),
2. the strength of *technical means* used for verification in the *particular authentication instance* (**authentication instance** will be used in the text).

Each level of assurance is then discussed using those two aspects.

Level 0 This is not officially defined and thus can be considered non-standard, but we use it as a conceptual baseline in case no identity verification has been done at all, while still having a notion of "a user". This can be used, e.g., for storing personal preferences that are not considered personal at all, or for tracking behavior of the user.

- **Identity verification:** No explicit registration (e.g., user agreeing to the terms and conditions of the service, use of website using cookies).
- **Authentication instance:** Private token directly provided by a user, e.g., a cookie in a web browser. No action is expected by the user. No secure communication is required and the token can be sent as plain text over the network (e.g., in HTTP protocol).

Level 1 Authentication on this level only demonstrates any kind of relation to the identity provider. This authentication is provided by Facebook and Google IdPs, but also various "hostel" services provided by eduID.xx IdPs, which are designed to serve users with no affiliation to any of the member institutions.

A secure communication channel is not required, it may be prone to attacks such as dictionary password attacks. However, this is intentionally chosen as a compromise between security and convenience for the users.

Note that any higher LoA also fulfills requirements of LoA 1.

- **Identity verification:** No identity proof is required at this level and any type of relation with the identity provider is acceptable (e.g., user self-registers using her email address).
- **Authentication instance:** Successful authentication requires user to demonstrate she/he is in possession of the token (e.g., knows a password). It is only required that plain-text passwords or tokens are not sent over the network (utilizing, e.g., simple challenge-response protocols), but there is no requirement to use a secure communication channel.

[21] `https://tuakiri.ac.nz/confluence/display/Tuakiri/Levels+of+Assurance`

ADOPT BBMRI-ERIC is funded by the European Union (EU)
Horizon 2020 under Grant number 676550.

Horizon 2020

Level 2 This is the minimum LoA for which the identity of a person is validated. However, as it is still prone to stealing credentials of the user because of just a single factor (e.g., password), it should not be used for access to really sensitive data.

- **Identity verification:** Presentation of personal identifying materials is required, supporting both in-person and remote registrations. For in-person registrations, the applicant must present a government-issued photo ID. For remote registrations, the applicant provides references to and asserts to current possession of a government-issued photo ID and a secondary ID or another secondary identification. The applicant must provide at minimum their name, date of birth, address and phone number.
- **Authentication instance:** Single factor is used for remote authenticated network access. It allows for passwords and PINs, as well as for any other token methods of higher LoAs. Secure communication channel is required; eavesdropping, replay attack and on-line token guessing attacks must be prevented.

Level 3 This is the first practical implementation of the multi-factor authentication, with the identity card of the person checked against records as a part of the registration process.

- **Identity verification:** All the requirements of LoA 2 must be fulfilled, but additional validation of IDs by the registrar is required, implemented by doing record checks.
- **Authentication instance:** Possession of a cryptographic tokens must be proved using cryptographic protocol. Three kinds of tokens are acceptable for LoA 3: (a) soft cryptographic tokens, (b) hard cryptographic tokens, (c) one time passwords. The secure communication channel must be protected against eavesdropping, replay attacks, on-line token guessing attacks, verifier impersonation, and man-in-the-middle attacks. Two-factor authentication is required: password or biometric must be used as an addition to the primary cryptographic token.

Level 4 This is the highest practical level of assurance for remote access, with mandatory multi-factor authentication and biometric recording of non-repudiation of the registration process. Because of FIPS 140-2 Level 2 and Level 3 requirements on the hardware and physical security, this may be hard to deploy in practice in distributed infrastructures spanning multiple administrative domains.

- **Identity verification:** All the requirements of LoA 3 must be fulfilled, but remote registration is not allowed and the applicant must appear in person before the registration officer. Two independent ID documents must be also presented and verified. One of these ID documents must be a current government issued ID card with (a) photo, (b) either address or nationality. In order to ensure non-repudiation by the applicant, a new biometric recording must be performed as a part of registration.
- **Authentication instance:** Authentication is intended to provide the highest practical authentication assurance that still allows for remote network access. All of the requirements of LoA 3 must be fulfilled, but only hard cryptographic tokens are allowed, FIPS 140-2 cryptographic module validation requirements are

ADOPT BBMRI-ERIC is funded by the European Union (EU)
Horizon 2020 under Grant number 676550.
Horizon 2020

stronger, and the subsequent critical data transfer processes must be authenticated using a key created as a part of the authentication process. The tokens must be validated by a hardware cryptographic module at FIPS 140-2 Level 2 or higher, with at least FIPS 140-2 Level 3 physical security.

Another set of LoAs has been proposed[22] by The Interoperable Global Trust Federation (IGTF)[23]: ASPEN, BIRCH, CEDAR, and DOGWOOD. The textual levels are used to avoid confusion with the number-based LoAs described above.

There is an ongoing work [47] of extending simple scalar LoAs to vectors describing *identity proofing*, *primary credential usage*, *primary credential management*, and *assertion presentation* as orthogonal elements of a vector. This approach is designed to be backward compatible with the scalar LoA by mapping certain vectors to the LoA scalars. But practical adoption in AAI is still an open question.

For access to public information, LoA 0 or 1 is sufficient. LoA 1 is often also used for accessing private information (e.g., projects proposals including information about people and budget stored in Google Documents with access based on Google ID), but such practice should be avoided if possible. For any sensitive data or for consuming resources of an infrastructure, minimum of LoA 2 should be considered. Current implementations of academic identity federations routinely support LoA 2. As multi-factor authentication are often overly complicated for users, benefits of LoA 3 or 4 and the value of the protected resource/information should be carefully examined for each service on case-by-case basis. LoA 3 or 4 are now being discussed by some academic and research infrastructures, but practical availability is very limited.[24]

Support for LoA is available in SAML V2.0, as a part of the Identity Assurance Profiles Version 1.0 [48]. They are also available in practical implementations like Shibboleth [49], which are basis for implementation of academic identity federations such as eduID.

It is also supported in OpenID as a part of OpenID Provider Authentication Policy Extension 1.0 [50].

An interesting solution with widely available IdPs very appropriate for the BBMRI-ERIC purposes will be **government-backed identity**. This approach has been explored and prototyped by Secure idenTity acrOss boRders linked (STORK)[25] and Secure idenTity acrOss boRders linked 2.0 (STORK 2.0)[26] projects and needs a working robust implementation in place to become dependable for real-world SPs. In principle, a government-backed IdP should provide at least strong registration (verification of identity) of LoA, which may be either accompanied by strong authentication instance or not. If the government-backed IdPs comes with an

[22] `https://www.eugridpma.org/guidelines/loa/IGTF-LoA-authN-set-20150930-v11.docx`

[23] `https://www.igtf.net/`

[24] Multi-factor authentication has been deployed by *TSD: a Secure and Scalable Service for Sensitive Data and eBiobanks*, based on personal communication with the developers. Practical implementation is based on Google Authenticator.

[25] `https://www.eid-stork.eu/`

[26] `https://www.eid-stork2.eu/`

ADOPT BBMRI-ERIC is funded by the European Union (EU)
Horizon 2020 under Grant number 676550.

Horizon 2020

insufficiently strong authentication instance, it can be improved using alternate IdP together with identity linking (described in the appendix A.3.3 below).

A.3.3. Merging/Linking User Identities from Different Identity Providers

A common problem in the real world is that one person has several identities in the digital world: identity provided by government (national ID or social security IDs), identities provided by employee or school, identities provided by various services such as Google, Facebook, or Microsoft, etc. This does not map onto real world properly, as a single real person should have single digital identity, complemented by various attributes or additional assertions about the person, such as her employment status, etc.

A proper solution to this is introduction of user-centric approach to identity federations, such as ADITI [43], which is however still subject to research and cannot be easily deployed in real-world due to lack of production implementations. In these systems, the user is the maintainer of her identity and the current identity providers become just attributes/assertions providers, which provide time-limited signed assertions to the user, who may relay these assertions to the service providers upon her discretion.

Interim solution to this problem is often provided by additional AAI layer(s), such as the Perun system [7], implementing several authorization-related functionality at once: identity merging or linking (we will use term “merging” in this document), issuing of additional attributes issuing, as well as management of virtual groups (participation in the groups translates into issuing additional attributes about the user for the SP).

A.3.4. Increasing Robustness of Distributed Authentication Infrastructures

As already mentioned in description of federated authentication architectures, another important practical problem is the need for online (synchronous) availability of multiple entities of a distributed system: identity provider, service provider, and possibly other systems such as WAYF, DS, or attribute authorities (see appendix A.3.5). It is a well-known property of distributed systems, however, that the more synchronous dependencies are in the distributed system, the more the system becomes fragile [51]. The user may then easily start blaming service provider for not ensuring appropriate/agreed service availability, while the actual problems lie out of the reach of both service provider and the user. Especially in large institutions, the user have very limited options to ask for increased availability of their institutional IdP. Increasing availability of federation infrastructure elements such as WAYF may easily be out of reach of both user and service provider.

This problem has given rise to concept of **Proxy IdP** in EGI, Authentication and Authorisation for Research and Collaboration (AARC)/VO Platform as a Service provided by GÉANT (VOPaaS) [11, 12], or ELIXIR, where the identities from the originating IdPs are cached by the

ADOPT BBMRI-ERIC is funded by the European Union (EU)
Horizon 2020 under Grant number 676550.
Horizon 2020

Proxy IdP, which is either in the same administrative domain as the SPs, or at least should be easier to deal with from the SP's or user's side.

Furthermore, the Proxy IdP can also inject additional attributes. This may help if the originating IdP does not provide all the attributes that are needed; this should be, however, relied upon with caution, as only a limited set of attributes can be issued: Proxy IdP cannot make assertions that are inherent to the user's home institution (e.g., employee or student status).

A.3.5. Issuing of Attributes

Attributes can be issued either by the IdPs, or they can be issued by third party services such as Perun-based management of virtual user groups mentioned above. In either case because of the privacy protection, the user needs to be "in charge", i.e., has to be able to approve or disapprove the attributes that are being released about her from IdPs or attribute services to the SPs. Current implementations of such a system for Shibboleth include uApprove[27] and uApproveJP[28] [52].

For environments like BBMRI-ERIC, the following attribute-related assertions are relevant:

institutional affiliations/roles which assert the user has certain relation to the given organization, e.g., an employee, a student, or a faculty member of an educational institution,

project affiliations/roles which assert the user has affiliation to a project or even more specifically that the user has certain role in a project,

group affiliation which could be understood as generalization of the previous two approaches, where it is possible to describe adherence of the user also to any other virtual group or subgroup.

The project-based affiliations are of particular interest in environments like BBMRI-ERIC, where access to samples/data is often governed by the adherence of the users to the projects that have been examined by ethical committees, and whose research intents must be compared to the informed consent that is available for given samples/data. See also discussion of project-based Role-Based Access Control (RBAC) in appendix A.4.3.

A.3.6. Delegation of Roles

A person may wish to delegate his/her role to another person. Typically, a PhD student may be entitled by his supervisor to take over some of simple technical tasks. Therefore, it is necessary to *distinguish between the role and the attributes which were used to assign the role*

[27] https://www.switch.ch/aai/support/tools/uapprove/

[28] https://meatwiki.nii.ac.jp/confluence/x/aQLO

ADOPT BBMRI-ERIC is funded by the European Union (EU)
Horizon 2020 under Grant number 676550.

Horizon 2020

to the person initially. While the person receiving the delegation will receive the role including all related entitlements, he/she will not receive the attributes.

Another important aspect is to distinguish between *delegable roles and non-delegable roles*. It is, however, recommended to minimize the non-delegable roles, as the delegation of roles is necessary in practice and making roles non-delegable often results in impersonation of users by sharing their credentials, which is much riskier behavior.

Another aspect is that delegation may introduce need for finer granularization of roles, as the delegator may need to *delegate only a subset of his/her entitlements*.

A.3.7. Legal Requirements for Security & Privacy

In the European Union (EU), the following regulations apply:

- Directive on the protection of personal data 95/46/EC [53],
- Directive 1999/93/EC on a Community framework for electronic signatures [54],
- Directive 2006/123/EC on services in the internal market [55],
- Directive 2002/58/EC concerning the processing of personal data and the protection of privacy in the electronic communication sector [56].

Another part of the framework will be General Data Protection Regulation (GDPR), obsoleting 95/46/EC. Consensus has been reached[29] between the European Commission, Parliament, and Council (so-called ’trilogue’ meetings) on December 15, 2015 and the GDPR has been submitted for approval process in Parliament. Consequences of GDPR are yet to be understood.

A.4. Modes of Access and Authorization

This section deals with the mode of access to the samples and data and with the concept of authorization, related to any restricted access. The basic access modes are discussed in appendix A.4.1, including open access, restricted access and committee-controlled access.

Authorization is the process of granting or denying access to given object or service. We particularly describe two main automated authorization approaches relevant for purposes of the BBMRI-ERIC: rule-based access control in appendix A.4.2 and role-based access control in appendix A.4.3.

[29] http://europa.eu/rapid/press-release_IP-15-6321_en.htm

ADOPT BBMRI-ERIC is funded by the European Union (EU)
Horizon 2020 under Grant number 676550.

Horizon 2020

A.4.1. Access modes to the data/samples

Based on sensitivity of the data and associated risks, as well as on access policies, the access control to the information and material can be divided into the following classes:

Open/public access Access is not restricted and the data is publicly available.

Restricted access This includes both RBAC and Mandatory Access Control (MAC), as well as committee-controlled access described below. Choice of specific strategy depends on practical implementability, as discussed in appendix A.4.

For practical purposes of implementation in the BBMRI-ERIC context, such minimization of user annoyance by more complicated security procedures, we will differentiate between the two levels of restricted access:

High-security restricted access requires higher level of assurance of the accessing person (implementation requirements discussed later in this document), existence of ethically approved project and ensuring that samples/data use in the project is compliant with the informed consent accompanying the samples/data.

High-security restricted access is used for controlling access to the IT services implementing use cases with high risk of security threats (covered by STRIDE) or privacy threats (covered by LINDDUN). See section 3.5 on page 40 for results of risk analysis.

Low/medium-security restricted access covers all other types of restricted access.

Low/medium-security restricted access covers low/medium risks, see again section 3.5 on page 40 for results of risk analysis for use cases. See also comment on the specifics of S+UCs-1 in that section, as some services may be available in both open access mode and low/medium security mode, sharing different level of information.

Committee-controlled access Is a specific subclass of restricted access, where the access is decided for a specific user or user group and/or for a specific purpose by a (Data|Samples) Access Committee (AC). Such a committee typically consists of representatives of custodians of samples/data: e.g., when a researcher has samples hosted by a biobank, the AC may be the researcher, or the biobank, or both, depending on the contract between the researcher and the biobank hosting the samples.

Primary reason for committee-controlled access is to give sample/data custodians greater degree of control (i.e., manual) for what purposes these are used. Typically, it is combined with high-security restricted access—but not necessarily always.

Technically, the committee-controlled access can be implemented, e.g., by Resource Entitlement Management System (REMS) [8].

ADOPT BBMRI-ERIC is funded by the European Union (EU)
Horizon 2020 under Grant number 676550.

Horizon 2020

A.4.2. Rule-based access control: Discretionary Access Control (DAC) and Mandatory Access Control (MAC)

Discretionary Access Control (DAC) and MAC approaches are rule-based authorization systems, which differ mainly in who sets the rules for a given object or service [30].

DAC is an approach where each object has an owner and the owner specifies access rules for individual people to the selected objects.

MAC is an approach where the system administrator sets up access control rules for individual people to selected objects. Inheritance of access control is typically supported, so that the child object inherits permissions from parents, unless explicitly stated otherwise. It is called mandatory, since the owner of the data is not allowed to alter the access control rules.

A.4.3. Role-Based Access Control (RBAC)

RBAC is an approach based on the roles that are assigned to the person and the authorization is done based on the person's role.

Attribute-based RBAC Roles can be also derived from the attributes that are released from IdPs or attribute services as discussed in appendix A.3.5.

In practice, there might be problems with this approach due to insufficient attributes being released by the IdPs to the SPs, mostly because of privacy concerns in the non-user-centric federated identity systems. Similar to reliability issue described above, the individual user may not be able to influence policy of her IdP, especially in larger institutions. Therefore concept of additional attribute authorities (or Proxy IdP) may need to be used, increasing formal burdens as the attributes must be issues on provable basis.

Example of attributes available in practical academic federations include[30]:

- identifier of the person: `eduPersonTargetedID`,
- name of the person: `commonName`, `displayName` (while some federations also request `givenName`, `surname`, `commonNameASCII`),
- organization with which the person is affiliated: `schacHomeOrganization`,
- type of affiliation of the person: `eduPersonScopedAffiliation`, which can be `{faculty, student, staff, alum, member, affiliate, employee, library-walk-in}@organization.org`
- other attributes: `mail`.

[30] This list of examples is based on eduGAIN recommended attributes, `https://wiki.edugain.org/IDP_Attribute_Profile:_recommended_attributes`

ADOPT BBMRI-ERIC is funded by the European Union (EU)
Horizon 2020 under Grant number 676550.
Horizon 2020

Another problem with pure attribute-based RBAC is delegation (see appendix A.3.6), where a person needs to delegate his/her role to some other person (if the person to receive the delegation does not have the same attributes as the delegator). Hence the RBAC based directly on attributes from IdPs is more useful for initial assignment of roles to the people, and then working explicitly with roles to allow also for delegation.

Project-based RBAC This is a variant of the RBAC where each user is strictly related to one or more projects, and the access control is based on those projects. This model often comes with additional non-interlinking condition, where the same user has permission to work with data set A for project 1 and data set B for project 2 respectively, but is not allowed to merge or correlate A and B. In order to map such requirements on existing access control systems, the common approach is to introduce new identities, comprised of a subset of Cartesian product of users and projects; i.e., identities like user1_project1, user1_project2, user2_project1, etc. The access control is then set based on the project affiliation of the identity. Such an approach has been implemented BiobankCloud platform[31] [57, 58], MOSLER[32] and TSD.[33]

A.4.4. Semantic development of committee-controlled access

Note that there is a subtle semantic shift since BioMedBridges Deliverable 5.3 [31] in how we work with committee-controlled access.

The Deliverable used the committee-controlled access as a further risk reduction mechanism beyond normal restricted access. Based on additional experience with the practical use of committee-controlled access in biobanks, we consider it rather an organizational measure for manual evaluation of compliance of the informed consent with the research intent of the project or to allow for prioritization of projects for resources that can be depleted (typically biological samples).

Hence we opted for separation of the risk management from the committee-controlled access, which resulted in introduction of high-security restricted access and low/medium-security restricted access introduced in appendix A.4.1. The committee-controlled access then remains orthogonal and can be combined with any restricted access mode.

A.5. Privacy-Enhancing Technologies (PET)

Privacy-Enhancing Technologies (PET), defined, e.g., in ISO 29100 [59] and [35]), deal with problems of protecting privacy of individuals in information technologies and information systems. As a part of the PET, we introduce the following definitions:

[31] http://www.biobankcloud.com/

[32] https://bils.se/resources/mosler.html

[33] https://www.norstore.no/services/TSD

ADOPT BBMRI-ERIC is funded by the European Union (EU)
Horizon 2020 under Grant number 676550.
Horizon 2020

DT-1 **Personal data.** According to the definition of the GDPR: "'personal data' means any information relating to an identified or identifiable natural person ('data subject')" [60]. This data type can be further divided into:

DT-1a **Data related to individual identifiable person.**

This typically includes original data in the patients healthcare records, questionnaires, etc., including patients identifiers.

DT-1b **Coded data**, which typically means that some identifying information (e.g., names, civic number or social security ID) has been removed and potentially replaced with a code (a "pseudonym", but the removal of the information may not be sufficient in the sense of GDPR pseudonymization, see DT-3). This is an auxiliary type introduced in this paper, which is not directly described by the GDPR but which is often used in practice.

DT-2 **(De facto) anonymized data** *Anonymity* of a subject from the perspective of an attacker means that the attacker cannot sufficiently identify the subject within a set of subjects, the anonymity set [35]. This data is therefore no longer personal, but it bears non-zero risk of re-identification. Anonymization must be always understood in a given context considering likelihood of attacks, e.g., from adversaries with specific background knowledge.

DT-3 **Pseudonymized data** In the strict interpretation of GDPR, this is data which if the key is not known, it can be considered anonymous[34] (i.e., with the same requirements as for DT-2).

This definition differs from previously used definitions of pseudonymization, see, e.g., [1, 35], and there is pending debate on implications of such definition (c.f. DT-1b "Coded data").

DT-4 **Data from deceased people** does not fall under General Data Protection Regulation but enjoys legal protection under different national jurisdictions. Also professional secrecy does not end with the death of a person (patient).

DT-5 Non-human data that does not contain any trace of personal/human data and thus is not privacy sensitive (e.g., temperature monitoring data from sample storage systems).

Furthermore, we introduce the following auxiliary definitions to simplify the text:

Privacy-enhanced data is data, for which identifiers have been removed or replaced using a method that is either impossible to revert or that would require unreasonable amount of time and manpower without knowing the initial information.

[34] GDPR definition reads as follows: "means the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or *identifiable* natural person;"

ADOPT BBMRI-ERIC is funded by the European Union (EU)
Horizon 2020 under Grant number 676550.

Horizon 2020

> This term can be used for denoting (de facto) anonymized data or pseudonymized data or coded data, and we will use it in this document to cover both. This is consistent with the specification in ISO 21089 [61].

It is worth mentioning there is disagreement among different authors regarding PET terminology. Namely ISO 25237 [1] understands pseudonymization as a particular type of anonymization – see the definition of pseudonymization:

> pseudonymization: particular type of anonymization that both removes the association with a data subject and adds an association between a particular set of characteristics relating to the data subject and one or more pseudonyms

and a similar view is shared by Holmes in [62, slide 16ff]. This is inconsistent with the notion of anonymization in the mathematical sense (see definitions above) and will not be used in this document.

It is also important to understand that anonymization is not a definitive process, it is relative to the risks, and thus it is expected to evolve into a procedural definition that is time-dependent and circumstances-dependent. The newly prepared GDPR already assumes this and Recital 23 states as follows[35]:

> The principles of data protection should apply to any information concerning an identified or identifiable natural person. To determine whether a person is identifiable, account should be taken of all the means reasonably likely to be used either by the controller or by any other person to identify or single out the individual directly or indirectly. To ascertain whether means are reasonably likely to be used to identify the individual, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration both available technology at the time of the processing and technological development.

A.5.1. Anonymization

As described in [63] and [64], anonymization is typically applied to a table which contains microdata in the form of records (rows) that correspond to an individual and have a number of attributes (columns) each. These attributes can be divided into three categories:

1. Explicit identifiers are attributes that clearly identify individuals (e.g., name, address).
2. Quasi-identifiers are attributes whose values taken together could potentially identify an individual (e.g., birthday, ZIP code).
3. Attributes that are considered sensitive (e.g., disease, salary).

[35] http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+TA+P7-TA-2014-0212+0+DOC+XML+V0//EN

ADOPT BBMRI-ERIC is funded by the European Union (EU)
Horizon 2020 under Grant number 676550.
Horizon 2020

Anonymization aims at processing such a microdata table in a way that it can be released without disclosing sensitive information about the individuals. In particular, three threats are commonly considered in the literature that can be mitigated using different anonymization methods:

1. Identity disclosure, which means that an individual can be linked to a particular record in the released table [63].
2. Attribute disclosure, which means that additional information about an individual can be inferred without necessarily having to linking it to a specific record in the released table [63].
3. Membership disclosure, which means that it is possible to determine whether or not an individual is contained in the released table utilizing quasi-identifiers [65].

According to [63], as a first step in the data anonymization process, explicit identifiers are removed. However, this is not enough, since an adversary may already know identifiers and quasi-identifiers of some individuals, for example from public datasets such as voter registration lists. This knowledge can enable the adversary to re-identify individuals in the released table by linking known quasi-identifiers to corresponding attributes in the table. Thus, further anonymization techniques should be employed, such as **suppression** or **generalization**. Suppression denotes the deletion of values from the table that is to be released. Generalization basically means the replacement of quasi-identifiers with less specific, but still semantically consistent values. It is worth noting that both suppression and generalization decrease the information content of the table, so in practice, these techniques should be applied to the extent that an acceptable level of anonymization is achieved while as much information as possible is preserved.

In order to quantify the degree of anonymization, multiple metrics have been proposed:

***k*-anonymity** meaning that, regarding the quasi-identifiers, each data item within a given data set cannot be distinguished from at least $k-1$ other data items [66].

***l*-diversity** meaning that for each group of records sharing a combination of quasi-identifiers, there are at least l “well represented” values for each sensitive attribute [67]. l-diversity implies l-anonymity.

***t*-closeness** meaning that for each group of records sharing a combination of quasi-identifiers, the distance between the distribution of a sensitive attribute in the group and the distribution of the attribute in the whole data set is no more than a threshold t [63].

δ-presence which basically models the disclosed dataset as a subset of larger dataset that represents the attacker’s background knowledge. A dataset is called $(\delta_{\text{min}}, \delta_{\text{max}})$-present if the probability that an individual from the global dataset is contained in the disclosed subset lies between δ_{min} and δ_{max} [65].

Different variants of l-diversity have been proposed, such as entropy-l-diversity and recursive-(c, l)-diversity, which implement different measures of diversity. It was shown that recursive-(c, l)-diversity delivers the best trade-off between data quality and privacy [67]. Different

ADOPT BBMRI-ERIC is funded by the European Union (EU)
Horizon 2020 under Grant number 676550.
Horizon 2020

variants exist also for t-closeness, e.g., equal-distance-t-closeness, which considers all values to be equally distant from each other, and hierarchical-distance-t-closeness, which utilizes generalization hierarchies to determine the distance between data items [63].

Both k-anonymity and l-diversity mitigate identity disclosure, while l-diversity additionally counters attribute disclosure. t-closeness is an alternative for protecting against attribute disclosure, while δ-presence mitigates membership disclosure. Regarding the LINDDUN threats, k-anonymity and l-diversity mitigate identifiability and linkability threats according to [4].

An open source tool that implements all of the anonymization metrics described above is the ARX toolkit and software library.[36]

Another anonymization method called Query-Set-Size Control can be used in order to dynamically answer statistical queries in a privacy preserving manner. The basic functional principle of this method is to return answers only if the number of entities contributing to the query result exceeds a given value k [68]. While it has been shown that this measure can be defeated by trackers [69], the susceptibility to tracker attacks can be prevented by only allowing predefined/restricted queries to be issued.

For the future, we recommend to investigate further approaches to anonymization, e.g., perturbation, which basically means the insertion of noise into microdata that is to be released [70].

Practical Recommendation for Anonymization There is no universal rule that applies to all the cases. Authors of guidelines for sharing clinical trials data [71] have performed an extensive survey of literature and existing guidelines, what is considered anonymous data based on the minimum cell size, which is equivalent to k for k-anonymity on the level of individual cells of source data [71, Appendix B, page 187]. Most commonly used value is 5, which means risk of re-identifying the data of $\frac{1}{5} = 20\,\%$. Some custodians use smaller values down to 3 [72–76], while others require larger values of 11 (in USA [77–80]) to 20 (in Canada [81, 82]). The maximum found in the literature was 25 [81]. Obviously the higher the k, the more suppression occurs or the more generalization is required.

A.5.2. Pseudonymization

Compared with anonymization as described in appendix A.5.1, pseudonymization also mitigates the LINDDUN threat types identifiability and linkability according to [4]. However, unlike anonymization, it does not remove the association between the identifying data set and the data subject, but rather replaces it with an association to one or more pseudonyms that usually enable only a restricted audience to re-identify the respective data subject. Typically, the possibility to re-identify subjects of pseudonymized data is restricted to members of the organizational entity that shared the pseudonymized data.

[36] arx.deidentifier.org/

Pseudonymization is required whenever the re-identification of data subjects from whom data has been shared might be necessary, for example in the case that research leads to new scientific findings the data subject requested to be informed about, or in case the data subject wants to withdraw or modify informed consent regarding data sharing.

Pseudonymization of data may be conducted by a data provider using encryption of identifiers before the data is sent to a particular consumer with a consumer specific secret key that was created ahead of time. This measure mitigates privacy threats arising from the linking of data sets that were sent to different data consumers because the same records have different identifiers in different data sets. Furthermore, the consumer specific identifiers could allow for the identification data leaks.

A.6. Accounting, Auditing, Provenance

Accounting and audit trails. Accountability is one of the key aspects of every infrastructure dealing with human biological material or data sets. Accounting means that actions of users should be recorded in the audit trails (logs), and these audit trails should be stored for a long time in order to be able to reconstruct flow of events in case of any investigation.

A common approach to this is distributed logging, that uses secure loggers, which are typically single-purpose computers with high physical security and software security and strong integrity measures. They provide unidirectional "sink interface" for other entities of the distributed system used to log events. Availability aspect is also very important in such setups, in oder to make them resistant to denial of service attacks.

Provenance. The goal of provenance is to provide consistent and complete information about history of both physical objects (biological samples) and digital objects (data sets, images, etc.). This goes well beyond the security & privacy (accountability), as provenance is also needed for quality management and for repeatability and reproducibility of results achieved using samples, data, and services provided by BBMRI-ERIC.

Common approaches to provenance include Open Provenance Model (OPM) and PROV Data Model (PROV-DM), as discussed in the results from EHR4CR and TRANSFoRm in [83]. OPM is graph-based where edges describe relations and vertices describe entities: artifacts (specific fixed data with context), processes (data transformations), agents (execution controllers – humans or immutable software). PROV-DM builds on OPM and adds attributions and extends support for evolution of entities over the time.

A.7. Protection of Storage and Communication Channels

Protection of storage and communication covers several aspects:

ADOPT BBMRI-ERIC is funded by the European Union (EU)
Horizon 2020 under Grant number 676550.

Horizon 2020

Protection against communication eavesdropping and storage intrusion both of which rely on sufficient encryption.

For network communication because of performance reasons, this typically combines asymmetric cryptography and symmetric. Computationally demanding asymmetric cryptography is used for exchange of randomly generated keys for computationally less demanding symmetric cryptography, which is in turn used for high-throughput communication.

For storage applications, similar approach can be used, protecting a key for symmetric cryptography using asymmetric encryption. The storage may also use distributed encryption, where the resulting system of k nodes may be resilient up to m security-compromised nodes (without compromising security of data) as well as up to n of unavailable nodes (without compromising security). Such approach has been demonstrated previously by Hydra FS[37] and Charon FS.[38]

Protection against man-in-the-middle attacks requiring authentication of all the communicating parties. This is typically part of the secure network communication protocols, where certificates issued by well-established CAs are used for server authentication by the client, while password-based or certificate-based approach is used for client authentication by the server. The certificate-based approach for client authentication is still in practice limited because of limited access of users to certificates, and also because of more complicated operations for non-technical users (although it is required for LoA > 2).

Countermeasures against vulnerability exploitation which focus mostly on avoiding access of the users to all the unnecessary services. This includes deployment and maintenance of network firewalls as well as limiting both physical and remote access to the computational and storage systems.

Vulnerabilities of systems should be continuously monitored and systems should be updated for all relevant vulnerabilities. Systems should be also proactively tested against known vulnerabilities (using tools like Nessus[39] [84]).

Practical implementation needs to pay close attention to the state-of-the-art of the approaches and tools, as some previously accepted techniques may become obsolete or deprecated. An example of this may be the use of all versions of Secure Socket Layer (SSL) due to their inherent deficiencies [85], so that for reasonably secure communication the service providers are expected to have switched to TLS 1.1 or newer (TLS 1.0 is also considered deprecated[40] [86]).

[37] https://twiki.cern.ch/twiki/bin/view/EGEE/DMEDS

[38] https://github.com/biobankcloud/charon-chef

[39] http://www.nessus.org/

[40] https://forums.juniper.net/t5/Security-Now/NIST-Deprecates-TLS-1-0-for-Government-Use/ba-p/242052

ADOPT BBMRI-ERIC is funded by the European Union (EU)
Horizon 2020 under Grant number 676550.
Horizon 2020

A.8. Organizational Aspects of Security

ISO/IEC 27000 is a series of standards for information security management, aiming at implementing and operating an Information Security Management System (ISMS). The core part of the standard is ISO/IEC 27001 which provides the minimum requirements for an ISMS, including a reference catalog of more than a hundred physical, technical and organizational information security controls that have to be implemented (if no exclusions apply) by any organization striving for compliance against the standard.

ISO/IEC 27018 is a code of practice for controls to protect PII processed in public cloud computing services. It may be used in conjunction with the requirements and security controls provided by ISO/IEC 27001. That means, for example, that the core ISMS of a public cloud services provider will be established according to ISO/IEC 27001 with the mandatory security controls from this standard, and the extended and additional controls listed in ISO/IEC 27018 will be added to the scope of this ISMS.

The main controls focus on the following areas relevant for trusted PII processing (the list not exhaustive):

- *contractually defined purpose of data processing:* data may be only used for the purposes defined in the contract between the service provider and consumer (i.e., service provider may not use them for any other purposes, such as data mining or advertisin, unless allowed in the contract);

- *provable removal of data:* removal of temporary files after processing, as well as provable removal of data after termination of the contract; additionaly, there are requirements on data encryption, restrictions on making hardcopy material, and on availability of tools to show the data distribution in the cloud infrastructure for the customer;

- *incident handling & transparency:* including notification of customer about any relevant security incidents, recoding to whom the data has been disclosed.

A.9. Other Terminology

The key words “MUST”, “MUST NOT”, “REQUIRED”, “SHOULD”, “SHOULD NOT”, “RECOMMENDED”, “MAY”, and “OPTIONAL” in all further sections of this document are to be interpreted as described in RFC 2119 [87]. “SHALL” and “SHALL NOT” will not be used as reserved words in this document for the sake of simplicity.

As common in IGTF documents,[41] if a “SHOULD” or “SHOULD NOT” is not followed, the reasoning for this exception must be explained to relevant accrediting bodies to make an informed

[41] `https://www.igtf.net/`

ADOPT BBMRI-ERIC is funded by the European Union (EU)
Horizon 2020 under Grant number 676550.

Horizon 2020

decision about accepting the exception, or the applicant must demonstrate to the accrediting bodies that an equivalent or better solution is in place.

Individual-level data is data about individual persons (participants = patients + donors) contributing their data and biological material for biobanks.

Sample-level data is data related to the individual samples stored in the biobanks.

ADOPT BBMRI-ERIC is funded by the European Union (EU)
Horizon 2020 under Grant number 676550.

Horizon 2020

B. General Requirements

Privacy and security requirements represent the current state of understanding of what are recommended approaches to mitigate risks inherent to processing human and medical data. These requirements must be reviewed and updated as state of the art evolves. They can be both strenghened if demonstrated insufficient, but can be also relaxed if less strict approach is proven (or becomes generally accepted) as sufficient. An initial set of requirements has been published as a part of EGI-Engage Milestone M6.2 document[42] and then continuously refined as an appendix of this architecture document.

When implementing these requirements, the risks should be evaluated specifically for every case and requirements adjusted accordingly.

B.1. Requirements on Personal Information Protection

Because of the particular importance of protection of personal information for BBMRI-ERIC, this section summarized general requirements:

Req-1 Unless exempted by requirement Req-2, any directly identifying data SHOULD stay at the originating institutions (formally defined as "data owners" by data protection regulations), which MUST implement either rule-based access control, or RBAC, or committee-based access control.

Req-2 It is only allowed to transfer data outside of a custodian's infrastructure, the data recipient ("processor") MUST assure at least the same level of data protection.

Req-3 Persons entitled to data access MUST NOT attempt to re-identify the person or otherwise counteract the de-identification of data. This SHOULD be covered by data access conditions if data is accessed locally in the biobank (requirement Req-1), or by DTA or MTA if data is transferred to recipient (requirement Req-2).

Req-4 For the data to be considered **(de facto) anonymized data** in BBMRI-ERIC infrastructure, the data MUST be at least k-anonymized, SHOULD be set to $k \geq 5$, and all the parameters SHOULD be considered quasi-identifiers.

It is of a particular note here that data custodians/owners may increase the k and/or apply other technical protection measures (see appendix A.5.1) if their national ethical and legal environment demands so or if they perceive the residual risks unacceptable.

$k \geq 5$ has been selected as the minimum commonly acceptable value based on literature survey discussed in appendix A.5.1, so that we don't impose unnecessary data suppression and generalization where not necessary. If data needs to be protected also against attribute disclosure when correlated with additional knowledge available from elsewhere, the k-anonymity is insufficient and additional measures (such as l-diversity, t-closeness, or δ-presence discussed in appendix A.5.1) need to be considered.

[42] https://documents.egi.eu/document/2677

ADOPT BBMRI-ERIC is funded by the European Union (EU)
Horizon 2020 under Grant number 676550.
Horizon 2020

Req-5 High security restricted access (see page 63) *(a)* MUST incorporate LoA ≥ 2 for both identity verification and authentication instance, *(b)* MUST include support for access control based on persons affiliated to projects, and *(c)* MUST include assessment of compliance of the projects with informed consent.

Req-6 The following table summarizes minimum requirements for different types of privacy-sensitive data

Table 9: Minimum requirements for basic data types. Non-personal data is used to denote data that does not contain any traces of privacy-sensitive data (e.g., data about operation of the biobank storage systems).

	directly identifying data (DT-1a)	coded data (DT-1b)	(de facto) anonymized data (DT-2)	non-human data (DT-5)
Authentication and authorization				
Identity verification	LoA ≥ 2	LoA ≥ 2	LoA ≥ 0	open
Authentication instance	LoA ≥ 3	LoA ≥ 2	LoA ≥ 0	open
Assessing project & informed consent compliance	not available for research	MANDATORY	RECOMMENDED	–
Restricted access	high security	high security	medium-low security	open
DTA/MTA	REQUIRED	REQUIRED	RECOMMENDED	open
Authentication and authorization				
Access log archive since last access	≥ 10 years	≥ 10 years	≥ 3 years	–
Data transfers and storage				
Encrypted storage Encrypted transfers	REQUIRED REQUIRED	REQUIRED REQUIRED		

Req-7 The BBMRI-ERIC policies MUST be compatible with GÉANT Data Protection Code of Conduct[43] [25].

B.2. Requirements on Accountability and Archiving

Req-8 Acceptation of a DTA or a MTA MUST be stored in non-repudiable way by both parties of the agreement. The document MUST contain agreed starting date and lifespan of the contract.

Possible implementation is PDF documents signed electronically by both parties using visible signature stamp, so that it can be also printed for archival purposes.

Req-9 Release process of any samples or any data containing person-level information (i.e., including (de facto) anonymized data and pseudonymized data and coded data) MUST be documented in non-repudiable way by the biobank.

Req-10 Link MUST be maintained between the DTA/MTA and the samples and data sent to the requesting party.

[43] `http://www.geant.net/uri/dataprotection-code-of-conduct/Pages/default.aspx`

ADOPT BBMRI-ERIC is funded by the European Union (EU)
Horizon 2020 under Grant number 676550.

Horizon 2020

Req-11 Access logs to any data that involves information on the level of individuals (e.g., sample-level data including (de facto) anonymized data) MUST be kept for minimum of 3 years.

Note that this is a minimum which may be increased for specific cases, such as requirement Req-12.

Req-12 Access logs to any directly identifying data or coded data MUST be kept at least for the same time as medical records in the following countries: the country of the participant (donor or patient), country of the data custodian, country of the data processing institution. RECOMMENDED minimum value is 10 years. Access logs MUST be kept for each BBMRI-ERIC Identity at least on the level of (a) date/time of beginning of access (signing DTA/MTA), (b) last date/time of access.

10 years recommended threshold has been selected as the minimum commonly found in the medical records retention, so that we don't impose unnecessary data suppression and generalization where not necessary. This is based on the following findings:

- 10 years since the last record in the patient care journal in Sweden,[44]
- 10 years for images in Italy and "forever" for clinical records (since the latter are considered legal documents)[45]
- 10 years in Norway by default, with some specific cases extended up to 60 years (such as exposure to carcinogens),
- 5 years of ambulant care, 10–40 years for various types of common care, 100 years for specific records (infectious diseases, mental disorders) in the Czech Republic,[46]
- 15 year in Netherlands,
- 10 years in a private medical center for personal medical record, 20 years in a public medical center for personal medical record, except if the patient is dead, 10 years after the death or 10 years after the last examination in the hospital in France,
- 25 years in United Kingdom,[47]
- 30 year in Germany.[48]

It is of a particular note here that national nodes may increase this threshold if their national ethical and legal environment implies so.

B.3. Requirements of Protection of Users Privacy

Req-13 BBMRI-ERIC MUST NOT use tracking of users[49] beyond auditing, understanding user's behavior and individual optimize services, and providing information about the impact of BBMRI-ERIC infrastructure. BBMRI-ERIC policy which describes the user tracking

[44] https://www.socialstyrelsen.se/fragorochsvar/patientjournaler (available in Swedish)

[45] Regulation Min.San.Dg.Osp./Div.III/n.900.2/AG./464/280 19.12.86, see also Regulation DL179/2012/a.13/c.5, http://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:decreto.legge:2012;179~art13-com5 (available in Italian). See http://www.slideshare.net/DigitalLaw/la-cartella-clinica-elettronica-lisi (available in Italian) for a discussion.

[46] Regulation 98/2012, https://www.zakonyprolidi.cz/cs/2012-98 (available in Czech).

[47] http://www.nhs.uk/chq/Pages/1889.aspx?CategoryID=68

[48] http://www.kvhb.de/aufbewahrungsfristen (available only in German)

[49] Following users both in individual services and across different IT services, see, e.g., [88–93] for more discussion of various techniques.

ADOPT BBMRI-ERIC is funded by the European Union (EU)
Horizon 2020 under Grant number 676550.
Horizon 2020

MUST be publicly available and MUST be written in simple terms understandable also for non-technical users.

Req-14 Whenever requested by regulations, the user MUST be clearly notified that tracking is in place and consent with the this policy. If the user does not provide consent with the tracking policy, he MUST be notified that those services will not be available to him/her.

Req-15 While BBMRI-ERIC MAY use external services to analyze user behavior, use of these services MUST NOT include those services dealing with privacy-sensitive data from biobanks. Users MUST be clearly notified about use of such external services.

This allows cautious use of third party tools such as Google Analytics for analysis of web-based applications, as BBMRI-ERIC will not have capacity to develop/operate such services in-house.

Req-16 The data coming from user tracking MUST be treated as confidential by BBMRI-ERIC.

Corollary: This does not say—on purpose—that the data must be collected inside of BBMRI-ERIC infrastructure, as this would rule out Google Analytics and similar services. But once the data is transferred to BBMRI-ERIC, it MUST NOT be published outside.

B.4. Requirements on Data Storage, Transfers, and Computer Networks

Req-17 Directly identifying data and coded data SHOULD be stored encrypted with state-of-the-art encryption strength appropriate to the sensitivity of the data.

See appendix A.7 for brief discussion of available technologies.

Req-18 Computer networks used for processing directly identifying data and coded data SHOULD use traffic filtering to lower risks of attacks from outside. Devices connected to the computer networks SHOULD be protected on their own (i.e., end-device security) in order to minimize damage when an attacker makes it into the protected network perimeters.

Req-19 Secure network protocols MUST be used when transferring privacy-sensitive data (directly identifying data and coded data) over the network. For (de facto) anonymized data it is RECOMMENDED.

See appendix A.7 for brief discussion of the state of the art, deprecation of SSL, etc.

B.5. Requirements on Software Design and Development

Req-20 All software developed within BBMRI-ERIC MUST have clearly defined license.

This requirement is also a prerequisite or at least a facilitating element for other subsequent requirements.

Req-21 Software developed within BBMRI-ERIC SHOULD use open-source license of either BSD/Apache/MIT style or LGPL/GPL style.

Choice of particular license needs to consider preferences of the development teams, dependency on other software, as well as external requirements (e.g., if software is developed as a part of broader collaboration in externally funded projects).

Req-22 Software developed within BBMRI-ERIC SHOULD undergo peer-review of the design as well as of the implementation. The peer-review SHOULD involve individuals or teams external to the development team of the given software (at least another development group in the BBMRI-ERIC CS IT).

Req-23 Choice of programming language and third-party libraries and frameworks for the development SHOULD consider security aspects and SHOULD facilitate requirements Req-21 and Req-22.

Req-24 Software development SHOULD use available static code analysis tools (and security-oriented analysis tools in particular) such as Coverity Scan.[50]

Use of such tools is facilitated by the open-source requirement Req-21 and choice of programming language and various frameworks requirement Req-23.

Req-25 Software developed within BBMRI-ERIC dealing with user's input MUST implement sufficient validation of the input, including prevention of code injection and prevention of cross-site scripting whenever appropriate.

Req-26 Software developed within BBMRI-ERIC is RECOMMENDED to use publicly available code repositories with version management, such as SourceForge[51] or GitHub.[52]

It is allowed to use also publicly available repositories maintained by the development teams.

Req-27 Software developed within BBMRI-ERIC SHOULD support versioning as a part of the configuration management.

Req-28 Software not developed within BBMRI-ERIC but integrated into the BBMRI-ERIC services is RECOMMENDED to adhere to the same principles as software developed within BBMRI-ERIC.

[50] `https://scan.coverity.com/`, as of writing available for free for analysis of open-source software.

[51] `https://sf.net`

[52] `https://github.com/`

C. Requirements on Use Cases

C.1. S+UCs-1: Biobank browsing/lookup

This use case typically does not deal with the privacy-sensitive information, because of the highly aggregated metadata. When generating the metadata, and particularly for small collections where natural sparseness combined with increasing dimensionality of the data can introduce privacy issues because of "dimensionality curse" [29], we require that the data must adhere to the anonymity guidelines.

Req-29 When extracting metadata about sample/data collections from the biobanks, the metadata generator MUST ensure the data is anonymized to the level of being considered *(de facto) anonymized data*: see requirement Req-4 on page 74.

C.2. S+UCs-{2,3}: Sample/Data Negotiator

Req-30 Sample/Data Negotiator MUST require user to sign MTA or DTA before positively concluding negotiation of access to samples or data respectively.

Req-31 Sample/Data Negotiator MUST require that all the sample/data requests are done with a user affiliated to a project. *This does not apply for sample reservations, see requirement Req-32.*

Req-32 As a part of the Sample/Data Negotiator workflow, compliance of project (or project proposal for reservations) with informed consent for samples/data MUST be evaluated, before enable requester access to the data or samples.

Req-33 Sample/Data Negotiator MUST require biobankers to consent with treating all the sample/data requests as well as reservations as confidential.

C.3. S+UCs-{5,6}: Sample Locator

Req-34 Sample Locator MUST also fulfill requirements of the Sample/Data Negotiator (appendix C.2).

Req-35 Users MUST require users to consent to the terms and conditions, including refraining from any person re-identification attempts, before using Sample Locator.

Req-36 Sample Locator MUST require user to sign MTA or DTA before positively concluding negotiation of access to samples or data respectively.

ADOPT BBMRI-ERIC is funded by the European Union (EU)
Horizon 2020 under Grant number 676550.
Horizon 2020

C.4. S+UCs-14: Data Processing

General requirements apply for this use case, and particular attention should be paid to requirements Req-2 and Req-6.

Req-37 Any third party computing and storage infrastructures (particularly cloud infrastructures) considered for offloading storage and computing applications MUST be risk-analyzed and results of this analysis must be stored for future reviews.

Req-38 Any third party computing/storage infrastructure used for processing and storing the data MUST provide sufficient liability.

Req-39 Physical computing resources used for processing privacy sensitive data (at least directly identifying data or coded data) SHOULD NOT be used for other simultaneous applications with lower risk level.

This requirement is particularly focused on minimizing risk of attacks, where an attacker gains access to the virtual machines on the same physical host or even to the host of the virtual machines to attack the virtual machines used for processing of privacy-sensitive data. Note that the requirement uses "SHOULD NOT" semantics, i.e., exception can be provided if the operator, e.g., Infrastructure as a Service (IaaS) provider, is able demonstrate the same or better level of security as if dedicated hardware infrastructure is used.[53]

C.5. Organization Security

Req-40 The security measures SHOULD be clearly documented as a part of the organizational measures on the institutional level (e.g., level of the biobank).

[53] This requirement is formulated as generic at the moment. Solutions using private/public cloud providers together with security-related certifications will be explored as a part of BBMRI-ERIC activities, e.g., in EGI-Engage and PhenoMeNal projects, also related to legal requirements and liability aspects.

ADOPT BBMRI-ERIC is funded by the European Union (EU)
Horizon 2020 under Grant number 676550.

Horizon 2020