EGI-Engage

Security and privacy requirements and secure storage architecture

M6.2

Date	01 March 2016
Activity	SA2
Lead Partner	BBMRI-ERIC
Document Status	FINAL
Document Link	https://documents.egi.eu/document/2677

Abstract

Privacy and security are fundamental concepts that must be built into BBMRI-ERIC IT services by design, as trust and transparency are the key element of each medical research infrastructure. This document focuses on providing a comprehensive list of requirements for implementing IT services of BBMRI-ERIC as well as for interacting with other infrastructures which will provide services to BBMRI-ERIC. It also provides risk analysis of the most important services and pays particular attention to authentication and authorization, as these are supposed to be built jointly with other infrastructures. Last but not least, it summarizes cloud-based architecture for processing of privacy-sensitive data related to biobanking and architecture for their secure storage. This document is written by the BBMRI Competence Centre of the EGI-ENGAGE project, building upon internal document of BBMRI-ERIC on “Security and Privacy Requirements”.

The EGI-Engage project is co-funded by the European Union (EU) Horizon 2020 program under Grant number 654142 http://go.egi.eu/eng

COPYRIGHT NOTICE

DELIVERY SLIP

	Name	***Partner/Activity***	***Date***
From:	Petr Holub, Jim Dowling, Salman Muhammad Khan Niazi, Kamal Hakimzadeh, Boris Parák	BBMRI Competence Centre	2016-02-26
Moderated by:	Małgorzata Krakowian	EGI.eu/NA1	2016-02-26
Reviewed by	Florian Kohlmayer Raffael Bild Michael Hummel Gianluigi Zanetti Philip Quinlan Alexandre Bonvin	TUM-MED TUM-MED CHARITÉ CRS4/BBMRI.it University of Nottingham/BBMRI.uk Universiteit Utrecht/PMB	2015-12-10,2016-02-01 2015-12-10,2016-02-01 2016-02-15 2016-02-19 2016-02-25 2016-02-17
Approved by:	AMB and PMB		2016-03-01

DOCUMENT LOG

Issue	***Date***	***Comment***	***Author/Partner***
v.1	2016-02-16	First version of the document for review. Petr Holub contributed Sections 1-7 and contributed also to Section 8, Jim Dowling, Salman Muhammad Khan Niazi, and Kamal Hakimzadeh, contributed Sections 8-9, Boris Parák contributed to Section 8.	Petr Holub, Jim Dowling, Salman Muhammad Khan Niazi, Kamal Hakimzadeh, Boris Parák
v.2	2016-02-26	Version after external review	Petr Holub
FINAL	2016-02-29	Final version	Petr Holub

Contents

Executive Summary

Privacy and security are fundamental concepts that must be built into all BBMRI-ERIC IT services *by design*, as trust and transparency are the key elements of medical research infrastructures dealing with privacy-sensitive human data. Hence this document focuses on providing a comprehensive list of requirements for implementing IT services of BBMRI-ERIC, as well as to provide input for other infrastructures which will deliver services to BBMRI-ERIC. It refers to the design documents of individual BBMRI-ERIC services and only provides design description of the privacy and security related services that are "middleware" shared among multiple BBMRI-ERIC services, such as authentication and authorization infrastructure. Another important aspect of the document is description of the data storage architecture that is suitable for storing and retrieving privacy-sensitive data.

Section 2 provides an overview of the most important concepts in security & privacy related to BBMRI-ERIC, such as risk analysis and sensitivity of information and material, authentication and authorization, as well as privacy enhancing technologies such as pseudonymization and anonymization. This section is intended to harmonize initial knowledge among the readers of different backgrounds. It is based on observation that even experts in the specialized sub-domains of privacy and security persons lack sometimes up-to-date information about other parts of the field; readers fully familiar with privacy and security can skip this section. Section 3 provides high-level architectural and functional description of BBMRI-ERIC IT services, organization of the data, employed data formats and standards, as well as APIs of services (where already defined). Section 4 models the most imminent use cases for BBMRI-ERIC IT services using Data Flow Diagrams, in order to help analysis of risks and threats using STRIDE and LINDDUN methodologies. Such analysis forms foundation for defining requirements, as these are intended to set minimum standards for minimization of risks related to processing privacy-sensitive data in the workflows specific for BBMRI-ERIC.

Actual requirements start with general security and privacy requirements in Section 5 and use-case specific requirements in Section 6. Particular attention is paid to requirements on AAI in Section 7, which is intended as an input for the AAI services provided by eInfrastructures (such as GÉANT) and government-backed identity providers (such as successor of STORK). Architecture for BBMRI-ERIC AAI is drafted in Section 8, which is understood as an interim solution before the services with the required extent of functionality and dependability are provided by eInfrastructures and government-backed identity providers.

Overview of architecture of the cloud-related processing of sensitive data for biobanks is described in Section 9, with primary focus on enabling private clouds in biobanks using EGI Federated Cloud and BiobankCloud technologies. This is understood as a first step, where EGI technologies will be used for building private clouds, typically inside the biobanks, to support scalable processing of the privacy-sensitive data. Scaling outside of the private clouds to third-party cloud providers will be explored later during implementation of the BBMRI Competence Center.

Similarly for storage architecture, Section 10 describes the basic secure storage model for the biobanking data and their interaction with cloud architectures. Further options will be explored later with particular

attention to the sensitivity of data and also changing regulatory frameworks,[1] which are expected to have profound impact on this field.

[1] At the time of writing this document, new General Data Protection Regulation (GDPR) regulation has been approved in the trilogue and sent for the legislative process into the European Parliament, see Section 2.3.7 for more details.

Glossary

AAI	Authentication and Authorization Infrastructure. 4, 6, 13, 24, 59, 61, 63–65, 67
AARC	Authentication and Authorisation for Research and Collaboration. See `https://aarc-project.eu/` and GÉANT Association (GÉANT), 25, 67
AC	(Data|Samples) Access Committee. 27
anonymous data	Anonymous data is such data, that is is no longer identifiable. 30, 38
Apache jclouds®	Apache jclouds® is an open source multi-cloud toolkit for the Java platform, see `https://jclouds.apache.org/`. 68, 71, 74
BBMRI Competence Center	BBMRI Competence Center is a part of WP6 of EGI-Engage project.. 3, 6, 12, 13, 68
BIMS	Biobank Information Management System. 43
CA	Certification Authority. 35
Common Service	A formal way of organizing full member countries of BBMRI-ERIC to provide services of common interest. 8
CS ELSI	Common Service ELSI. See Common Service and ELSI, 18, *see* ELSI
CS IT	Common Service IT. See Common Service
DAC	Discretionary Access Control. 3, 28, 60, 65
deidentified data	Data in which identifiers have been removed or replaced, such as in case of anonymized or pseudonymized data. See Section 2.5 for more detailed discussion.. 30
DFD	Data Flow Diagram. [1], 3, 14, 16, 43
DoS	Denial of Service. 15
DS	Discovery Service. See Shibboleth, 19, 20, 24
DTA	Data Transfer Agreement. 3, 18, 38, 47, 52, 53, 57
eduID	Research and educational identity federations, represented by national federations such as eduID.se, eduID.hu, eduID.cz, etc.. 20, 23
EGI	`http://www.egi.eu/`. 4, 12, 25, 67–69, 71, 74
EGI-Engage	EGI-Engage project. `https://www.egi.eu/about/egi-engage/`, 3, 8, 12, 58, 67, 68, 70
ELSI	Ethical, Legal, and Social Issues. 8
EoP	Elevation of Privilege. 15
EU	European Union. 26
FIPS	Federal Information Processing Standard. 23

GDPR	General Data Protection Regulation. 7, 13, 26, 31
GÉANT	GÉANT Association. http://www.geant.net/, 8, 10, 20, 25, 53, 67, 70
HTTP	Hypertext Transfer Protocol. 21
IdP	Identity Provider. See Shibboleth, 19–22, 24, 25, 28, 29, 62, 64, 65, 67
IoI	Item of Interest. 15
ISMS	Information Security Management System. 35
LINDDUN	Linkability, Identifiability, Non-repudiation, Detectability, Disclosure of information, Content Unawareness, Policy and consent non-compliance. [2], 4, 6, 13–16, 27, 32, 33, 43, 49, 50
LoA	Level of Assurance. 3, 20–24, 35, 52, 53, 59–63
MAC	Mandatory Access Control. 3, 27, 28, 60, 65
MOSLER	Secure platform for processing sensitive data. See https://bils.se/resources/mosler.html. 29, 48, *see* TSD
MTA	Material Transfer Agreement. 3, 18, 38, 47, 52, 53, 57
non-deidentified data	Data which has not been deidentified, e.g., raw patient records. See Section 2.5. 30, 38, 52, 53, 55, 58
ODbL	Open Data Commons Open Database License. http://opendatacommons.org/licenses/odbl/, 17
OpenID	standard decentralized protocol for authentication with substantial support in commercial environments. See http://openid.net/, 20, 23
OPM	Open Provenance Model. http://openprovenance.org/, 34
PDP	Policy Decision Point. 65
PEP	Policy Enforcement Point. 65
Perun	Virtual group management system with support for virtual identity consolidation [3]. 24, 25
PET	Privacy-Enhancing Technologies. 3, 30
practically anonymous	Data which has been processed to the level that they can be considered anonymous for practical purposes. See Requirement Req-3. 38, 43, 45, 47, 52, 53, 55, 57
PROV-DM	PROV Data Model. http://www.w3.org/TR/prov-dm/, 34
pseudonymous data	Pseudonymous data is such data for which identifiers of persons have been replaced by a pseudonym (code) [4]. 30, 38
RBAC	Role-Based Access Control. 3, 25, 27–29, 52, 60, 65, 66
REMS	Resource Entitlement Management System. http://www.csc.fi/rems and [5], 28

SAML V2.0	Security Assertion Markup Language, Version 2.0. See https://www.oasis-open.org/committees/security/, 19, 23
Shibboleth	Federated identity system [6, 7], https://shibboleth.net/. 4, 8–10, 19, 25
SP	Service Provider. See Shibboleth, 4, 19, 20, 24, 25, 28, 65
SSL	Secure Socket Layer. 35
SSO	Single Sign On. 20
STORK	Secure idenTity acrOss boRders linked. https://www.eid-stork.eu/, 24
STORK 2.0	Secure idenTity acrOss boRders linked 2.0. https://www.eid-stork2.eu/, 24
STRIDE	Spoofing, Tampering, Repudiation, Information Disclosure, Denial of service, Elevation of privilege. [1], 4, 6, 13, 14, 16, 27, 43, 49
TLS	Transport Level Security. 35
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/. 29, 48
UI	user interface. 70, 74
VOPaaS	VO Platform as a Service provided by GÉANT. GÉANT and [8, 9], 25, 67
WAYF	Where Are You From service. See Shibboleth, 19, 20, 24, 25

1 Introduction

1.1 Biobanks and BBMRI-ERIC

Biobanks have become a major source of biosamples as well as data for the biomedical and bioinformatics research. Biobanks are used by the researcher not only to request samples and data, but also to provide the researchers with long-term sample and data repositories for material used in their research. Data collection, harmonization and processing has been part of biobanks since their inception, as biosamples without the data is of little use. The data collection started with the phenotype, clinical, and lifestyle data (with focus on specific data types given by the type of the biobanks, such as population biobanks or clinical biobanks). Unprecedented growth of omics data generation in recent 15 years have brought biobanks into the domain of big data, processing and storing genomics, proteomics, metabolomics and other types of data.

After about ten years of preparations, BBMRI-ERIC has become one of the first European Research Infrastructure Consortia, with the mission of providing high-quality samples, data, and biomolecular resources from biobanks to support healthcare advancement in Europe and beyond. The major goals of BBMRI-ERIC are:

- to *increase use of material and data* stored in European biobanks, while adhering to strong *privacy protection* of patients and donors contributing the material and data,
- to *improve quality and traceability* of the material and data in European biobanks, referring to the infamous recent publications demonstrating that large portions of biomedical research are not reproducible [10, 11, 12, 13, 14] and this has been even demonstrated specifically for the process of generating data from samples [15],
- to *improve data harmonization* and contribute to the standardization processes,
- to *contribute to the ethical, legal, and social issues*, with particular focus on cross-border exchanges of human biological resources and data attached for research use.

Although biomedical and bioinformatics researchers (coming from both academia and industry) as well as biobankers are mostly seen as the primary users of BBMRI-ERIC. Other users and stakeholders are also embraced and supported, such as research participants (= patients/donors) and their organizations, data protection agencies and research funding agencies are also part of the target users. Furthermore, even for the researchers, the use cases go beyond well-known sample/data request use case: recent investigations by BBMRI.uk[2] have shown that sample/data storage and curation requests may be as frequent, and industry is specifically known for joint prospective studies with biobanks instead of requesting existing samples.[3]

The IT infrastructure of BBMRI-ERIC will be developed and operated using the Common Service IT instrument, to which all the full-member countries of BBMRI-ERIC contribute. It follows up on experience from

[2] Results have not been published yet.

[3] The reasons for this range from the informed consent signed by the research participants to tighter control over the sample collection/processing/storage requirements.

the BBMRI Preparatory Phase[4] as well as collaboration within other projects in the BBMRI ecosystem, such as BBMRI-LPC,[5] BioSHaRE,[6] BioMedBridges,[7] or BiobankCloud.[8]

1.2 BBMRI Competence Center in EGI-Engage

Based on the specifics of large-scale privacy-sensitive data, EGI-Engage project has proposed to use BBMRI-ERIC as one of the pilot applications to focus on when evolving EGI services. This led to setting up BBMRI Competence Center as a part of WP6 (SA2) Knowledge Commons of EGI-Engage.

BBMRI Competence Center focuses on the following main tasks:

- defining security and privacy requirements on the BBMRI-ERIC services, with particular focus on computing and storage of biobank data (handled by this Milestone),
- defining storage architecture of the storage of privacy-sensitive data processed as a part of the biobanking workflows (also covered in part by this Milestone),
- implement scalable processing of privacy-sensitive data (with genomic data taken as an example) using EGI Federated Cloud platform,
- showcase a pilot deployment of the integrated system,
- disseminate information about achieved results.

The BBMRI Competence Center is expected to build upon the above mentioned BiobankCloud platform and services and technologies of EUDAT to support local biobanks by connecting data resources in a federated cloud infrastructure in coordination with the ELIXIR cloud working group and BBMRI-ERIC Common Service IT.

1.3 How To Read This Document

This document was created as to demonstrate achievement of the EGI-Engage Milestone M6.2, but it has utilized synergy between needs of BBMRI Competence Center and the internal needs of BBMRI-ERIC to develop fundamental document describing privacy and security requirements on IT infrastrcture. While availability of the first version of Security & Privacy Requirements document constituted the milestone of BBMRI Competence Center, the document is expected to be further developed and updated over the time based on gained practical experiences, as well as development of regulatory framework and developments in the IT domains related to privacy and security.

[4] Material from BBMRI Preparatory Phase can be found at `http://bbmri-eric.eu/reports`
[5] `http://www.bbmri-lpc.org/`
[6] `https://www.bioshare.eu/`
[7] `http://www.biomedbridges.eu/`
[8] `http://www.biobankcloud.com/`

Section 2 provides an overview of the most important concepts in security & privacy related to BBMRI-ERIC, such as risk analysis and sensitivity of information and material, authentication and authorization, as well as privacy enhancing technologies such as pseudonymization and anonymization. This section is intended to harmonize initial knowledge among the readers of different backgrounds. It is based on observation that even experts in the specialized sub-domains of privacy and security persons lack sometimes up-to-date information about other parts of the field; readers fully familiar with privacy and security can skip this section. Section 3 provides high-level architectural and functional description of BBMRI-ERIC IT services, organization of the data, employed data formats and standards, as well as APIs of services (where already defined). Section 4 models the most imminent use cases for BBMRI-ERIC IT services using Data Flow Diagrams, in order to help analysis of risks and threats using STRIDE and LINDDUN methodologies. Such analysis forms foundation for defining requirements, as these are intended to set minimum standards for minimization of risks related to processing privacy-sensitive data in the workflows specific for BBMRI-ERIC.

Actual requirements start with general security and privacy requirements in Section 5 and use-case specific requirements in Section 6. Particular attention is paid to requirements on AAI in Section 7, which is intended as an input for the AAI services provided by eInfrastructures (such as GÉANT) and government-backed identity providers (such as successor of STORK). Architecture for BBMRI-ERIC AAI is drafted in Section 8, which is understood as an interim solution before the services with the required extent of functionality and dependability are provided by eInfrastructures and government-backed identity providers.

Overview of architecture of the cloud-related processing of sensitive data for biobanks is described in Section 9, with primary focus on enabling private clouds in biobanks using EGI Federated Cloud and BiobankCloud technologies. This is understood as a first step, where EGI technologies will be used for building private clouds, typically inside the biobanks, to support scalable processing of the privacy-sensitive data. Scaling outside of the private clouds to third-party cloud providers will be explored later during implementation of the BBMRI Competence Center.

Similarly for storage architecture, Section 10 describes the basic secure storage model for the biobanking data and their interaction with cloud architectures. Further options will be explored later with particular attention to the sensitivity of data and also changing regulatory frameworks,[9] which are expected to have profound impact on this field.

[9] At the time of writing this document, new General Data Protection Regulation (GDPR) regulation has been approved in the trilogue and sent for the legislative process into the European Parliament, see Section 2.3.7 for more details.

2 Relevant Security & Privacy Concepts

This section provides 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 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., [16] as well as literature referred to throughout this section).

Parts of this section, namely Sections 2.1, 2.2, and 2.5, use excerpts from Deliverable 5.3 [17] of 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 Section 2.5 for further discussion, including explicitly stated incompatibility with ISO 25237 [4], 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 Section 2.4.4 for further discussion).

2.1 Risk Analysis and Management

As proposed in BioMedBridges Deliverable 5.3 [17], we will use Data Flow Diagrams (DFDs) [18] 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, [17] 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. A sample DFD is shown in Figure 1.

Figure 1: Sample DFD with color coding proposed in [17]. This DFD is only intended as an example of entities without any real-world meaning.

The risks will be analyzed using Spoofing, Tampering, Repudiation, Information Disclosure, Denial of service, Elevation of privilege (STRIDE) [1] and Linkability, Identifiability, Non-repudiation, Detectability, Disclosure of information, Content Unawareness, Policy and consent non-compliance (LINDDUN) [2] methodologies. The STRIDE focuses on security threats, while LINDDUN focuses on privacy threats.

STRIDE [1] identifies the following security risks, connected to the imperiled security properties [19, 20]:

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 [19].

Tampering threats involve malicious modification of data or code. This threatens **integrity**, which is property of correctness and completeness of assets [19].

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 [19].

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 [19].

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 [19].

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 [20].

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 [2, 21].

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" [2]. Please note we are using slightly different definition of anonymity as discussed in the Section 2.5.

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 [2, 21].

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 [21].

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 [22].

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 [2].

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. [2].

Mapping of risks described by STRIDE and LINDDUN to the DFD entities is shown in Tables 2 and 3.

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 2: Mapping STRIDE security threats and countermeasures to data flow diagram element types (see Tables 9-5 and 9-8 in Chapter 9 of [1]).

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 3: Mapping LINDDUN privacy threats and objectives to DFD element types (see Tables 4 and 6 in [2])

.

The overall **risk level** is qualitatively assessed using **likelihood of a threat** and **level of impact** as shown Table 4.

2.2 Sensitivity of Information and Biological Material (Samples)

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

Table 4: Qualitative risk assessment.

2.2.1 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).[10]

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 Section 2.4.1.

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.).

2.2.2 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 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 not uncommon that results of the discussion have unintended impact into

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

biomedical research [23, 24, 25, 26, 27]. This field is the expertise of Common Service ELSI[11] of BBMRI-ERIC and any issues should be consulted with this body.

2.2.3 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 Data Transfer Agreements (DTAs), while biological material is covered by Material Transfer Agreements (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.

2.3 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 (Section 2.3.1), commonly used levels of assurance of persons physical and virtual identities (Section 2.3.2), problems of identity merging for persons possessing multiple virtual identities (Section 2.3.3), as well as aspects related to the robustness of the authentication systems (Section 2.3.4. Since authentication often provides additional means for authorization, we will discuss also attribute issuing as a part of the authentication (Section 2.3.5). Finally, we will conclude with references to the regulations that constitute legal framework to the authentication (Section 2.3.7).

2.3.1 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 assur-

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

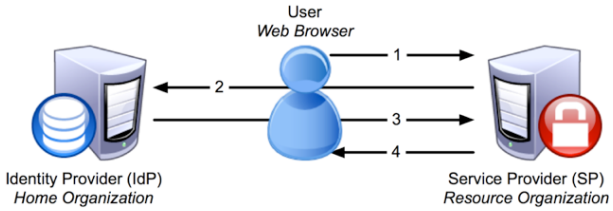

Figure 2: Simple interaction of an IdP and a SP (without WAYF/DS). The diagram starts with user accessing the Resource (1). See https://wiki.shibboleth.net/confluence/display/CONCEPT/Home for more details.
Source: https://wiki.shibboleth.net/confluence/download/attachments/4358538/sso-flow.png?version=2&modificationDate=1249311729063&api=v2

ance of user identity (see Section 2.3.2) 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 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) [28] respectively, but they are applicable more generally. IdP is the actual authentication service at an institution which verifies a person's virtual identity and SP is any service provided to the person that consumes the virtual identity and uses it for authorization purposes, as shown in Figure 2. 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 3 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).

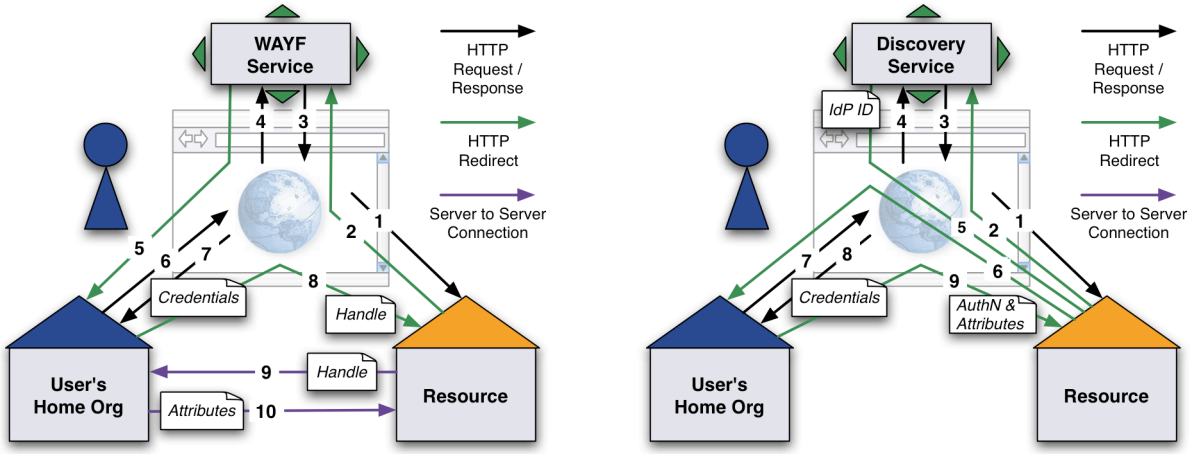

Figure 3: Interaction of an IdP (User's Home Org), a SP (Resource), and a WAYF or DS. The diagram starts with user accessing the Resource (1). See `https://www.switch.ch/aai/support/tools/wayf/` for more details.
Source: `https://www.switch.ch/aai/support/tools/wayf/wayf-vs-ds.png`

These systems are now becoming widely available in the various flavors: research and educational communities have successfully established identity federations such as eduID[12]; commercial companies having organized themselves in OpenID[13] or at least providing comparable interfaces such as Facebook Connect[14]; and there are pilot efforts of government-backed identity federations called STORK discussed in Section 2.3.2 on page 23.

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 Section 2.3.2 below). Another advantage of 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 into the need for some "catch-all" IdPs, which may be hard to implement at the same LoA as "normal" IdPs.

[12] 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.
[13] `http://openid.net/`
[14] `https://developers.facebook.com/blog/post/2008/05/09/announcing-facebook-connect/`, `https://developers.facebook.com/docs/facebook-login`

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 Section 2.3.5. 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 [29]. 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.

2.3.2 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 comes from NIST SP 800-63-2 [30], while a nice summary of implementation in practical federated authentication systems is available on the Tuakiri Federation website[15] and in [31].

There are two main aspects of level of assurance:

1. the strength of the process of *identity proofing and verification* (see [32, Article 8 and 9(1)]) of the person during registration of the user (we will use **identity verification** in the following text),
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 that no identity verification has been done at all, while still having a notion of "a user". This can be used, e.g., 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 pro-

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

vided 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.

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 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 stronger, and the subsequent critical data transfer processes must be authenticated using a key created as a part of the authentication process. The tokes 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[16] by The Interoperable Global Trust Federation (IGTF)[17]: 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 [33] 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 used also 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 is 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.[18]

Support for LoA is available in SAML V2.0, as a part of the Identity Assurance Profiles Version 1.0 [34]. They are also available in practical implementations like Shibboleth [35], 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 [36].

[16] `https://www.eugridpma.org/guidelines/loa/IGTF-LoA-authN-set-20150930-v11.docx`
[17] `https://www.igtf.net/`
[18] 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.

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)[19] and Secure idenTity acrOss boRders linked 2.0 (STORK 2.0)[20] 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 insufficiently strong authentication instance, it can be improved using alternate IdP together with identity linking (described in the Section 2.3.3 below).

2.3.3 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 [29], 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 [3], 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).

2.3.4 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 Section 2.3.5). 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 [37]. 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

[19] https://www.eid-stork.eu/
[20] https://www.eid-stork2.eu/

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) [8, 9], or ELIXIR, where the identities from the originating IdPs are cached by the 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).

2.3.5 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[21] and uApproveJP[22] [38].

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 of 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 Section 2.4.3.

[21] https://www.switch.ch/aai/support/tools/uapprove/
[22] https://meatwiki.nii.ac.jp/confluence/x/aQLO

2.3.6 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 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*.

2.3.7 Legal Requirements for Security & Privacy

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

- Directive on the protection of personal data 95/46/EC [39],
- Directive 1999/93/EC on a Community framework for electronic signatures [40],
- Directive 2006/123/EC on services in the internal market [41],
- Directive 2002/58/EC concerning the processing of personal data and the protection of privacy in the electronic communication sector [42].

Another part of the framework will be General Data Protection Regulation (GDPR), obsoleting 95/46/EC. Consensus has been reached[23] by between by the European Commission, Parliament, and Council (so-called 'trilogue' meetings) on December 15, 2015 and the General Data Protection Regulation (GDPR) has been submitted for approval process in Parliament. Consequences of GDPR are yet to be understood.

2.4 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 Section 2.4.1, including open access, restricted access and committee-controlled access.

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

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 Section 2.4.2 and role-based access control in Section 2.4.3.

2.4.1 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 Section 2.4.

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 4.2 on page 49 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 4.2 on page 49 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) [5].

2.4.2 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 [16].

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.

2.4.3 Role-Based Access Control (RBAC)

RBAC is an approach based on the roles that is 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 release from IdPs or attribute services as discussed in Section 2.3.5.

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[24]:

- 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`,

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

- 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`.

Another problem with pure attribute-based RBAC is delegation (see Section 2.3.6), 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[25] [43, 44], MOSLER[26] and TSD.[27]

2.4.4 Semantic development of committee-controlled access

Note that there is a subtle semantic shift since BioMedBridges Deliverable 5.3 [17] in how we work with committee-controlled access.

The Deliverable used the committee-controlled access as 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 Section 2.4.1. The committee-controlled access then remains orthogonal and can be combined with any restricted access mode.

[25] `http://www.biobankcloud.com/`
[26] `https://bils.se/resources/mosler.html`
[27] `https://www.norstore.no/services/TSD`

2.5 Privacy-Enhancing Technologies (PET)

Privacy-Enhancing Technologies (PET), defined, e.g., in ISO 29100 [45] and [21]), deal with problem of protecting privacy of individuals in information technologies and information systems. As a part of the PET, we introduce the following definitions:

Anonymous data is a data in that *attacker cannot sufficiently identify the subject within a set of subjects, the anonymity set* [2, 21].

This is both practical and mathematically sound definition. Alternatively, there is a simpler common-sense definition: *data that is is no longer identifiable*. This simpler definition comes from Directive on the protection of personal data 95/46/EC [39] and can be seen as intuitively equivalent, but it lacks the rigor of working with the anonymity set.

Anonymization is a transformation which makes the data anonymous.

Anonymization of data can be performed dynamically as a data release preparation, or data can already be anonymized before persisting it.

Pseudonymous data is such data for which identifiers of persons have been replaced by a pseudonym (code) [4].

Note that pseudonymous data is *not* a subset of anonymous data, as the pseudonymous data is not anonymous: there is even no notion of anonymity set. The data is still uniquely identifying, albeit linking (or translation) might be known only to some trusted subject. This is consistent with [46, 21].

Pseudonymization is a transformation which makes the data pseudonymous by both removing the association with a data subject and adding an association between a particular set of characteristics relating to the data subject and one or more pseudonyms [4].

Deidentified data is data, for which identifiers have been removed or replaced.

This term can be used for denoting anonymous data or pseudonymous data, and we will use it in this document to cover both.

Non-deidentified data is complement to deidentified data; i.e., it is data, for which identifiers have not been removed.

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

It is worth mentioning there is disagreement among different authors regarding PET terminology. Namely ISO 25237 [4] 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 [47, 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[28]:

> 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.

2.5.1 Anonymization

As described in [48] and [49], 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).

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 [48].
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 [48].
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 [50].

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

According to [48], 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 [51].

***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 [52]. 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 [48].

***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 $(\delta_{\min}, \delta_{\max})$-present if the probability that an individual from the global dataset is contained in the disclosed subset lies between $\delta_{\min}$ and $\delta_{\max}$ [50].

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 [52]. Different 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 [48].

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 [2].

An open source tool that implements all of the anonymization metrics described above is the ARX toolkit and software library.[29]

[29] arx.deidentifier.org/

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* [53]. While it has been shown that this measure can be defeated by trackers [54], 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 [55].

Practical Recommendation for Anonymization There is no universal rule that applies to all the cases. Authors of guidelines for sharing clinical trials data [56] 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 [56, 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 [57, 58, 59, 60, 61], while others require larger values of 11 (in USA [62, 63, 64, 65]) to 20 (in Canada [66, 67]). The maximum found in the literature was 25 [66]. Obviously the higher the *k*, the more suppression occurs or the more generalization is required.

2.5.2 Pseudonymization

Compared with anonymization as described in Section 2.5.1, pseudonymization also mitigates the LINDDUN threat types identifiability and linkability according to [2]. 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.

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.

2.6 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 long time in order to be able to reconstruct flow of events in case of any investigation.

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 [68]. 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.

2.7 Protection of Storage and Communication Channels

Protection of storage and communication covers several aspects:

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 com-

promising security of data) as well as up to *n* of unavailable nodes (without compromising security). Such approach has been demonstrated previously by Hydra FS[30] and Charon FS.[31]

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 Certification Authorities (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, as well as 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[32] [69]).

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 use of all versions of Secure Socket Layer (SSL) due to their inherent deficiencies [70], so that for reasonably secure communication the service providers are expected to have switched to Transport Level Security (TLS) 1.1 or newer (TLS 1.0 is also considered deprecated[33] [71]).

2.8 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 personally identifiable information 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.

[30] https://twiki.cern.ch/twiki/bin/view/EGEE/DMEDS
[31] https://github.com/biobankcloud/charon-chef
[32] http://www.nessus.org/
[33] https://forums.juniper.net/t5/Security-Now/NIST-Deprecates-TLS-1-0-for-Government-Use/ba-p/242052

2.9 Other Terminology

The key words “MUST”, “MUST NOT”, “REQUIRED”, “SHOULD”, “SHOULD NOT”, “RECOMMENDED”, “MAY”, and “OPTIONAL” in all further sections of this document (i.e., starting with Section 4) are to be interpreted as described in RFC 2119 [72]. “SHALL” and “SHALL NOT” will not be used as reserved words in this document for the sake of simplicity.

As common in IGTF documents,[34] 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 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.

[34] https://www.igtf.net/

3 IT Architecture and Data Management Strategy of BBMRI-ERIC

3.1 Functional Description

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 4 and the components are described in further detail in Section 3.2. Architecture of the system is fully distributed, following distributed architecture of BBMRI-ERIC itself, where it is called "hub and spokes" with central level, National Nodes level, and individual biobanks level. This architecture is applied to all the aspects including the long-term data storage and curation, querying data, migration of computations to data, etc. The architecture, however, must support temporary data caching for performance reasons. 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.

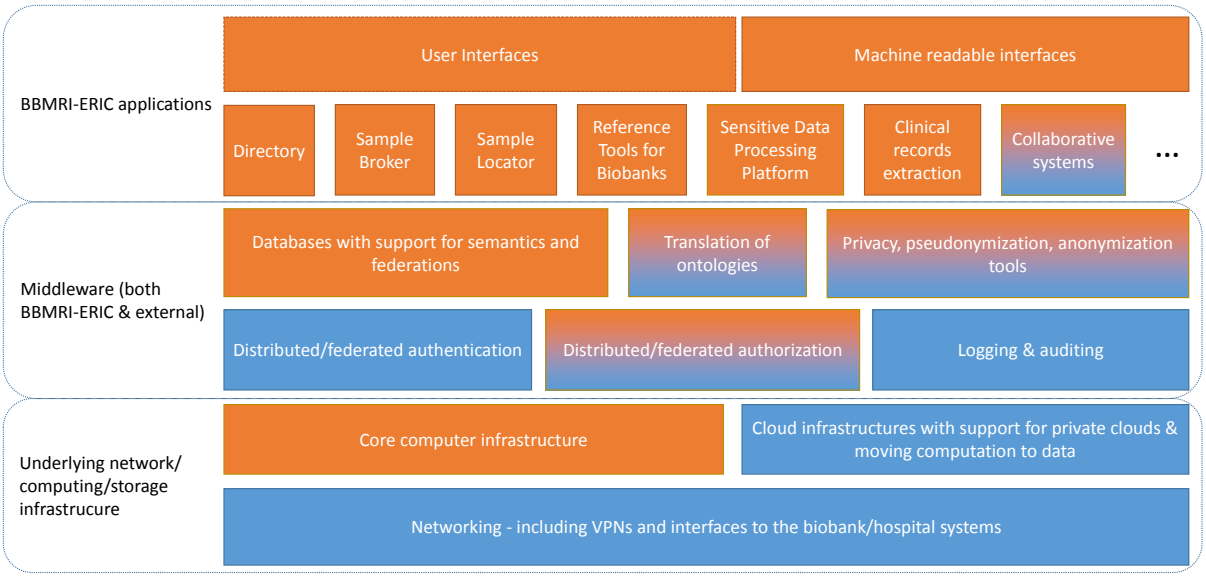

Figure 4: Software stack of BBMRI-ERIC IT system. Orange components are assumed to be build by BBMRI-ERIC, blue components are expected from other e-Infrastructures. Orange-blue components are assumed to be developed jointly with other e-Infrastructures.

From the data exchange perspective, BBMRI-ERIC is committed to FAIR principles[35] (Findable, Accessible, Interoperable, Reusable), with accessibility limited by privacy protection of patients and donors given the nature of data in BBMRI-ERIC infrastructure. This implies that access is only provided to the authorized people, i.e., typically researchers who work on research projects that have been reviewed by a competent ethical review board.

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

Typical workflow for the user starts with authenticated user[36] searching for the samples and/or data, or trying to identify biobanks to start collaboration with (see the Directory and Sample Broker/Locator components described in Section 3.2). Before accessing samples and/or actual privacy-sensitive data (data that is personal and not anonymous – see Requirement Req-3 on page 52 for definition and discussion of practically anonymous data), the user must submit a project that 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 fit for the intended purposes, 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 user's request is approved, the user signs MTA and/or DTA and the sample/data is given to the user.

When processing privacy-sensitive data, it is typically required that non-deidentified data never leaves biobank. Depending on the type of the request, the biobank can transfer either anonymous data or pseudonymous 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 pseudonymous data or even non-deidentified 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[37] [73, 74, 75].

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 clouds 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.[38] An extended version of this scenario is targeted by the Sensitive Data Processing Platform component in the software stack diagram.

Another specific aspect of BBMRI-ERIC infrastructure is the heterogeneity of data that are coming into the biobanks and that need to be mapped into consistent data sets. Therefore BBMRI-ERIC works with the federated databases with semantic data support (triple store systems) and translation of ontologies, which has been being worked upon, e.g., in the BioMedBridges project.[39] Specific issue for the clinical biobanks is the unstructured parts of clinical records that are on one hand one of the most valuable sources of information, but on the other hand that in many cases require reliable extraction including natural language processing, which is still a research challenge.

3.2 Description of Main Components

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 re-

[36] Strong authentication is needed, preferably multi-factor, because of the privacy and security aspects.
[37] http://www.p3g.org/biobank-toolkit/datashaper
[38] http://www.biobankcloud.com/
[39] http://www.biomedbridges.eu/

searchers to identify biobanks that might 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.[40]

Sample Broker This tool is intended for the researchers who already have their research intent/project and need samples or data to implement it. Inquiries by the researchers for the samples often span multiple biobanks and they are subject to iterative refinement. As a part of this process, the biobankers must understand various aspects of the expected methods to be used in the planned research, in order to evaluate whether their samples are fit for the particular purpose (e.g., analytical method). This is by its nature a M:N communication between researchers and biobankers, generating large overhead that can be simplified by employing efficient tools for group communication.

Sample 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 BBMRI-ERIC, the situation is, however, more complicated because of various strategies related to differential privacy [78, 79, 80] need to be in place. Approaches such as *k*-anonymity, *l*-diversity, and *t*-closeness together with generalization and suppression may result in substantial "hidden black matter" because in practice the high-dimensional data is sparse [81]. An alternative solution to avoid too much suppression is by reducing dimensionality, which may in turn result in users being unable to ask queries as specific as they need. Another aspect is competing interests of biobankers and researchers, which results in biobankers being reluctant to put all of their samples into a system that can identify individual samples. 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.

Ontology Translation Service With distributed nature of BBMRI-ERIC, the data come in many different ontologies even in a single domain.[42] As data harmonization and ontology translation is an extremely important service for many other tools, we define it as a separate component with well-defined interface to be incorporated into other applications.

Sensitive Data Processing and Sharing Platform This component is composed of two parts: one is the private cloud-based tools for biobanks and the other is a platform where sensitive data can be collected and shared, such as TSD[43] or MOSLER.[44]

Clinical Records Extraction Clinical records are a valuable source of information especially for the clinical biobanks, which take biosamples from the clinical practice. Typical clinical records, however, contain only limited structured information and large portions are written as free text in natural

[40] See, e.g., BioResource Impact Factor (BRIF)[41] [76, 77].

[42] 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.

[43] `https://www.uio.no/tjenester/it/forskning/sensitiv/`

[44] `https://wiki.bils.se/wiki/Mosler_user_documentation`

language, often with some particular domain specifics. In many cases, there is further complication for the biobanks that they are detached from the hospital information systems and may not access this data online. While very important and characteristic for BBMRI-ERIC, reliable extraction from the unstructured clinical records is still an open basic research problem to a large extent and therefore it is in the optional components list.

Reference Tools for National Nodes and Biobanks Because biobanks and BBMRI-ERIC national nodes have often very limited IT personnel capacity, BBMRI-ERIC is committed to provide reference tools for both of these levels. These tools are assumed to be distributed either as software packages or even as pre-installed and mostly pre-configured virtual machines.

An important aspect of the reference tools will be documentation of APIs and file formats used for the data exchange, as biobanks and national nodes will be free to replace any of the components of the reference tool set by the tools of their preference, only retaining the API interoperability.

3.3 Data Organization Description

The schema below tries to provide an overview of data organization. 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 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.**

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 information 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: omic 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 technical or administrative basis.[45] 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 to the hospital information system managers.

Example data: clinical records of patients.

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

Amount 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).

(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 avoid storing any privacy-sensitive data on the central level.

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

(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 fully federated 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 is 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

[45] 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.

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 [82, 83].

3.4 Data Formats and APIs

The most common interfaces in the BBMRI-ERIC community are REST interfaces. For linked data, JSON-LD and less frequently RDF is being used with Virtuoso[46] used as triple store database.

Other interfaces are used as appropriate for given applications. For example Directory 1.0 relies on hierarchy of LDAP servers (national nodes can run their own LDAP servers, or can upload LDIF/JSON data directly to the central server) and LDIF data format for distributed data queries and JSON translators are available in/out for the LDAP.

When dealing with the clinical data, hospital information systems rely on HL7 (Health Level 7)[47] as well as custom interfaces. Data often utilizes specialized formats such as DICOM[48] for imaging modalities. There is ongoing work on harmonization of Electronic Health Records (EHR) within HL7 called Fast Healthcare Interoperability Resources (FHIR),[49] which in turn relies again on REST.

National nodes and local biobanks run a variety of systems and APIs and it is one of the major goals of BBMRI-ERIC to simplify the situation by providing reference tools for the national nodes and biobanks.

As part of the efforts to improve quality and interoperability of APIs and data formats, BBMRI-ERIC actively participates in ISO TC 276[50] Working Group 5 (WG5) "Data processing and integration", which aims at (a) definition of data and model formats and their interfaces; (b) definition of metadata and relations of data and models; (c) quality management of processed data and models. In order to provide consistent input, BBMRI-ERIC also participates in ISO TC 276 WG1 (terminology) and WG2 (biobanking).

[46] http://virtuoso.openlinksw.com/
[47] http://www.hl7.org/
[48] http://dicom.nema.org/
[49] Pronounced "fire", http://hl7.org/implement/standards/fhir/ .
[50] http://www.iso.org/iso/home/standards_development/list_of_iso_technical_committees/iso_technical_committee.htm?commid=4514241

4 Use Cases

This section uses DFD to model use cases of BBMRI-ERIC [84] (Section 4.1), in order to evaluate them using STRIDE and LINDDUN (Section 2.1), as described in the previous section. This analysis results in definition of requirements for implementation of those services.

4.1 DFD-Based Modeling of BBMRI-ERIC Use Cases

4.1.1 S+UCs-1: Biobank browsing/lookup

This use case deals with publishing highly aggregated information about biobanks, collection, biobank networks, and possibly other 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 information that is considered practically anonymous (see Section 2.5.1 on page 33 and Requirement Req-3 on page 52 for discussion and definition of the term practically anonymous). In practice, this use case is implemented by the BBMRI-ERIC Directory.[51]

As shown in a DFD in Figure 5, 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 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 get also 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.

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 5. 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.

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.);

[51] http://bbmri-eric.eu/bbmri-eric-directory

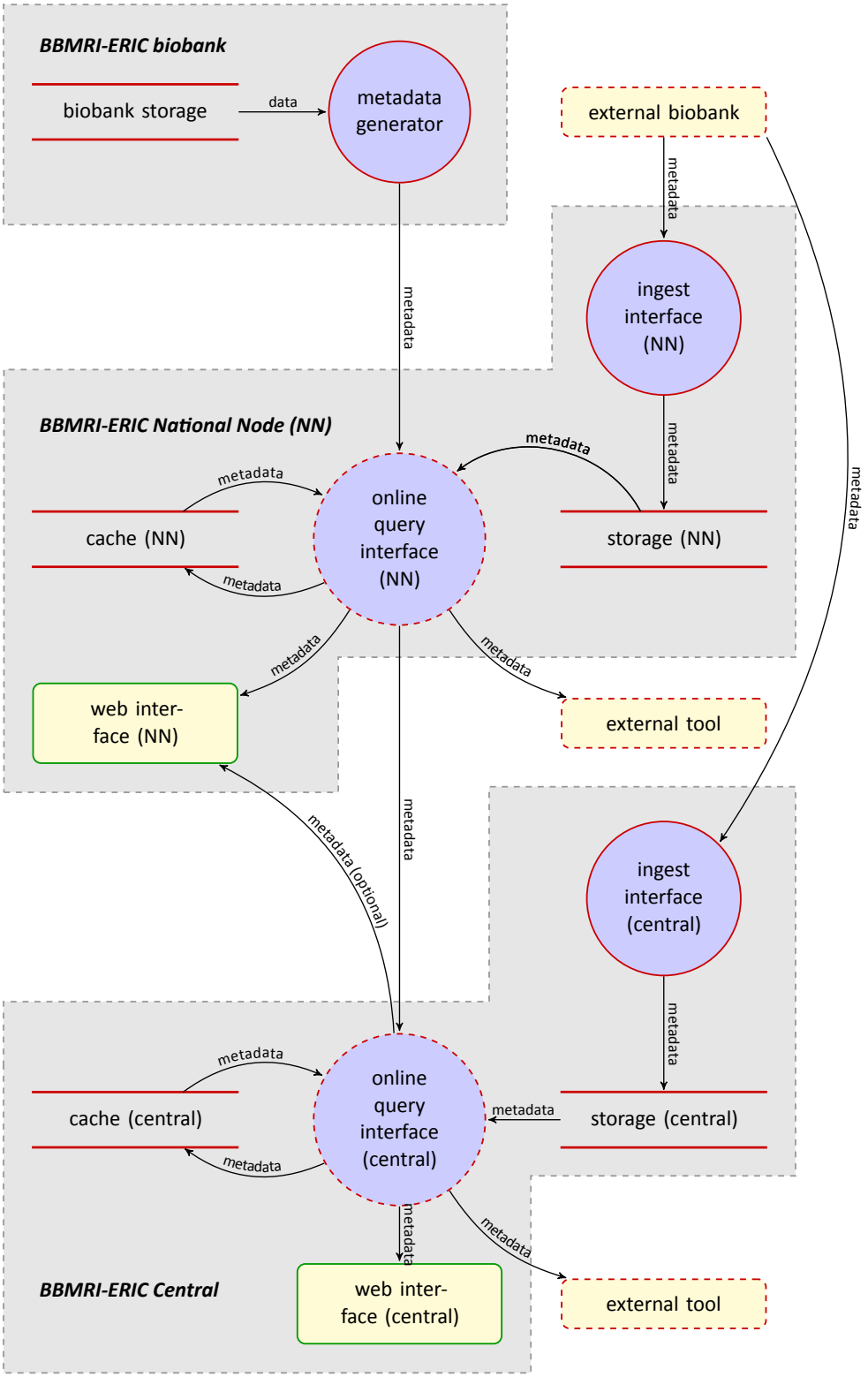

Figure 5: S+UCs-1: Biobank browsing/lookup

- *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. The data is expected to become more granular in the future, resulting in number of samples for each combination of parameters, while ensuring the data is still practically anonymous.

4.1.2 S+UCs-{2,3}: 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 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 $N + M$. The workflow will 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.

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 actuall 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.

As shown in Figure 6, the whole process starts with the requester communicating via BBMRI-ERIC web interface with the request tracker/broker 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 samples/data sets.

As can be seen from the DFD, during the sample/data brokering (negotiation), no sample-level or individual-level data leaves the biobank. The restricted access to the services is in place for the following

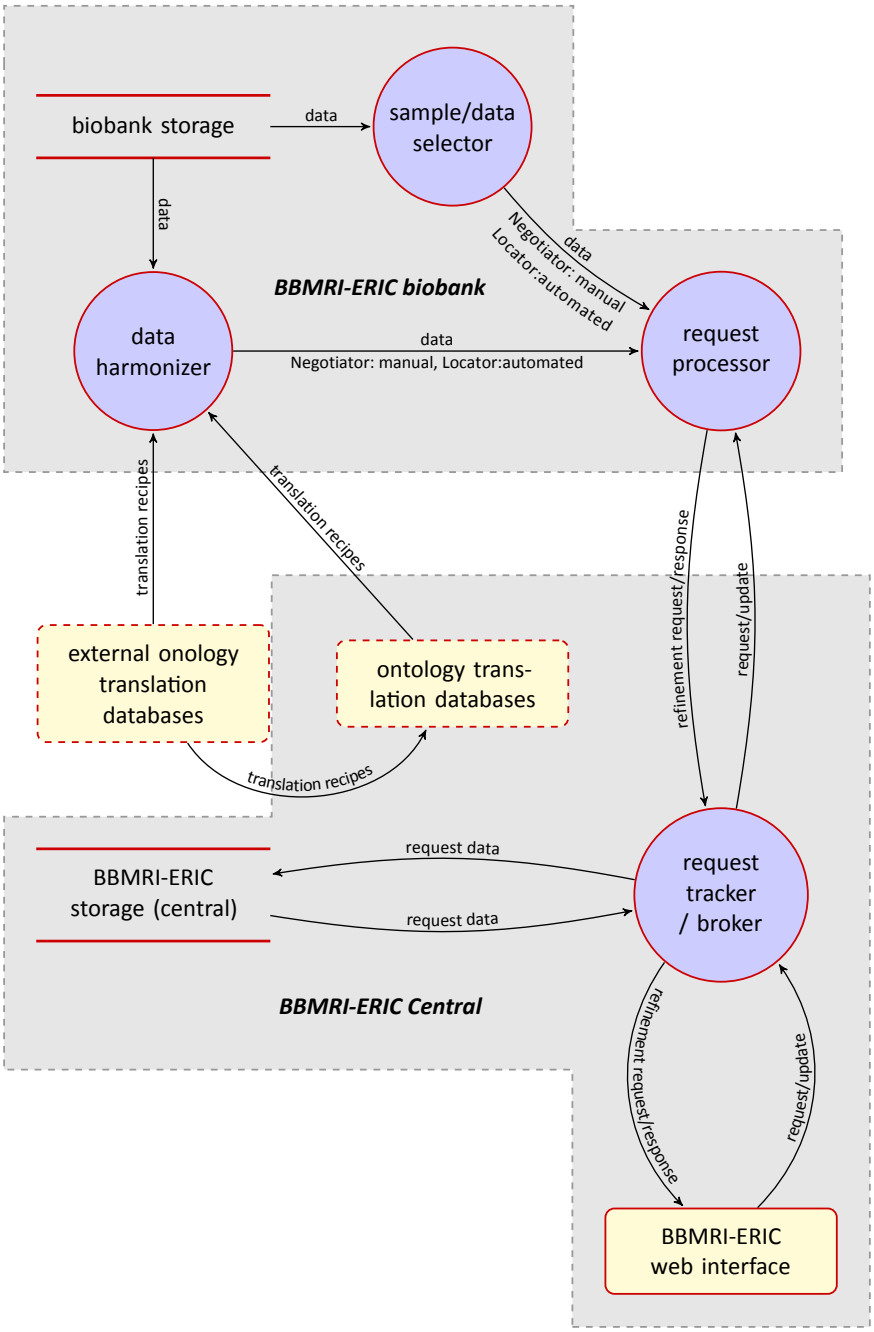

Figure 6: S+UCs-{2,3}: Sample/Data Negotiator and S+UCs-{5,6}: Sample Locator.
At the high-level component architecture used for DFD modeling, both use cases share the same data flow. The difference is in automation of the communication: while the data/sample selector is manual for S+UCs-{2,3}: Sample/Data Negotiator operated by the biobanker, it is automated for S+UCs-{5,6}: Sample Locator.

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 5, 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.

From the risk analysis perspective, an important aspect is that the requesters cannot browse automatically through informations about individual samples, which is functionality reserved for the biobankers. The sample/data selector module can be entirely detached/disconnected from the request processor, and even if there is online connection between the two, the transfer of the data from the selector to the request processor is a manually controlled step (similar 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 they will be used by the biobank as their own novel research ideas.

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

This use case deals with access of requesters to the sample-level data: browsing and search through individual samples stored in the biobanks and data sets related to individuals. Data may be either practically anonymous or even only pseudonymized, depending on dimensionality of data (the higher the worse) and acceptable level of suppression (the lower the harder). It is related to previous use case S+UCs-{2,3} and shares the same DFD in Figure 6 in page 46. The major difference is its automated access to the sample-level data or individual-level data, which may be highly multi-dimensional and thus problematic to achieve practical anonymity without very high suppression/generalization levels. Automated access to sample-level data is particularly sensitive from the privacy perspective, as it might be abused for reverse-engineering of de-identification (e.g., using statistical inference). Therefore it must be the subject of high-security restricted access and acceptance of liability by the user (researcher, possible requester).

4.1.4 S+UCs-14: Data Processing

This use case deals with processing very privacy-sensitive data, such as pseudonymized (individual-level) data. Potentially, this can be very large data sets, such as omics data (genomics, proteomics, metabolomics including time series, etc.) or processing of large imagery (often more than Gpix per image) in digital pathology.

From the scope of the source data, we can distinguish two types of analyses:

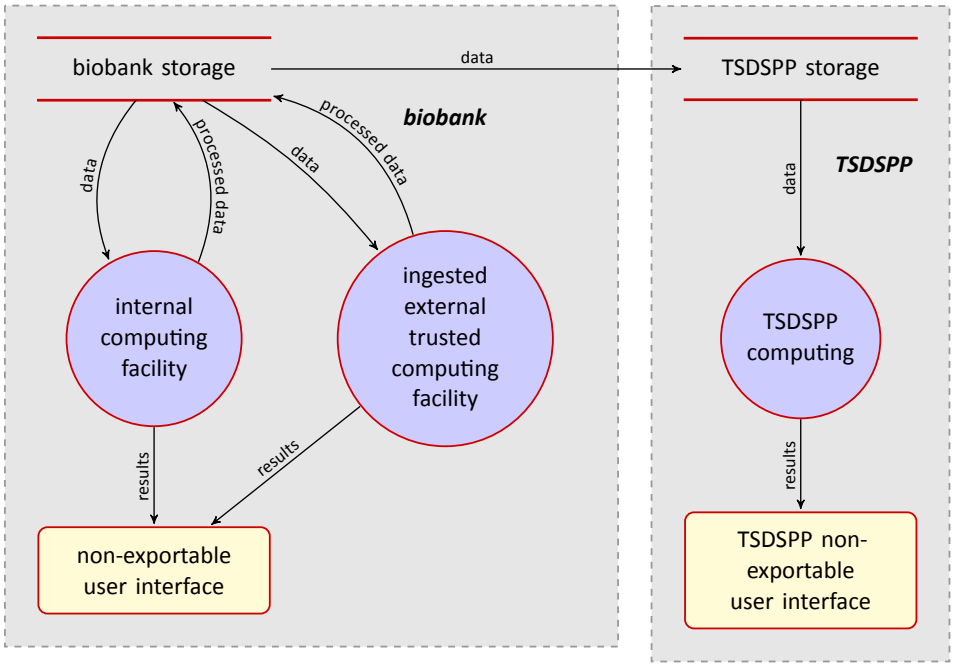

Figure 7: S+UCs-14: Data Processing.
TSDSPP stands for Trusted Sensitive Data Sharing and Processing Platform, such as MOSLER or TSD.

(1) analysis single source (biobank/bioresouce) data,

(2) data analysis across multiple (independent) sources, which can be of further two subtypes:

(2a) data pooled together on single location,

(2b) federated data analysis where only aggregate data leave the source.

The following scenarios can be considered for the processing of the data:

(1) data processing entirely inside the biobanks,
(2) data processing in dedicated Trusted Sensitive Data Sharing and Processing Platform (TSDSPP) such as MOSLER or TSD (see project-based RBAC description in Section 2.4.3), which allows for storage of the data with a possibility of extracting it from the TSDSPP (including running only trusted/certified processing software to avoid users dumping data to the user interfaces) and supporting access control based on users belonging to projects (multi-tenancy),
(3) ingestion of trusted computing infrastructures into the logical scope of the biobanks.

The first approach should be always feasible, while the remaining two depend on legal/ethical requirements in the given countries, as well as on availability of technologies and services (such as provisioning of certified cloud resources, see discussion of ISO 27018 in Section 2.8).

If legal framework supports it in given countries, scalability of this can be further improved by using distributed data storage systems which are secure despite using public (grid/cloud) resources such as Overbank[52] [85] or Hydra[53] [86, 87, 88, 89].

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

Table 5: 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 6: 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
– (not relevant), + (low), ++ (medium), +++ (high)						

Continued on next page...

[52] http://www.biobankcloud.com/?q=node/45

[53] https://twiki.cern.ch/twiki/bin/view/EGEE/DMEDS#What_is_Hydra

... continued from previous page.

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

Table 7: 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.

4.3 Relation to Business Model of BBMRI-ERIC Services

This section is tentative as an updated business model of BBMRI-ERIC is under preparation, which will provide information on specific conditions of access for BBMRI-ERIC services.

The business model of BBMRI-ERIC differentiates several access policies for BBMRI-ERIC services, based namely on membership in BBMRI-ERIC.

- *Services open for free to all users, irrespective of country origin.*
  Specific examples:
  - BBMRI-ERIC Directory for browsing/lookup of aggregate information about biobanks and collections, as described in S+UCs-1: Biobank browsing/lookup in Section 4.1.1.
- *Services available for free for BBMRI-ERIC full members and observers; users from other countries may be charged for their use.*
  Specific examples:
  - BBMRI-ERIC Sample Broker for negotiating access to the samples and data sets, as described in S+UCs-{2,3}: Sample/Data Broker in Section 4.1.2.
- *Services available for free for BBMRI-ERIC full members; users from other countries including BBMRI-ERIC observers may be charged for their use.*
  Specific examples:
  - BBMRI-ERIC Sample Locator for browsing and searching through sample-level databases as well as individual data sets, as described in S+UCs-{5,6}: Sample Locator in Section 4.1.3.
  - BBMRI-ERIC Data Processing platform for secure processing of sensitive data, as described in S+UCs-14: Data Processing in Section .
- *Services paid by all users.*
  Specific examples:
  - no examples available yet.

5 General Requirements

Privacy and security requirements represent 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.

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

5.1 Requirements on Personal Information Protection

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

Req-1 Unless exempted by Requirement Req-2, any non-deidentified 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. The data recipient also MUST NOT attempt to re-identify the person or otherwise counteract the de-identification of data, which SHOULD be covered by DTA or MTA.

Req-3 For the data to be considered **practically anonymous** 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.

$k \geq 5$ has been selected as the minimum commonly acceptable value based on literature survey discussed in Section 2.5.1, so that we don't impose unnecessary data suppression and generalization where not necessary. *It is of a particular note here that data custodians/owners may increase the k and/or apply other technical protection measures (see Section 2.5.1) if their national ethical and legal environment demands so or if they perceive the residual risks unacceptable.*

Req-4 High security restricted access (see page 27) *(a)* MUST incorporate LoA $\geq$ 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-5 The following table summarizes minimum requirements for different types of privacy-sensitive data

Table 8: Minimum requirements for basic data types. Non-personal data is used to denote data the does not contain any traces of privacy-sensitive data (e.g., data about operation of the biobank storage systems).

	raw (non-deindentifed)	pseudonynous	practically anonymous	non-personal
Authentication and authorization				
Identity verification	LoA $\geq$ 2	LoA $\geq$ 2	LoA $\geq$ 0	open
Authentication instance	LoA $\geq$ 3	LoA $\geq$ 2	LoA $\geq$ 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	$\geq$ 10 years	$\geq$ 10 years	$\geq$ 3 years	–
Data transfers and storage				
Encrypted storage Encrypted transfers	REQUIRED REQUIRED	REQUIRED REQUIRED		

Req-6 The BBMRI-ERIC policies MUST be compatible with GÉANT Data Protection Code of Conduct[54] [90].

5.2 Requirements on Accountability and Archiving

Req-7 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-8 Release of any samples or any data containing person-level information (i.e., including anonymous and pseudonymous data) MUST be stored in non-repudiable way by the biobank.

Req-9 Link MUST be maintained between the DTA/MTA and the samples and data sent to the requesting party.

Req-10 Access logs to any data that involves information on the level of individuals (e.g., sample-level data including practically anonymous 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-11.

Req-11 Access logs to any non-deidentified data or pseudonymized 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.

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

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,[55]
- 10 years for images in Italy and "forever" for clinical records (since the latter are considered legal documents)[56]
- 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,[57]
- 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,[58]
- 30 year in Germany.[59]

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

5.3 Requirements of Protection of Users Privacy

Req-12 BBMRI-ERIC MUST NOT use tracking of users[60] 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 MUST be publicly available and MUST be written in simple terms understandable also for non-technical users.

Req-13 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-14 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-15 The data coming from user tracking MUST be treated as confidential by BBMRI-ERIC.

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

[56] 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.

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

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

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

[60] Following users both in individual services and across different IT services, see, e.g., [91, 92, 93, 94, 95, 96] for more discussion of various techniques.

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.

5.4 Requirements on Data Storage, Transfers, and Computer Networks

Req-16 Non-deidentified data and pseudonymized data SHOULD be stored encrypted with state-of-the-art encryption strength appropriate to the sensitivity of the data.

See Section 2.7 for brief discussion of available technologies.

Req-17 Computer networks used for processing non-deidentified data and pseudonymized 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-18 Secure network protocols MUST be used when transferring privacy-sensitive data (non-deidentified data and pseudonymized data) over the network. For practically anonymous data it is RECOMMENDED.

See Section 2.7 for brief discussion of the state of the art, deprecation of Secure Socket Layers (SSL), etc.

5.5 Requirements on Software Design and Development

Req-19 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-20 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-21 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-22 Choice of programming language and third-party libraries and frameworks for the development SHOULD consider security aspects and SHOULD facilitate Requirements Req-20 and Req-21.

Req-23 Software development SHOULD use available static code analysis tools (and security-oriented analysis tools in particular) such as Coverity Scan.[61]

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

Use of such tools is facilitated by the open-source Requirement Req-20 and choice of programming language and various frameworks Requirement Req-22.

Req-24 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-25 Software developed within BBMRI-ERIC is RECOMMENDED to use publicly available code repositories with version management, such as SourceForge[62] or GitHub.[63]

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

Req-26 Software developed within BBMRI-ERIC SHOULD support versioning as a part of the configuration management.

Req-27 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.

[62] https://sf.net
[63] https://github.com/

6 Requirements on Use Cases

6.1 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" [81], we require that the data must adhere to the anonymity guidelines.

Req-28 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 *practically anonymous*: see Requirement Req-3 on page 52.

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

Req-29 Sample/Data Negotiator MUST require user to sign MTA or DTA before positively concluding negotiation of access to samples or data respectively.

Req-30 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-31.*

Req-31 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-32 Sample/Data Negotiator MUST require biobankers to consent with treating all the sample/data requests as well as reservations as confidential.

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

Req-33 Sample Locator MUST also fulfill requirements of the Sample/Data Negotiator (Section 6.2).

Req-34 Users MUST require users to consent to the terms and conditions, including refraining from any person re-identification attempts, before using Sample Locator.

Req-35 Sample Locator MUST require user to sign MTA or DTA before positively concluding negotiation of access to samples or data respectively.

6.4 S+UCs-14: Data Processing

General requirements apply for this use case, and particular attention should be paid to Requirements Req-2 and Req-5.

Req-36 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-37 Any third party computing/storage infrastructure used for processing and storing the data MUST provide sufficient liability.

Req-38 Physical computing resources used for processing privacy sensitive data (at least non-deidentified data or pseudonymized 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.[64]

6.5 Organization Security

Req-39 The security measures SHOULD be clearly documented as a part of the organizational measures on the institutional level (e.g., level of the biobank).

[64] 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.

7 Requirements on AAI

7.1 AAI Support for BBMRI-ERIC Business Model

In order to implement its business model (Section 4.3 on page 51), BBMRI-ERIC has the following requirements:

Req-40 Authentication supporting BBMRI-ERIC member affiliation, i.e., whether the user has work contract or owns business in any of the member countries, MUST be in place for any services that differentiate between paid/free model based on BBMRI-ERIC membership.

As of January 2016, BBMRI-ERIC members include (1) Austria (AT), (2) Belgium (BE), (3) Switzerland (CH), (4) Czech Republic (CZ), (5) Germany (DE), (6) Estonia (EE), (7) Finland (FI), (8) France (FR), (9) United Kingdom (GB), (10) Greece (GR), (11) Italy (IT), (12) Malta (MT), (13) Netherlands (NL), (14) Norway (NO), (15) Poland (PL), (16) Sweden (SE), (17) Turkey (TR).

7.2 Use Cases for AAI

7.2.1 Public/Open Services

Openly accessible services of BBMRI-ERIC may support LoA 0 authentication (e.g., cookies for web-based applications) to store user preferences and possibly also to track and analyse user behavior. This is important in order to demonstrate impact of BBMRI-ERIC infrastructure, to analyze behavior of users and their use of BBMRI-ERIC services so that the services can be optimized in the future.

If the LoA 0 authentication is in place and whenever technically feasible, there must be optional authentication (LoA $\geq$ 1, using any IdP that supports it) to allow for explicit account management if the user wishes to do so. This approach may be preferred by the privacy-conscious users who want to have control of their accounts.

Tracking of users on public services without explicit login, such as services where LoA 0 authentication is used for storing personal preferences, can be considered intrusion into user's privacy. Hence tracking of users must be clearly documented in publicly available BBMRI-ERIC policy for transparency reasons in as simple terms as possible, so that even non-technical users understand to reasonable extent what is collected and why. Wherever required by the regulations, the user must be requested to provide explicit consent with regulations.

Target use cases:

- S+UCs-1: Biobank browsing/lookup (Section 4.1.1),
- BBMRI-ERIC web site.

List of requirements:

Req-41 Openly accessible services of BBMRI-ERIC MAY support LoA 0 authentication to store user preferences and to track behavior of users.

Req-42 If the LoA 0 authentication is in place and whenever technically feasible, there MUST be OPTIONAL authentication LoA ≥ 1 for users who prefer explicit account management.

7.2.2 Restricted Services

This use case deals with all the services that deal with sample-level anonymized data, i.e., restricted services (DAC, MAC, or RBAC) or services subject to committee-controlled access. All such services must be bound to LoA ≥ 2 authentication supporting also project affiliations. As for authorization, all such services must support project-based RBAC or must provide project affiliations to support decisions in committee-controlled access.

In order to provide access to those users whose home institutions do not participate in any accepted identity federation, BBMRI-ERIC will provide a fallback IdP with authentication instance of LoA 2 and BBMRI-ERIC itself and its BBMRI-ERIC National Nodes must provide registration with minimum LoA 2, see Section 7.3.1 for more in-depth discussion.

Existence of projects is ideally asserted by the hosting institution. For accessing BBMRI-ERIC infrastructure, it is important to ensure that the project exists and that it has been favorably reviewed (accepted) by an ethical board. We call this process **project validation**. It is *not expected* that BBMRI-ERIC performs reviews of the projects (neither scientific reviews nor ethical reviews). As some institution may be unable to provide such assertions, BBMRI-ERIC and its National Nodes must implement additional services as described in Section 7.3.1.

In order to asses compliance of the research intent to the informed consent for the samples and data sets, the authentication must support also project affiliation of users. Project management, i.e., affiliation of the persons to the projects, can be done either on institutional basis (i.e., the institution asserts existence of the project and affiliation of persons to that project), or it may be done individually by project investigators (PIs) of the projects. The latter approach requires BBMRI-ERIC to implement additional services as described in Section 7.3.1.

In order to support user identities bound to project affiliations in legacy systems (operating systems and legacy applications), the AAI must support user identities in the form of userID_projectID.

This scenario must also support role delegation as described in Section 2.3.6. That means that a person should be able to delegate his/her role to another person, unless the role is marked as non-delegable (default should be delegable roles).

Target use cases:

- S+UCs-{2,3}: Sample/Data Broker (Section 4.1.2),
- S+UCs-{5,6}: Sample Locator (Section 4.1.3).

List of requirements:

Req-43 All the services with restricted access or committee-controlled access MUST be bound to LoA ≥ 2 authentication.

Req-44 AAI MUST support attribute-based project affiliation assertions.

Req-45 AAI MUST support management of project affiliations by the PI of the project or a person to which PI delegates the management.

Req-46 AAI MUST support user identities in form of userID_projectID for common operating systems and legacy applications.

Req-47 AAI MUST support role delegation as well as optional marking of roles as non-delegable.

Req-48 All requirements of use case described in Section 7.3.1 apply.

7.2.3 Highly-Secure Authenticated User Access

This use case deals with any system where the user directly deals with either pseudonymized data or even the raw source data (typically inside the biobanks). This requires strong assertions of user identity both on the level of registration and authentication instances; hence we require LoA ≥ 3. Other requirements are the same as for the previous use case (Section 7.2.2).

Target use cases:

- S+UCs-14: Data Processing (Section 4.1.4),
- signing DTA by the requester and transferring data from the biobank to the requester.

List of requirements:

Req-49 Any system which allows direct access or processing of pseudonymized or raw source data MUST require authentication at LoA ≥ 3.

Req-50 All other requirements (with exception of LoA 2 authentication) of use case described in Section 7.2.2 apply.

7.3 Additional Requirements on AAI

7.3.1 Access of "Homeless" Users and "Homeless" Projects

This use case deals with support for users whose home institution does not participate in the identity federations supported by BBMRI-ERIC, or institutions which do not provide sufficient information about their users. In order to provide access to those users whose home institutions do not participate in any accepted identity federation, the BBMRI-ERIC will provide a fallback IdP with authentication instance of LoA 2 and BBMRI-ERIC itself and its BBMRI-ERIC National Nodes must provide registration with minimum LoA 2, which means that the "homeless" users must either physically visit the National Node and present government-issued photo ID card for in-person registration, or the applicant must provide references to and must assert to current possession of government-issued photo ID and a second form of identification. The National Node must be provided by the user with minimum of name, date of birth, address and personal phone number. It is envisioned that in-person registrations may be required by the National Nodes if they cannot reliably implement remote registrations. See Section 2.3.2 for more discussion.

This use case also solves situations when project validation and project affiliation assertions cannot be provided by the hosting institution. BBMRI-ERIC and its National Nodes must be able to perform such project validation: National Nodes for the projects hosted by their country (either national projects or international projects where the project coordinator resides in the given country), while BBMRI-ERIC needs to implement validation of remaining projects. Such process is also important for projects that are bound to the identity of the PI and not to the institution (e.g., ERC grants). This can be implemented, e.g., by a Proxy IdP instances running BBMRI-ERIC and its National Nodes.

For project affiliations, BBMRI-ERIC will implement a project membership management system, which allows PIs of the projects to assign people to the projects. This solves the problem with institutions not providing the project affiliations, as well as cross-institutional projects where contributing institutions do not have sufficient information to provide the project affiliations.

Req-51 BBMRI-ERIC MUST provide registration authority with registration level of LoA 2.

Req-52 BBMRI-ERIC National Nodes MUST provide registration authority with registration level of LoA 2.

Req-53 BBMRI-ERIC and BBMRI-ERIC National Nodes MUST support project validation (existence check and check of approval by an ethical board) if these cannot be asserted by the hosting institution as a part of the federated AAI.

BBMRI-ERIC National Nodes are responsible for projects hosted in their countries (for international projects, this includes projects where project coordinator is in the given country), while central BBMRI-ERIC is responsible for projects hosted in countries where there is no National Node or for projects that do not have any hosting country.

Req-54 BBMRI-ERIC MUST provide fallback IdP with authentication instance of LoA 2.

7.3.2 BBMRI-ERIC Member Affiliation of Users

Members of BBMRI-ERIC are primarily *countries* (both European and non-European), where they can be either *full members* and *observers*.

There is also a special status of *organizational members* of BBMRI-ERIC, which allows international organizations to become members.

Business model of BBMRI-ERIC assumes that not all the services are available for free for everybody and that full members enjoy most benefits, while observers should still enjoy some benefits over non-members.

As BBMRI-ERIC IT infrastructure deals with individual users and not with countries (or organizational members of BBMRI-ERIC), AAI must be able to provide information on affiliations of individual users.

Affiliation of the users is decided based on the following rules:

1. A user is considered affiliated with the given country, if he is employed by an institution residing solely in the given country.

2. For institutions that span more than one country (e.g., having subsidies in various countries), the user is considered affiliated with given country if he is employed by a subsidy in that country.

3. BBMRI-ERIC National Node may approve exceptional (adopted) country affiliation for a user, which may not be affiliated with the given country otherwise. The National Node may only act with respect to its own country.

4. A user is considered affiliated with an organizational member of BBMRI-ERIC, if he is employed by given international organization.

Req-55 AAI MUST provide means to determine the countries with which the user is affiliated, at least when LoA $\geq$ 2 is in place and the user is not employed by an international organization.

Req-56 AAI MUST be able to indicate that the user is affiliated with the international organization, that is member of BBMRI-ERIC.

Req-57 AAI MUST provide means to add additional (exceptional) affiliation of a user to the country, if it is approved (adopted) by the BBMRI-ERIC National Node in the given country.

Req-58 AAI MUST support integration of all the countries that are members (both full members and observers) of BBMRI-ERIC.

As of January 2016, BBMRI-ERIC members include (1) Austria (AT), (2) Belgium (BE), (3) Switzerland (CH), (4) Czech Republic (CZ), (5) Germany (DE), (6) Estonia (EE), (7) Finland (FI), (8) France (FR), (9) United Kingdom (GB), (10) Greece (GR), (11) Italy (IT), (12) Malta (MT), (13) Netherlands (NL), (14) Norway (NO), (15) Poland (PL), (16) Sweden (SE), (17) Turkey (TR).

Req-59 AAI MUST support extension to other countries in the future, including also non-European countries, as BBMRI-ERIC is likely to expand.

7.3.3 BBMRI-ERIC Identity and Robustness/Performance Enhancements

In order to support merging of various virtual identities of a single person, BBMRI-ERIC will introduce a **BBMRI-ERIC Identity**. Each physical person is recommended to merge their virtual identities coming from different identity providers supported by BBMRI-ERIC, into this identity. Each BBMRI-ERIC Identity will receive a unique non-reassignable identifier and optional nickname, which can be changed by the user.

In this section, we only list requirements on the BBMRI-ERIC Identity, such as support for translation of credentials, at least to X.509 certificates, SSH keys, and Kerberos tickets, in order to support authentication to services such as web applications and web services, virtual machines, or access to data archives. For design decisions, the reader should refer to Section 8.1.1.

Req-60 BBMRI-ERIC MUST provide a digital identity for each physical person that uses its infrastructure. This identity is called BBMRI-ERIC Identity.

Req-61 BBMRI-ERIC Identity MUST allow merging (linking) of other user's identities. Each physical person is RECOMMENDED to have a single BBMRI-ERIC Identity and merge their existing virtual identities into it.

Req-62 BBMRI-ERIC Identity MUST support merging (linking) with ORCID.

Req-63 An opaque and non-reassignable identifier in the form of `[a-f0-9]{32,}@identity.bbmri-eric.eu` MUST be assigned to each BBMRI-ERIC Identity.

Req-64 A special testing identity of `00000000000000000000000000000000@identity.bbmri-eric.eu` MUST NOT be assigned to anybody and MUST NOT be used for allowing access to any services with restricted access.

Req-65 A user MAY choose a nickname in the form of `[a-z_][a-z0-9_-]*@users.bbmri-eric.eu` which MUST be unique in any given time. However, the nickname MAY be reassigned to another user after the original user chooses a different one. Because of reassignment option, each assignment must be persistently logged, see Requirement Req-66.

Req-66 Assignment of BBMRI-ERIC Identity identifiers and nicknames together timestamps of assignment must be stored permanently by BBMRI-ERIC for auditing purposes.

Req-67 When multiple identities are merged (linked) together into BBMRI-ERIC Identity, the resulting attribute set MUST be constructed as a union of attributes of the merged identities. Scoped attributes, such as `eduPersonScopedAffiliation`,[65] can be merged directly, while the scope MUST be appended for unscoped parameters as a part of the merging.

Req-68 In order to mitigate problems with temporal unavailability of users home institution, BBMRI-ERIC AAI SHOULD operate a Proxy IdP that allows for "caching" of user identities including their attributes.

[65] `https://www.internet2.edu/media/medialibrary/2013/09/04/internet2-mace-dir-eduperson-201203.html#eduPersonScopedAffiliation`

Req-69 Caching of attributes must be done for maximum of 7 days and must provide a mechanism for explicit immediate invalidation.

Req-70 BBMRI-ERIC Identity MUST support translation of credentials, at least to X.509 certificates, SSH keys, and Kerberos tickets.

7.3.4 BBMRI-ERIC AAI Data Retention Policy

In order to make it compatible with existing legal frameworks, the BBMRI-ERIC AAI policy must meet at least the following requirements:

Req-71 Because of compatibility with European data protection regulations [39], requiring personal data to be deleted when no longer needed, the BBMRI-ERIC Identity will be considered inactive after being unused for authentication for 24 months. BBMRI-ERIC Identity identifier, nickname, group/project membership, and logs MUST be retained for inactive identities, while all other personal information including attributes MUST be deleted.

Req-72 Because of compatibility with European data protection regulations requiring *"right to be forgotten"*[66] (see [39, Articles 11 and 12], proposal of upcoming General Data Protection Regulation, and May 13, 2014 ruling of European Court of Justice in Google vs. Costeja case[67]), requiring personal data to be discarded per user's request, the BBMRI-ERIC Identity MUST be deactivated if requested by that person.

This requirement only involves deactivation of the account, but does not imply removal of account and access logs, in order to comply with Requirements Req-10 and Req-11.

7.3.5 Authentication Interfaces for SPs

Req-73 In order to make BBMRI-ERIC AAI compatible with legacy software systems as well as with newly developed application, the AAI MUST provide at least SAML IdP interface and LDAP directory interface for querying attributes (such as affiliation and group/project membership).

7.3.6 Authorization

Access control layer of BBMRI-ERIC infrastructure assumes conceptual separation of Policy Decision Points (PDPs), where the access is decided, and Policy Enforcement Points (PEPs), where access is actually enforced and implemented. Typically, it is possible to have a single PDP and multiple PEPs for distributed services (e.g., distributed storage facilities).

Req-74 For restricted access services, at least one of the following access control mechanisms MUST be implemented: DAC or MAC or RBAC or committee-controlled access.

[66] https://en.wikipedia.org/wiki/Right_to_be_forgotten

[67] ECLI:EU:C:2014:317, http://curia.europa.eu/juris/documents.jsf?critereEcli=ECLI:EU:C:2014:317

Req-75 RBAC is RECOMMENDED to be implemented for services that do not require committee-controlled access.

Req-76 Any changes to access control decisions must be available for logging.

Additionally, for committee-controlled access, the following rules apply:

Req-77 Committee-controlled access must store the decisions persistently.

Req-78 Committee-controlled access must log decision outcomes (any changes) for minimum of 3 years.

8 AAI Architecture

BBMRI-ERIC will rely on identity federations provided by other e-Infrastructures and with government-backed federations, while supplementing it with its own infrastructure (kept to the minimum extent possible) to deal with situation where users come from organization that do not participate in identity federations or their identity providers do not provide sufficient information.

For requirements on identity federations, see discussion of use cases for AAI and list of requirements in Section 7.2.

In order to implement authentication, BBMRI-ERIC is expected to work with:

- GÉANT in AARC/AARC2 and VOPaaS,
- EGI in EGI-Engage,
- government-backed identity federations such as successors of STORK pilots (see Section 2.3.2 on page 23).

8.1 Authentication

As discussed in the requirements, BBMRI-ERIC will rely on federated identity management in order to ensure scalability and validation of real-world identities. *These services should be provided by e-Infrastructures such as GÉANT and BBMRI-ERIC Identity described below is understood only as a interim solution circumventing temporary availability problems.* For list of requirements on BBMRI-ERIC Identity, the reader should refer to Section 7.3.3.

8.1.1 BBMRI-ERIC Identity

BBMRI-ERIC Identity will be supported also via Proxy IdP operated by BBMRI-ERIC, which will be introduced for several reasons:

- to overcome temporal IdP availability problems at the user's home institutions,
- to support effectively users who are registered directly by BBMRI-ERIC or its National Nodes as discussed in Requirements Req-51 and Req-52,
- to insert additional attributes (e.g., project affiliation) that is not provided by the user's home institution, as discussed in Requirement Req-53.

This is seen as a temporal solution before the BBMRI-ERIC Identity is provided by one of the infrastructures that have this as their primary scope and able of providing acceptable Service Level Agreement or at least Service Level Declaration.

9 Cloud-Based Data Processing Architecture

BBMRI Competence Center deals with the processing the data stored in the biobanks, and hence privacy protection of the participants (patients or donors who have decided to donate their samples and data for the research purposes). The goal of the Competence Center is to employ EGI Federated Cloud technologies to enable biobanks to store and process the large volumes of their privacy-sensitive data (see Sections 3 and 4.1.4) in a scalable way.

In order to employ synergy with the tools developed previously in the broader BBMRI-ERIC context, the Competence Center has decided to use BiobankCloud technology,[68] which has been primarily focusing on employing private clouds built inside the biobanks in order to do the scalable processing of genomics and other types of privacy-sensitive omics data. Within the EGI-Engage project, the following aspects will be dealth with:

- make BiobankCloud interoperable with the EGI Federated Cloud platform, which includes implementation of OGF's Open Cloud Computing Interface (OCCI) support and support for EGI API in Apache jclouds® project,
- make BiobankCloud integrated with the Shibboleth federated authentication system,
- use EGI Federated Cloud platform to deploy pilot private clouds in the select biobanks of BBMRI-ERIC national nodes participating in BBMRI Competence Center of EGI-Engage,
- explore the possibility to use services of cloud providers (such as providers contributing to EGI Federated Clouds platform or even other cloud providers) for processing of the privacy-sensitive data,
- explore the use of secure storage platform provided by BiobankCloud to store the data in a scalable way using cloud reseources (see Section 10).

9.1 BiobankCloud Data Processing Platform

BiobankCloud is a front-end to Hadoop that provides a new model for multi-tenancy in Hadoop, based around studies. The owner of the study manages membership himself/herself (without the need for system administrator involvement), and users can have different roles in the study. The two roles[69] we support are data scientists, who can run programs, and data owners, who can also curate, import, and export data. Users are prevented from copying data between studies or running programs that process data from different studies, even if the user is a member of those studies. That is, we prevent the cross-linking of data across studies. A more security-oriented way of describing this is that we implement

[68] `http://www.biobankcloud.com/`

[69] There is also one more auxiliary/technical role to support software updates via `karamelbor`, which is to be done by the IT administration of the biobanks if BiobankCloud is used within private clouds. The scenario with third-party clouds is yet to be explored.

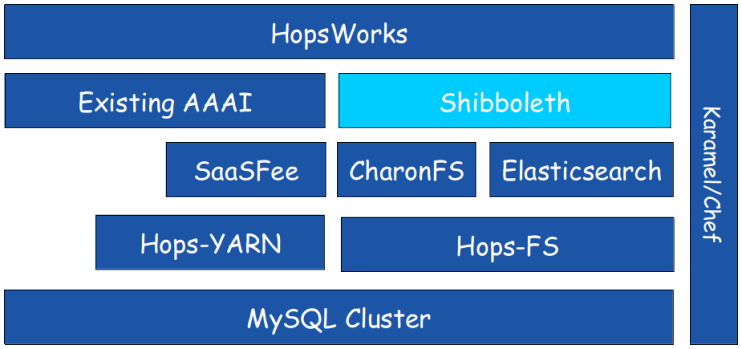

Figure 8: BiobankCloud Architecture

multi-tenancy using dynamic roles, where the user's role is based on the currently active study. Users are still able to share datasets between studies, however. Sharing of datasets between studies without copying is currently only supported within the context of a single BiobankCloud cluster. Datasets can also be copied securely between studies, although this involves copying the data between clusters, using a tool called CharonFS [43].

BiobankCloud also supports the processing of data using Hadoop data parallel processing frameworks, such as MapReduce, Spark, Flink, Adam, and SaasFee. Adam and SaasFee are of particular interest to bioinformaticians as they support many popular scalable workflow pipelines for bioinformatics, such as variant-calling for whole genome sequence data, and RNA-Seq. As Figure 8 shows different components of BiobankCloud, it is built on a new distribution of Hadoop called Hops. BiobankCloud is open-source and licensed as Apache v2, with database connectors licensed as GPL v2. BiobankCloud can be deployed on-premises (bare-metal), on private clouds and public clouds. In this project, we will focus on private cloud deployments using Karamel.

9.2 BiobankCloud on Private Clouds

We have decided to use EGI as our default private cloud platform for deploying BiobankCloud. Currently BiobankCloud has support to be run on known public clouds, Amazon AWS and Google Compute Engine, as well as in house premises and OpenStack however by adding EGI support as our default private cloud we are enabling BiobankCloud with a higher level of security for sensitive data.

Elastic private clouds: offloading to other Clouds

In BiobankCloud CharonFS [43] is used to share data using public clouds. CharonFS is a cloud-backed file system capable of storing and sharing big data in a secure and efficient way with minimal management and no dedicated server infrastructure. Charon builds upon on multi-cloud data replication to avoid having any single cloud service as a single point of failure, using instead distributed trust for operating

correctly even if a fraction of the providers are unavailable or misbehave. By leveraging CharonFS we aim for an elasticity solution that can alternate between public and private clouds.

9.3 Federated Authentication for BiobankCloud

BiobankCloud supports authentication using an identity local to a single BiobankCloud instance. User identity consists of a validated email address and a 2nd factor, generated either by a smartphone or a Yubikey dongle. BiobankCloud implements authentication using a JAAS authentication plugin for Glassfish (a J2EE application server). In EGI-Engage, we would, however, like to support federated authentication using Shibboleth. Shibboleth is an identity provider that implements widely used federated identity standards. It supports single sign-on services and extends its reach into other organizations and new services through authentication of users and securely providing appropriate data to requesting services. Shibboleth will enable users, who have a single federated identity, to log in to potentially any BiobankCloud cluster, given appropriate permissions. We will implement support for Shibboleth authentication by developing a new JAAS authentication plugin for Glassfish (a J2EE application server) that will enable users to login to BiobankCloud using their existing GÉANT or BBMRI-ERIC federated identity.

9.4 What is Karamel?

Karamel is a management tool for reproducibly deploying and provisioning distributed applications on bare-metal, cloud or multi-cloud environments. Users of Karamel experience the tool as an easy-to-use user interface (UI) driven approach to deploying distributed systems or orchestration distributed jobs in a cluster.

Karamel users can open a cluster definition file that describes a distributed system or jobs as:

- the application stacks used in the system, containing the set of services in each application stack
- the provider(s) for each application stack in the cluster (the cloud provider or IP addresses of the bare-metal hosts)
- the number of nodes that should be created and provisioned for each application stack
- configuration parameters to customize each application stack.

Karamel is an orchestration engine that orchestrates:

- the creation of virtual machines if a cloud provider is used
- the global order for installing and starting services on each node
- the injection of configuration parameters and passing of parameters between services
- connecting to hosts using ssh and running chef recipes using chef solo.

Karamel is built on the configuration framework, Chef. The distributed system or experiment is defined in YAML as a set of node groups that each implement a number of Chef recipes, where the Chef cookbooks are deployed on Github. Karamel orchestrates the execution of Chef recipes using a set of ordering

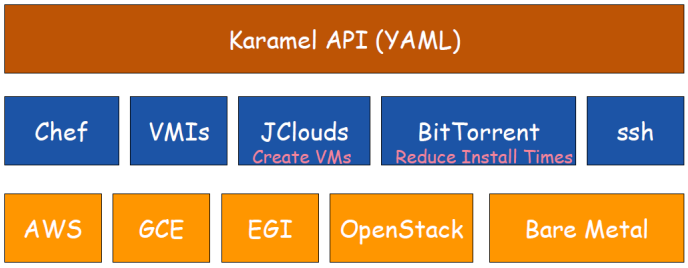

Figure 9: EGI in Karamel Stack

rules defined in a YAML file (Karamelfile) in each cookbook. For each recipe, the Karamelfile can define a set of dependent (possibly external) recipes that should be executed before it. At the system level, the set of Karamelfiles defines a directed acyclic graph (DAG) of service dependencies. Karamel system definitions are very compact. We leverage Berkshelf to transparently download and install transitive cookbook dependencies, so large systems can be defined in a few lines of code. Finally, the Karamel runtime builds and manages the execution of the DAG of Chef recipes, by first launching the virtual machines or configuring the bare-metal boxes and then executing recipes with Chef Solo. The Karamel runtime executes the node setup steps using Apache jclouds® and Ssh. Karamel is agentless, and only requires ssh to be installed on the target host. Karamel transparently handles faults by retrying, as virtual machine creation or configuration is not always reliable or timely.

9.5 BiobankCloud on Karamel

BiobankCloud powered by Karamel can easily be installed by non-technical users who can click-through an installation using only a file that defines a BiobankCloud cluster and account credentials for a cloud computing platform. Our solution is based on the configuration management platform Chef [97]. The main reason we adopted Chef is that it provides support for both upgrading long-lived stateful software and parametrized installations. Chef has, however, no support for orchestrating installations. For distributed systems with many services, such as BiobankCloud, there is often a need to start and initialize services in a well-defined order, that is, to orchestrate the installation and starting of services - that is basically what Karamel adds into Chef.

9.6 BiobankCloud Cluster Definition

We have written karamelized Chef cookbooks for installing all of the components of BiobankCloud, and we provide some sample cluster definitions for installing small, medium, and large BiobankCloud clusters. Users are, of course, expected to adapt these sample cluster definitions to their cloud provider or bare-metal environment as well as their needs.

The following is a brief description of the karmelized Chef cookbooks that we have developed to support the installation of BiobankCloud. The cookbooks are all publicly available at:

- http://github.com/hopshadoop/apache-hadoop-chef
- http://github.com/hopshadoop/hops-hadoop-chef
- http://github.com/hopshadoop/elasticsearch-chef
- http://github.com/hopshadoop/ndb-chef
- http://github.com/hopshadoop/zeppelin-chef
- http://github.com/hopshadoop/hopsworks-chef
- http://github.com/hopshadoop/spark-chef
- http://github.com/hopshadoop/flink-chef
- http://github.com/biobankcloud/charon-chef
- http://github.com/biobankcloud/hiway-chef

The Listing 1 is a cluster definition file that installs a very large, highly available, BiobankCloud cluster on 56 `m3.xlarge` instance on AWS/EC2:

Listing 1: Karamel Cluster Definition for BiobankCloud

```
name: BiobankCloudMediumAws
ec2:
    type: m3.xlarge
    region: eu-west-1
cookbooks:
  hops:
    github: "hopshadoop/hops-hadoop-chef"
    branch: "master"
  hadoop:
    github: "hopshadoop/apache-hadoop-chef"
    branch: "master"
  hopsworks:
    github: "hopshadoop/hopsworks-chef"
    branch: "master"
  ndb:
    github: "hopshadoop/ndb-chef"
    branch: "master"
  spark:
    github: "hopshadoop/spark-chef"
    branch: "hops"
  zeppelin:
    github: "hopshadoop/zeppelin-chef"
    branch: "master"
  elastic:
    github: "hopshadoop/elasticsearch-chef"
    branch: "master"
  charon:
    github: "biobankcloud/charon-chef"
    branch: "master"
  hiway:
    github: "biobankcloud/hiway-chef"
    branch: "master"
attrs:
  hdfs:
    user: glassfish
    conf_dir: /mnt/hadoop/etc/hadoop
  hadoop:
    dir: /mnt
    yarn:
```

```
        user: glassfish
        nm:
          memory_mbs: 9600
        vcores: 8
    mr:
        user: glassfish
  spark:
    user: glassfish
  hiway:
    home: /mnt/hiway
    user: glassfish
    release: false
    hiway:
      am:
        memory_mb: '512'
        vcores: '1'
      worker:
        memory_mb: '3072'
        vcores: '1'
  hopsworks:
    user: glassfish
    twofactor_auth: "true"
  hops:
    use_hopsworks: "true"
  ndb:
    DataMemory: '8000'
    IndexMemory: '1000'
    dir: "/mnt"
    shared_folder: "/mnt"
  mysql:
    dir: "/mnt"
  charon:
    user: glassfish
    group: hadoop
    user_email: jdowling@kth.se
    use_only_aws: true
groups:
  master:
    size: 1
    bbcui:
        - ndb::mgmd
        - ndb::mysqld
        - hops::ndb
        - hops::client
        - hopsworks
        - spark::yarn
        - charon
        - zeppelin
        - hiway::hiway_client
        - hiway::cuneiform_client
  metadata:
    size: 2
    recipes:
        - hops::ndb
        - hops::rm
        - hops::nn
        - ndb::mysqld
```

```
elastic:
  size: 1
  recipes:
      - elastic
database:
  size: 2
  recipes:
      - ndb::ndbd
workers:
  size: 50
  recipes:
      - hops::ndb
      - hops::dn
      - hops::nm
      - hiway::hiway_worker
      - hiway::cuneiform_worker
      - hiway::variantcall_worker
```

9.7 Plan to support EGI in Karamel

Adding support for a new Cloud such as EGI is straight forward in Karamel, like Figure 9 shows EGI will be one of the cloud providers . In orchestration layer Karamel has abstracted out the cloud provider and it has a unified API to handle different Cloud Providers. Underneath it uses Apache jclouds® API to communicate with the known cloud providers such as Amazon AWS, Google Compute Engine and OpenStack. Intuitively the following steps need to be taken to make EGI cloud available in Karamel:

1. Implement EGI API in Apache jclouds® project. This API should come with configurable resource provisioning functions, ssh key configuration support and handling exceptional situations such as repeating mechanism for failures.

2. The launcher (similar to Ec2Launcher class in Karamel) must have all the required phases of the cluster like pre-cleaning, forkgroups, forkmachines, purge. This class should be tested in isolation by mocking the access into EGI and also with access into it.

3. ClusterManager class in Karamel must be aware of new cloud type.

4. Add EGI in the cluster definition language and Karamel UI.

9.8 Application on EGI Federated Cloud Platform

In order to utilize existing architecture and software components built in the context of the EGI Federated Cloud initiative, all existing provisioning and orchestration tools of the Biobank Cloud platform have to support a specific communication protocol, namely OGF's Open Cloud Computing Interface (OCCI), used for unified compute resource provisioning in a heterogeneous environment. Introducing this support will ensure cross-platform compatibility and future extensibility of the proposed solution.

The primary extension target is the Karamel tool, serving as the dynamic orchestration and cluster management solution for various computing platforms used in the context of the BBMRI-ERIC project. By providing a so-called “launcher” component for OCCI, Karamel will be able to dynamically instantiate, provision and configure whole purpose-built on-demand compute clusters in any OCCI-enabled cloud platform.

10 Secure Storage Architectural Design

BioBankCloud provides a scalable storage service called HopsFS, see Figure 8. HopsFS is a distributed file system that scales to store petabyes of data, and HopsFS is a drop-in replacement for the Apache Hadoop Distributed Filesystem (HDFS). In BiobankCloud, HopsFS is primarly used to store genomic data. Genomic data is organized in DataSets accessible to different Studies. DataSet consists of a related group of directories, files, and metadata. To allow for access control of users to DataSets, which is not inherent in the DataSet concept, we introduce the notion of Studies. A Study is a grouping of researchers and DataSets with role-based access control where different researchers can be given different access rights to DataSets. Our storage model supports multitenancy where the Studies are completely isolated from each other. DataSets can be shared between Studies (when the necessary security, legal, and ethical conditions for sharing are in place) without violating the isolation of Studies for other DataSets. The point of interaction between Biobanker, Bio-informatician and the BiobankCloud is a HopsWorks and integrates all the software components from BiobankCloud.

10.1 Deploying BiobankCloud Storage

BiobankCloud storage service is provided by HopsFS, which is deployed using Karamel as discussed in section 9. Karamel and HopsWorks enable non-sophisticated users to deploy BiobankCloud on cloud infrastructures or on-premises, and immediately be able to use the software to curate data (Biobankers) or run workflows (Bioinformaticians), while storing petabytes of data in secure, isolated studies. HopsFS is a POSIX like distributed file system that stores the data in files organized in hierarchical folders. HopsFS uses Unix like file permissions to isolate users and their data. However, HopsFS is not fully POSIX compliant as it is an append only file system and it does not support random updates in a file. Currently HopsFS does not support federation, that is, each HopsFS cluster is independent and it does not support sharing data across different HopsFS clusters.

10.2 HopsFS

The storage model is centered around the files and folders that contain data, and with which metadata is associated. These files are stored in HopsFS [98], a scalable and highly available implementation of the Hadoop Distributed File System (HDFS) [99]. HopsFS offers all the basic functionality of a file system: storage and retrieval of files and a hierarchical directory structure in which to organize them. It also offers access control based on file and directory permissions. Files are replicated to minimize the chance of data loss. In HopsFS, contrary to Apache HDFS, the namenode is not a single machine that contains all the state in the system. Rather, there are multiple stateless namenodes which store state in a in-memory distributed database, MySQL cluster [100]. This eliminates drawback of having a single point of failure in the namenode and also takes away scalability concerns when too many files are stored in the system. Apart from the state of the file system, we also store the metadata, which is an essential part of our object model, in the MySQL cluster database. This allows us to maintain referential integrity between the metadata and files; metadata for non-existing files will never be occur.

10.3 Studies, DataSets and HopsFS

The entire access control system enforced by HopsFS. So in order to understand our implementation, we first need to clarify how Studies and DataSets are represented in HopsFS.

Both Studies and DataSets are fundamentally represented by subtrees in HopsFS. That is, they consist of one dedicated folder and all its children, i.e. all the folders and files it contains. The subtree may be of arbitrary depth.

The root of a Study subtree, which we call the *Study base directory* or the *Study root directory*, is a folder whose name is the same as the Study's name. (Note that this implies that Study names have to be unique in the entire system.) Study root directories are always created as a direct child of the */studies/* directory; the Study root directory is created upon Study creation. A user always operates in the context of a Study. This means that a user can be completely oblivious to the structure of the file system outside the Study subtree. Moreover, when a user is working within a Study, (s)he should be unable to access the subtree of any other Study.

The root of a DataSet subtree, analogously called the *DataSet base directory* or *DataSet root directory*, is a folder named after the DataSet name and is a direct child of the root directory of the Study to which it belongs. More, every direct child directory of the Study root directory is considered to be a DataSet. This implies that a DataSet name must be unique within the Study it belongs to. Other than that, there are no restrictions on the amount of DataSets that can be created within a Study.

An example of the resulting directory structure is shown in Figure 10. There are *N* Studies in the system ($study_1$ through $study_N$, and four DataSets ($dataset_A$ through $dataset_D$).

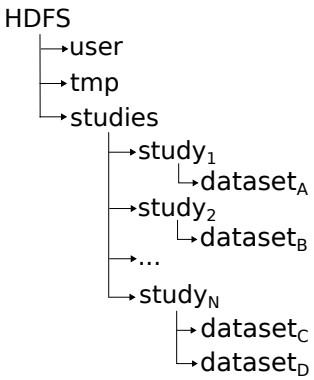

Figure 10: HopsWorks folder structure in HopsFS

10.3.1 DataSets

A Dataset in Hops:

1. is a directory in HopsFS and all the files and directories in its subgraph;

2. is any HopsFS directory that is a direct child of a study base directory;

3. contains basic metadata and optionally extended metadata that is associated with the DataSet's directory;

4. has a single owner with read/write privileges;

5. is readable by all other members of the study;

6. may be shared with other projects (remote projects).

The main use-cases for DataSets are search and sharing. Searching can be either at the level of searching for a DataSet (DataSet discovery) or searching within a DataSet (File Discovery). These use cases are analogous to a Sample Collection Availability Service (DataSet discovery) and Sample Availability Service (File Discovery).

1. DataSet Discovery: free-text search for what DataSets are available within the cluster. In BiobankCloud, this is equivalent to a sample collection discovery service.

2. File Discovery: free-text search for files or directories within DataSets local to a project. In BiobankCloud, this is equivalent to a service for searching for individual samples.

3. DataSet Browsing from within a Project (similar to a file-browser). In BiobankCloud, this is equivalent to browsing the catalog of samples and sample collections.

10.3.2 Studies

A Study is a grouping of DataSets and users, as illustrated in Figure 11, that also integrates a role-based access control mechanism. Users are granted different permissions on the DataSets in the Study based on their role in it. The different study-level roles are as follows:

- Data Provider (BBC_ADMIN): can add data and members to the Study.

- Researcher (BBC_RESEARCHER): can process the data in the platform through running workflows.

The creator of a Study, who is also its owner, is always assigned the Data Provider role. Additional users can be assigned both roles.

Each study can contain zero or more DataSets. From HopsFS' perspective, a DataSet is a directory within the study base directory that has one owner with read/write privileges and, depending on the type of DataSet, a group of users with either only read privileges, or both read and write privileges. A DataSet is associated with metadata, where the metadata is either:

Figure 11: Projects are groupings of DataSets and Users.

- minimal metadata (a free-text description and whether the DataSet's metadata should be discoverable by users outside the project) OR
- based on a metadata template (e.g., a Next-Generation Sequencing DataSet).

Metadata is used primarily to enable free-text search for DataSets and files within DataSets. A DataSet is implemented in HopsFS as a directory located in the base directory of the study along with a set of tables in the database that are subsequently exported to Elasticsearch for search functionality.

10.4 Multitenancy in BiobankCloud

Our access control model is based on three requirements. The first is that Studies are completely isolated from each other; a user operating in Study A should not be able to use data from Study B. The second requirement allows for a deviation from this rule: DataSets should be shareable with other Studies. But this action of course should not violate the isolation of Studies for other DataSets. The third requirement states that the Study-level roles are enforced. We discuss these requirements in a bit more detail in the following sections.

10.4.1 Isolation of Studies

The access control model should guarantee the integrity of Studies as isolated entities of authorization. Concretely, when a user A has been granted access to files in both Study X and Y, this does not imply that user A should have the right to use data from both Study X and Y in the same experiment. The vanilla HDFS permission scheme cannot guarantee this. To see why, let's consider how this would typically be implemented using HDFS permissions.

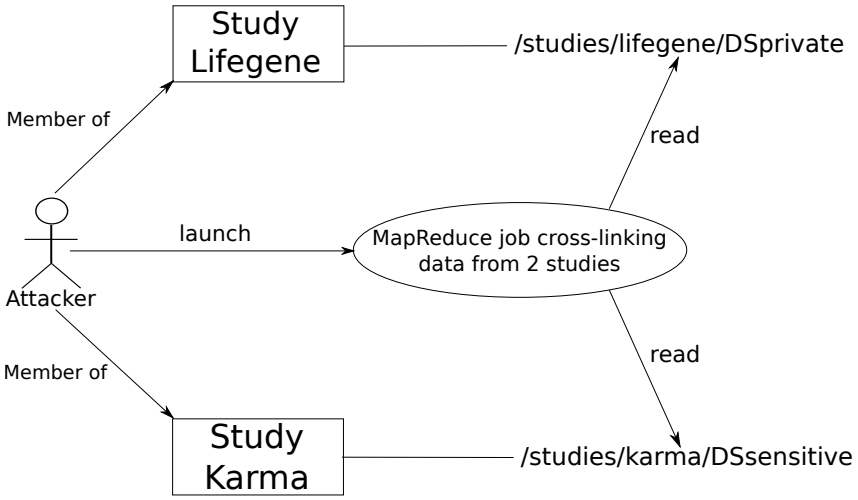

Figure 12: Access control is unsatisfactory with plain HDFS permissions: the attacker can mix data from both studies.

Since a user can be a member of many groups and an inode has a group associated with it, the natural solution would be to create a new group per Study. Members of the Study become member of the new group and all the files in the Study subtree are associated with this group. As an example, let's say an attacker who is a regular user of the platform, is a member of the Study *lifegene*. In HDFS, this translates into the subtree with root folder */studies/lifegene* where all files have the group `lifegene` and the attacker being a member of the group `lifegene`. If the attacker now also gains membership of the Study *karma* and hence group `karma`, she has access to all files in both the subtree for the Study *lifegene* and the Study *karma*.

Now, consider the situation illustrated in Figure 12. In her capacity as a member of the Study *karma*, the attacker launches a MapReduce job. This MapReduce job reads from both a DataSet in the Study *lifegene* and a DataSet in the Study *karma*. Because the attacker is a member of both groups `lifegene` and `karma`, this is allowed according to the HDFS permissions model.

Hence, if permissions for both Studies are based on the same HDFS user identity, there are no mechanisms to prevent users from writing applications that cross-link data in different Studies. It is clear that the simple user-group scheme is not adequate for our access control. In the next section, we discuss our solution of the problem. In short, for each user we create a new per-study HDFS user for each Study the user is a member of.

10.4.2 Shareable DataSets

DataSets are the finest grain of sharing in our platform. Suppose we have a Study *S* with two DataSets, DataSet *A* and DataSet *B*. We then want to be able to share DataSet *A* with another Study *T*, while keeping DataSet *B* private to Study *S*. The only way to do this is to add all users of Study *T* to the group

of DataSet *A*, but not of DataSet *B*. Hence, it is clear that we need a finer grouping level than the Study level. We solve this by creating a new group per DataSet.

10.4.3 Enforcing role permissions

Study members can have one of two roles: data provider or researcher. Data providers can create DataSets, while researchers can only read them. On the file system level, this translates into data providers being able to create subfolders in the Study root directory, while researchers should be able to list all files in the study root directory, and read the different DataSets.

Mapping this to the concepts of owner, group, and world permissions is done as follows. First, the owner, i.e. creator of the study, has all permissions on the folder. Second, the group has read and write permissions; this maps to data providers. Third, the world permissions allow to read the contents of the folder. This allows researchers to list the available DataSets in a specific Study. This entails creating a HopsFS group for each Study to contain its data providers.

The point of interaction between Biobanker, Bio-informatician and the BiobankCloud is a HopsWorks. To overcome the problem of users being able to cross-link Studies, we have to do away with the single HopsFS user identity per HopsWorks user. Instead, for each HopsWorks user, we create a per-Study HopsFS user identity for each Study the user is a member of. A HopsWorks user will always interact with HopsFS in the capacity of his/her per-Study identity. This implies that **a user identity in the HopsWorks has no privileges whatsoever in HopsFS; only the per-Study identies have HopsFS privileges**. For each interaction with HopsFS, the HopsWorks intercepts the operation and determines the HopsFS identity to use based on the logged in user and his/her active Study. The HopsWorks then passes the operation to HopsFS as the newly determined user.

With each user having a per-Study HopsFS user identity for each Study (s)he is a member of, each user has as many HopsFS user identities as the number of Studies (s)he is a member of. However, this solution does not allow for DataSets to be shared. To enable this, we need to define per-dataset groups as well. Finally, to enforce the data provider and researcher roles, we need to create a new role per Study.

References

[1] M. Howard and S. Lipner. *The security development lifecycle: SDL-A process for developing demonstrably more secure software*. 2006.

[2] M. Deng, K. Wuyts, R. Scandariato, B. Preneel, and W. Joosen. "A privacy threat analysis framework: supporting the elicitation and fulfillment of privacy requirements". In: *Requirements Engineering* 16.1 (2011), pp. 3–32.

[3] M. Procházka, S. Licehammer, and L. Matyska. "Perun—Modern approach for user and service management". In: *IST–Africa Conference Proceedings, 2014*. IEEE. 2014, pp. 1–11.

[4] *Health informatics – Pseudonymization*. ISO/TS 25237:2008. 2008.

[5] M. Linden, T. Nyrönen, and I. Lappalainen. "Resource Entitlement Management System". In: *TERENA Networking Conference 2013 (TNC2013)*. June 2013. URL: http://tnc2013.terena.org/getfile/870.

[6] S. Cantor and T. SCAVO. "Shibboleth architecture". In: *Protocols and Profiles* 10 (2005), p. 16.

[7] T. Barton, J. Basney, T. Freeman, T. Scavo, F. Siebenlist, V. Welch, R. Ananthakrishnan, B. Baker, M. Goode, and K. Keahey. "Identity federation and attribute-based authorization through the globus toolkit, shibboleth, gridshib, and myproxy". In: *5th Annual PKI R&D Workshop*. Vol. 4. 2006.

[8] N. van Dijk. "Virtual Organization - as a Service". In: *AARC kickoff*. June 2015. URL: https://aarc-project.eu/wp-content/uploads/2015/05/20150603-AARC_KICKOFF-Virtual-Organization-as-a-Service.pdf.

[9] N. van Dijk. "VOPaaS: Virtual Organisation Platform as a Service". In: *Internet2 2015 Technology Exchange*. Oct. 2015. URL: https://aarc-project.eu/wp-content/uploads/2015/05/20150603-AARC_KICKOFF-Virtual-Organization-as-a-Service.pdf.

[10] J. P. Ioannidis, D. B. Allison, C. A. Ball, I. Coulibaly, X. Cui, A. C. Culhane, M. Falchi, C. Furlanello, L. Game, G. Jurman, et al. "Repeatability of published microarray gene expression analyses". In: *Nature genetics* 41.2 (2009), pp. 149–155.

[11] F. Prinz, T. Schlange, and K. Asadullah. "Believe it or not: how much can we rely on published data on potential drug targets?" In: *Nature reviews Drug discovery* 10.9 (2011), p. 712.

[12] C. G. Begley and L. M. Ellis. "Drug development: Raise standards for preclinical cancer research". In: *Nature* 483.7391 (2012), pp. 531–533.

[13] M. Bissell. "Reproducibility: The risks of the replication drive". In: *Nature* 503 (2013), pp. 333–334.

[14] S. J. Morrison. "Time to do something about reproducibility". In: *eLife* 3 (2014), e03981.

[15] P. AC't Hoen, M. R. Friedländer, J. Almlöf, M. Sammeth, I. Pulyakhina, S. Y. Anvar, J. F. Laros, H. P. Buermans, O. Karlberg, M. Brännvall, et al. "Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories". In: *Nature biotechnology* 31.11 (2013), pp. 1015–1022.

[16] M. A. Bishop. *The Art and Science of Computer Security*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2002. ISBN: 0201440997.

[17] R. Bild, F. Kohlmayer, S. Brunner, K. Kuhn, B. Rodriguez, A. Lamichhane, S. Klein, M. Raess, C. Lengger, P. Gormanns, C. Ohmann, W. Kuchinke, U. Sarkans, M. Sariyar, C. Morris, T. Nyrönen, and J. Leinonen. *Report describing the security architecture and framework*. BioMedBridges Project, Deliverable 5.3. June 2014. URL: http://www.biomedbridges.eu/sites/biomedbridges.eu/files/documents/deliverables/d5-3_report_describing_the_security_architecture_and_framework-formatted_pdf.pdf.

[18] W. P. Stevens, G. J. Myers, and L. L. Constantine. "Structured design". In: *IBM Systems Journal* 13.2 (1974), pp. 115–139.

[19] *Information technology - Security techniques - Information security management systems - Overview and vocabulary*. ISO 27000. May 2009.

[20] R. W. Shirey. *Internet Security Glossary*. RFC 4949. IETF, Aug. 2007, pp. 1–365. URL: https://www.rfc-editor.org/rfc/rfc4949.txt.

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

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

[23] R. Chadwick and K. Berg. "Solidarity and equity: new ethical frameworks for genetic databases". In: *Nature Reviews Genetics* 2.4 (2001), pp. 318–321.

[24] A. T. Ewing, L. A. Erby, J. Bollinger, E. Tetteyfio, L. J. Ricks-Santi, and D. Kaufman. "Demographic Differences in Willingness to Provide Broad and Narrow Consent for Biobank Research". In: *Biopreservation and biobanking* 13.2 (2015), pp. 98–106.

[25] M. G. Hansson, J. Dillner, C. R. Bartram, J. A. Carlson, and G. Helgesson. "Should donors be allowed to give broad consent to future biobank research?" In: *The Lancet Oncology* 7.3 (2006), pp. 266–269.

[26] K. Hoeyer. "Informed consent: the making of a ubiquitous rule in medical practice". In: *Organization* 16.2 (2009), pp. 267–288.

[27] H. Williams, K. Spencer, C. Sanders, D. Lund, E. A. Whitley, J. Kaye, and W. G. Dixon. "Dynamic consent: a possible solution to improve patient confidence and trust in how electronic patient records are used in medical research". In: *JMIR medical informatics* 3.1 (2015).

[28] G. Fletcher, H. Lockhart, S. Anderson, J. Bohren, R. Philpott, C. Canales-Valenzuela, L. Thiyagarajan, A. Nadalin, S. Cantor, B. Morgan, E. Tiffany, T. Scavo, P. Davis, J. Hodges, F. Hirsch, A. Barbir, P. Madsen, P. Mishra, B. Campbell, A. Saldhana, E. Maler, E. Xu, K. Spaulding, and D. Staggs. *Identity Provider Discovery Service Protocol and Profile*. Committee Specification 01. Mar. 2008. URL: http://docs.oasis-open.org/security/saml/Post2.0/sstc-saml-idp-discovery.pdf.

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

[30] W. E. Burr, D. F. Dodson, E. M. Newton, R. A. Perlner, W. T. Polk, S. Gupta, and E. A. Nabbus. *Electronic Authentication Guideline*. NIST Special Publication 800-63-2. Aug. 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.

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

[32] 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". In: *OJ L 257, 28.8.2014, p. 73–114* (2014).

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

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

[35] S. Kumar, D. Walker, A. West, D. L. D. Fisher, M. Smith, C. Tompkins, T. Barton, M. Dunker, W. Curry, S. Carmody, and T. Scavo. *Assurance Enhancements for the Shibboleth Identity Provider Draft v17*. Apr. 2013. URL: https://spaces.internet2.edu/download/attachments/9185/AssuranceReqShibIdPv17.pdf.

[36] D. Recordon, M. Jones, J. Bufu, J. Daugherty, and N. Sakimura. *OpenID Provider Authentication Policy Extension 1.0*. Dec. 2008. URL: http://openid.net/specs/openid-provider-authentication-policy-extension-1_0.html.

[37] F. Brosch, H. Koziolek, B. Buhnová, and R. Reussner. "Parameterized Reliability Prediction for Component-Based Software Architectures". English. In: *Research into Practice: Reality and Gaps*. Ed. by G. T. Heineman, J. Kofroň, and F. Plašil. Vol. 6093. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2010, pp. 36–51. ISBN: 978-3-642-13820-1. DOI: 10.1007/978-3-642-13821-8_5. URL: http://dx.doi.org/10.1007/978-3-642-13821-8_5.

[38] T. Orawiwattanakul, K. Yamaji, M. Nakamura, T. Kataoka, and N. Sonehara. "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, pp. 243–249.

[39] 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". In: *Official Journal of the EC* 23.6 (1995).

[40] EU Directive. "Directive 1999/93/EC of the European Parliament and of the Council on a Community framework for electronic signatures". In: *Official Journal L* 13 (1999).

[41] 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.

[42] 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". In: *JL 201, 31.7. 2002, at 37.(Directive on Privacy and Electronic Communications)* (2002).

[43] A. Bessani, J. Brandt, M. Bux, V. Cogo, L. Dimitrova, J. Dowling, A. Gholami, K. Hakimzadeh, M. Hummel, M. Ismail, 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.

[44] M. Bux, J. Brandt, C. Lipka, K. Hakimzadeh, J. Dowling, and U. Leser. “SAASFEE: Scalable Scientific Workflow Execution Engine”. In: *Proc. VLDB Endow.* 8.12 (Aug. 2015), pp. 1892–1895. ISSN: 2150-8097. DOI: 10.14778/2824032.2824094. URL: http://dx.doi.org/10.14778/2824032.2824094.

[45] *Information technology – Security techniques – Privacy framework*. ISO/TS 29100:2011. 2011.

[46] S. M. Fullerton, N. R. Anderson, G. Guzauskas, D. Freeman, and K. Fryer-Edwards. “Meeting the governance challenges of next-generation biorepository research”. In: *Science Translational Medicine* 2.15 (2010), pp. 15cm3–15cm3.

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

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

[49] T. Li, N. Li, J. Zhang, and I. Molloy. “Slicing: A new approach for privacy preserving data publishing”. In: *Knowledge and Data Engineering, IEEE Transactions on* 24.3 (2012), pp. 561–574.

[50] M. E. Nergiz, M. Atzori, and C. Clifton. “Hiding the presence of individuals from shared databases”. In: *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*. ACM. 2007, pp. 665–676.

[51] L. Sweeney. “k-anonymity: A model for protecting privacy”. In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10.05 (2002), pp. 557–570.

[52] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. “l-diversity: Privacy beyond k-anonymity”. In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1.1 (2007), p. 3.

[53] N. R. Adam and J. C. Worthmann. “Security-control methods for statistical databases: a comparative study”. In: *ACM Computing Surveys (CSUR)* 21.4 (1989), pp. 515–556.

[54] D. E. Denning, P. J. Denning, and M. D. Schwartz. “The tracker: A threat to statistical database security”. In: *ACM Transactions on Database Systems (TODS)* 4.1 (1979), pp. 76–96.

[55] B. Zhou, J. Pei, and W. Luk. “A brief survey on anonymization techniques for privacy preserving publishing of social network data”. In: *ACM SIGKDD Explorations Newsletter* 10.2 (2008), pp. 12–22.

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

[57] 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 11/11/2015).

[58] A. De Waal and L. Willenborg. “A view on statistical disclosure control for microdata”. In: *Survey Methodology* 22 (1996), pp. 95–103.

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

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

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

[62] 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 11/11/2015).

[63] 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 11/11/2015).

[64] E. Erdem and S. I. Prada. "Creation of public use files: lessons learned from the comparative effectiveness research public use files data pilot project". In: *2011 JSM Proceedings* (2011), pp. 4095–4109. URL: http://mpra.ub.uni-muenchen.de/35478/ (visited on 11/11/2015).

[65] *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 11/11/2015).

[66] K. El Emam, D. Paton, F. Dankar, and G. Koru. "De-identifying a public use microdata file from the Canadian national discharge abstract database". In: *BMC Med Inform Decis Mak* 11 (2011), p. 53.

[67] K. El Emam, L. Arbuckle, G. Koru, B. Eze, L. Gaudette, E. Neri, S. Rose, J. Howard, and J. Gluck. "De-identification methods for open health data: the case of the Heritage Health Prize claims dataset". In: *J. Med. Internet Res.* 14.1 (2012), e33.

[68] V. Curcin, S. Miles, R. Danger, Y. Chen, R. Bache, and A. Taweel. "Implementing interoperable provenance in biomedical research". In: *Future Generation Computer Systems* 34 (2014), pp. 1–16.

[69] S. Jajodia, S. Noel, and B. O'Berry. "Topological Analysis of Network Attack Vulnerability". English. In: *Managing Cyber Threats*. Ed. by V. Kumar, J. Srivastava, and A. Lazarevic. Vol. 5. Massive Computing. Springer US, 2005, pp. 247–266. ISBN: 978-0-387-24226-2. DOI: 10.1007/0-387-24230-9_9. URL: http://dx.doi.org/10.1007/0-387-24230-9_9.

[70] R. Barnes, M. Thomson, A. Pironti, and A. Langley. *Deprecating Secure Sockets Layer Version 3.0*. RFC 7568. IETF, June 2015, pp. 1–7. URL: https://www.rfc-editor.org/rfc/rfc7568.txt.

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

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

[73] M. Wolfson, S. E. Wallace, N. Masca, G. Rowe, N. A. Sheehan, V. Ferretti, P. LaFlamme, M. D. Tobin, J. Macleod, J. Little, et al. “DataSHIELD: resolving a conflict in contemporary bioscience—performing a pooled analysis of individual-level data without sharing the data”. In: *International journal of epidemiology* (2010), dyq111.

[74] E. Jones, N. Sheehan, N. Masca, S. Wallace, M. Murtagh, and P. Burton. “DataSHIELD–shared individual-level analysis without sharing the data: a biostatistical perspective”. In: *Norsk epidemiologi* 21.2 (2012).

[75] A. Gaye, Y. Marcon, J. Isaeva, P. LaFlamme, A. Turner, E. M. Jones, J. Minion, A. W. Boyd, C. J. Newby, M.-L. Nuotio, et al. “DataSHIELD: taking the analysis to the data, not the data to the analysis”. In: *International journal of epidemiology* 43.6 (2014), pp. 1929–1944.

[76] A. Cambon-Thomsen, G. A. Thorisson, L. Mabile, S. Andrieu, G. Bertier, M. Boeckhout, J. Carpenter, G. Dagher, R. Dalgleish, M. Deschênes, et al. “The role of a bioresource research impact factor as an incentive to share human bioresources”. In: *Nature Genetics* 43.6 (2011), pp. 503–504.

[77] L. Mabile, R. Dalgleish, G. A. Thorisson, M. Deschênes, R. Hewitt, J. Carpenter, E. Bravo, M. Filocamo, P. A. Gourraud, J. R. Harris, et al. “Quantifying the use of bioresources for promoting their sharing in scientific research”. In: *Gigascience* 2.1 (2013), p. 7.

[78] C. Dwork. “Differential privacy: A survey of results”. In: *Theory and applications of models of computation*. Springer, 2008, pp. 1–19.

[79] N. Li, W. H. Qardaji, and D. Su. “Provably private data anonymization: Or, k-anonymity meets differential privacy”. In: *CoRR, abs/1101.2604* 49 (2011), p. 55.

[80] C. Dwork and A. Roth. “The algorithmic foundations of differential privacy”. In: *Theoretical Computer Science* 9.3-4 (2013), pp. 211–407.

[81] C. C. Aggarwal. “On k-anonymity and the curse of dimensionality”. In: *Proceedings of the 31st international conference on Very large data bases*. VLDB Endowment. 2005, pp. 901–909.

[82] G. Cuccuru, S. Leo, L. Lianas, M. Muggiri, A. Pinna, L. Pireddu, P. Uva, A. Angius, G. Fotia, and G. Zanetti. “An automated infrastructure to support high-throughput bioinformatics”. In: *High Performance Computing & Simulation (HPCS), 2014 International Conference on*. IEEE. 2014, pp. 600–607.

[83] G. Zanetti. *Data intensive biology and data provenance graphs*. BiobankCloud Project. Feb. 2015. URL: http://www.biobankcloud.com/sites/default/files/ngshadoop/hadoop_ngs_zanetti.pdf.

[84] A. DiezFraile, P. Holub, M. Hummel, K. Kuhn, J.-E. Litton, H. Müller, P. Quinlan, M. Swertz, F. Ückert, D. Valík, O. Vojtíšek, and G. Zanetti. *BBMRI-ERIC Use Cases*. 2015.

[85] A. Bessani, R. Mendes, V. Cogo, T. Oliveira, N. Neves, and R. Fonseca. *D4.2: The Overbank Cloud Architecture, Protocols and Middleware*. BiobankCloud Deliverable. Nov. 2014. URL: http://www.biobankcloud.com/sites/default/files/deliverables/2014/D4.2-final.pdf.

[86] Á. Frohner, T. Aspelien, J. Montagnat, D. Jouvenot, and C. Pera. “Encrypted Data Storage in EGEE”. In: (2006). URL: https://indico.cern.ch/event/430389/session/s2/contribution/s2t13/attachments/929813/1316802/EGEE-JRA1-All-Hands-EDS-MDM-20060711.pdf.

[87] K. Maheshwari, C. Goble, P. Missier, and J. Montagnat. “Medical image processing workflow support on the EGEE grid with taverna”. In: *Computer-Based Medical Systems, 2009. CBMS 2009. 22nd IEEE International Symposium on*. IEEE. 2009, pp. 1–7.

[88] Á. Frohner, J.-P. Baud, R. M. G. Rioja, G. Grosdidier, R. Mollon, D. Smith, and P. Tedesco. “Data management in EGEE”. In: *Journal of Physics: Conference Series*. Vol. 219. 6. IOP Publishing. 2010, p. 062012.

[89] I. Magnin and J. Montagnat. “The grid and the biomedical community: Achievements and open issues”. In: *EGEE User Forum, CERN, Geneva, Switzerland*. 2006.

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

[91] T. E. Graber, J. Kopelman, E. H. Watkeys III, and M. I. Weinberger. *Method and apparatus for tracking the navigation path of a user on the world wide web*. US Patent 5,717,860. Feb. 1998.

[92] T. Damico, J. Kopelman, S. F. Wamoglu, and M. I. Weinberger. *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. Oct. 1998.

[93] M. I. Ingrassia Jr, J. A. Shelton, and T. M. Rowland. *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. Mar. 2000.

[94] J. E. Allard, D. R. Treadwell III, and J. F. Ludeman. *Method, system and apparatus for client-side usage tracking of information server systems*. US Patent 6,018,619. Jan. 2000.

[95] D. Oberle, B. Berendt, A. Hotho, and J. Gonzalez. “Conceptual user tracking”. In: *Advances in Web Intelligence*. Springer, 2003, pp. 155–164.

[96] R. Atterer, M. Wnuk, and A. Schmidt. “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, pp. 203–212.

[97] S. Nelson-Smith. *Chef: The Definitive Guide*. O’Reilly Media, Inc., 2013.

[98] K. Hakimzadeh, H. Peiro Sajjad, and J. Dowling. “Scaling HDFS with a Strongly Consistent Relational Model for Metadata”. English. In: *Distributed Applications and Interoperable Systems*. Ed. by K. Magoutis and P. Pietzuch. Vol. 8460. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2014, pp. 38–51. ISBN: 978-3-662-43351-5. DOI: 10.1007/978-3-662-43352-2_4. URL: http://dx.doi.org/10.1007/978-3-662-43352-2_4.

[99] K. Shvachko, H. Kuang, S. Radia, and R. Chansler. “The Hadoop Distributed File System”. In: *Mass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium on*. May 2010, pp. 1–10. DOI: 10.1109/MSST.2010.5496972.

[100] O. Corporation. *MySQL Cluster NDB 7.2*. https://dev.mysql.com/doc/refman/5.5/en/mysql-cluster.html. [Online; Accessed: 22-05-2015].