

Deliverable 5.3

Project Title:	Building data bridges between biological and medical infrastructures in Europe	
Project Acronym:	BioMedBridges	
Grant agreement no.:	284209	
	Research Infrastructures, FP7 Capacities Specific Programme; [INFRA-2011-2.3.2.] "Implementation of common solutions for a cluster of ESFRI infrastructures in the field of "Life sciences"	
Deliverable title:	Report describing the security architecture and framework	
WP No.	5	
Lead Beneficiary:	7: Klinikum rechts der Isar der Technischen Universitaet Muenchen (TUM-MED)	
WP Title	Secure access	
Contractual delivery date:	30 June 2014	
Actual delivery date:	30 June 2014	
WP leader:	5: Christian Ohmann 7: Klaus Kuhn	5: UDUS 7: TUM-MED
Contributing partner(s):	1: EMBL, 4: STFC, 5: UDUS, 6: FVB, 7: TUM-MED, 9: ErasmusMC, 10:TMF, 11: HMGU, 14: INSERM	

Authors and contributors: Raffael Bild, Florian Kohlmayer, Sabine Brunner, Klaus Kuhn, Benedicto Rodriguez, Ashish Lamichhane, Stefan Klein, Michael Raess, Christoph Lengger, Philipp Gormanns, Christian Ohmann, Wolfgang Kuchinke, Ugis Sarkans, Murat Sariyar, Chris Morris, Tommi Nyrönen, Jaakko Leinonen



Contents

Figures	3
Tables	4
List of Abbreviations	5
1 Executive summary	6
2 Project objectives	8
3 Background.....	9
3.1 Related work and projects	9
4 Security and privacy concepts and definitions.....	14
4.1 Threat, vulnerability, and risk	14
4.2 Security threat modelling using STRIDE.....	16
4.2.1 Threats and security properties addressed by STRIDE	17
4.3 Privacy threat modelling using LINDDUN	18
4.3.1 Threats and privacy properties addressed by LINDDUN.....	19
5 Modelling a generic data bridge.....	21
6 Security requirements for a e-infrastructure addressing the use cases	24
6.1 Security requirements for the research infrastructures	25
6.2 Security requirements for the use case work packages	26
6.2.1 Overview of data flow diagrams	28
6.2.2 Survey questionnaire for the Use Case Work Packages.....	29
6.2.3 Survey results for use case work packages.....	30
6.2.3.1 Custom extensions to notation of data flow diagrams	31
6.2.3.2 Work package 6 “Imaging Bridge”	32
6.2.3.3 Work package 7 “PhenoBridge”	34
6.2.3.4 Work package 8 “Personalized Medicine”	36
6.2.3.5 Work package 10 “Biological Sample Data Integration”	38
6.2.4 Work package 9 “Structural Data Bridge”	40
6.2.4.1 Work package 9 workflow diagram	41
6.2.4.2 Data Bridge 1: From Researcher to INSTRUCT	43
6.2.4.3 Data Bridge 2: From ELIXIR to INSTRUCT/Researcher	43
6.2.4.4 Data Bridge 3: From INSTRUCT/Researcher to ELIXIR (EMDB) ..	44
6.3 Summary.....	45
7 Threat and risk analysis for sharing data or biomaterials.....	45
7.1 Methodology.....	45
7.2 Risk assessment results	49
8 Design of the security architecture and framework	54
8.1 Overview of the security architecture and framework.....	55
8.2 Countermeasures of the security architecture and framework	59
8.2.1 Authentication	59
8.2.2 Authorization	60
8.2.3 Secure data communication.....	62



8.2.4	Encryption of data	63
8.2.5	Anonymization.....	64
8.2.6	Pseudonymization	67
8.2.7	Auditing and provenance	67
8.2.8	Common data access policies	69
8.3	Secure workflow specified by the security architecture	71
8.3.1	Open data access tier	72
8.3.2	Restricted data access tier.....	72
8.3.3	Committee-controlled data access tier.....	73
9	Steps from continuous feedback to implementation.....	78
9.1	Reviews and feedback on interim progress	78
9.2	Implementation of the security architecture by use case work packages...	80
10	Delivery and schedule	82
11	Adjustments made	82
12	Appendices.....	83
12.1	Result Tables of Threat and Risk Analysis	83
12.1.1	Work package 6 threat and risk analysis results	83
12.1.2	Work package 7 threat and risk analysis results	86
12.1.3	Work package 8 threat and risk analysis results	87
12.1.4	Work package 10 threat and risk analysis results	90
13	Background information	93
14	References	98

Figures

Figure 1.	Relationship between vulnerabilities, threats, risks, and safeguards [6]	16
Figure 2.	Data flow in BioMedBridges Data Bridges from deliverable 5.2.....	22
Figure 3.	DFD of WP6 that resulted from the answers to the survey questionnaire ..	34
Figure 4.	DFD for WP7 "Phenobridge".....	36
Figure 5.	DFD for WP8 "Personalized Medicine"	38
Figure 6.	DFD for WP10 "Biological Sample Data Integration"	39
Figure 7.	Activity diagram for the Open Data Access Tier.....	75
Figure 8.	Activity diagram for the Restricted Data Access Tier	75
Figure 9.	Activity diagram for the Committee Controlled Data Access Tier.....	77



Tables

Table 1. Data security areas of “WP5 Survey 1” questionnaire.....	25
Table 2. WP5 Survey 1. Use/Sharing of individual level-data	26
Table 3. Components of a Data Flow Diagram.....	28
Table 4. Questionnaire of the survey distributed among the use case WPs.	29
Table 5. Custom extensions to DFD notation to represent security properties.	32
Table 6. Summary of answers to survey questionnaire by WP6.	33
Table 7. Summary of answers to survey questionnaire by WP7.	35
Table 8. Summary of answers to survey questionnaire by WP8.	37
Table 9. Summary of answers to survey questionnaire by WP10.	39
Table 10. Mapping STRIDE security threats and countermeasures to DFD element types (see Tables 9-5 and 9-8 in Chapter 9 of [8])	47
Table 11. Mapping LINDDUN privacy threats and objectives to DFD element types (see Tables 4 and 6 in [7])	47
Table 12. Template used to report the risk assessment of threats for BioMedBridges (cf. Table I-5 in Appendix I of [4]).....	48
Table 13. A qualitative measure of risk assessment.	49
Table 14. Risk assessment for threats (STRIDE and LINDDUN) to the “Data Flow” element of the DFD.	52
Table 15. Risk assessment for security (STRIDE) threats to the “Data Store”, “Process”, and “Entity” elements of the DFD associated to the Use Case WPs.	52
Table 16. Risk assessment for privacy (LINDDUN) threats to the “Data Store”, “Process”, and “Entity” elements of the DFD associated to the Use Case WPs.	53
Table 17. Components of the workflow activity diagrams.....	71
Table 18. Legend used by tables to report the threat and risk assessment results. ...	83
Table 19. Threat and risk assessment results for use case WP6.	83
Table 20. Threat and risk assessment results for use case WP7.	86
Table 21. Threat and risk assessment results for use case WP8.	87
Table 22. Threat and risk assessment results for use case WP10. The REMS is explicitly mentioned here. It is suggested to use it for all UC WPs.....	90



List of Abbreviations

AES	Advanced Encryption Standard
BBMRI	Biobanking and Biomolecular Resources Research Infrastructure
BMB	BioMedBridges
BMS	Biological and Medical Sciences
DAC	Data Access Committee
DAC	Discretionary Access Control
DFD	Data Flow Diagram
DoS	Denial of Service
DoW	Description of Work
DUA	Data Use Agreement
EC	Ethic Committee
EATRIS	European Advanced Translational Research Infrastructure in Medicine
ECRIN	European Clinical Research Infrastructures Network
EGA	European Genome-phenome Archive
EGI	European Grid Infrastructure
ELIXIR	European Life Sciences Infrastructure for Biological Information
ELSI	Ethical, Legal, and Social Implications
EMBRIC	European Marine Biological Resource Centre
EMDB	EMDB: Electron Microscopy Data Bank
EoP	Elevation of Privilege
ERINHA	European Research Infrastructure on Highly Pathogenic Agents
EU-OPENSREEN	European Infrastructure of Open Screening Platforms for Chemical Biology
Euro-Biolmaging	European Biomedical Imaging Infrastructure
FIMM	Institute for Molecular Medicine Finland
hSERN	Human Sample Exchange Regulation Navigator
HTTP	Hypertext Transfer Protocol
IC	Informed Consent
Infrafrontier	European Infrastructure for Phenotyping and Archiving of Model Mammalian Genomes
INSTRUCT	Integrated Structural Biology Infrastructure for Europe
LAT	Legal Assessment Tool
LINDDUN	Linkability, Identifiability, Non-repudiation, Detectability, Disclosure of information, Content Unawareness, Policy and consent non-compliance
MAC	Mandatory Access Control
MTA	Material Transfer Agreement
NIST	National Institute of Standards and Technology
OE	Organizational Entity
OPM	Open Provenance Model
PKI	Public Key Infrastructure
PROV-DM	PROV data model
REMS	Resource Entitlement Management System
RI	Research Infrastructure
RBAC	Role-Based Access Control
SAML	Security Assertion Markup Language
SDL	Secure Development Lifecycle
SSL	Secure Socket Layer
SSO	Single sign-on
STRIDE	Spoofing, Tampering, Repudiation, Information Disclosure, Denial of service, Elevation of privilege
TLS	Transport Layer Security
TUM-MED	Klinikum rechts der Isar der Technischen Universität München
UC	Use Case
UDUS	Heinrich-Heine-Universität Düsseldorf
WP	Work Package
WT	Work Task



1 Executive summary

This deliverable report describes the security architecture and framework of BioMedBridges (BMB). Deliverable 5.3 is based on three Work Tasks (WTs): WT 5, *Security requirements for an e-infrastructure addressing the use cases*, WT6, *Threat and risk analysis for sharing data or biomaterials*, and, particularly, WT7, *Design of the security architecture and framework*. Deliverable 5.3 also lays the groundwork for WT8, *Implementation of a pilot for the security framework*.

Deliverable 5.3 builds upon the previous deliverables 5.1 [1] and 5.2 [2], and one of its central elements has been cooperation with other Work Packages (WPs). In order to better understand the security requirements and threats, close contacts were established with all Use Case (UC) work packages, i.e. WPs 6-10, and with the other construction work packages, especially WP4 Technical integration. Contacts to WP11 e-advisory task force also established a connection beyond BioMedBridges to external partners, including the European Grid Infrastructure.

Feedback from Research Infrastructures and use case work packages was sought early. Building upon deliverable 5.1, relevant data bridges and related security requirements as well as threats were identified. This was based on two surveys which have been carried out and are described here. The first survey was conducted in Aug 2012. Its results were of relevance to both deliverables 5.1 and to 5.3 and were also described in the report to deliverable 5.1. The second survey was carried out in September/October 2013 and was aiming primarily at security questions. It resulted in data flow diagrams illustrating security threats.

This report covers a complete series of steps from requirements (identified by means of the two surveys and based upon requirement clusters that are explained in the report to deliverable 5.1) to data flow diagrams and threats (constructed from results of the second survey) to risks, further to countermeasures, and finally towards the architectural elements of the proposed security framework. An important characteristic of the security architecture described here is its bottom-up development. As presented in section 9, process elements and descriptions are based on feedback from the



research infrastructures and the use case work packages. Specifications in the form of activity diagrams (deliverable 5.1) and data flow diagrams (this deliverable report) were developed in direct interaction with representatives of the use case work packages. Domain specific (EGA) and domain independent (EGI) experience was sought for the definition of security specific process elements.

This document is structured as follows:

Section 3 looks at related work that has influenced this report, including previous deliverables of WP5 (i.e. deliverable 5.1 and 5.2), other WPs (e.g. WP4); and projects outside of BioMedBridges, yet within the same or a similar domain and with relevant data management, security and privacy requirements.

Section 4 reviews the essential definitions of the most relevant concepts in the domain of data security and privacy in information systems referred to throughout this report.

Section 5 introduces the notion and scope of a generic data bridge with a specific focus on the security architecture and framework to be defined. The concepts explained and reviewed in section 5 are also referenced throughout the deliverable.

Section 6 is named after WT5 of WP5: *Security requirements for an e-infrastructure addressing the use cases*. It relies on two surveys (WP5 survey 1, and WP5 survey 2), and discusses in detail the methodology and results of both. The data bridges of use case WPs are illustrated by data flow diagrams annotated with the answers to survey 2.

Section 7 corresponds to WT6 of WP5: *Threat and risk analysis for sharing data or biomaterials*. For each UC WP it presents (a) the relevant security and privacy threats to be addressed; (b) the results of the risk assessment associated to those threats; and (c) the applicable security and privacy countermeasures to minimize risk.

Section 8 aligns to WT7 of WP5, which is the central topic of Deliverable 5.3: *Design of the security architecture and framework*. It revisits the concept of a generic data bridge introduced in section 5 and augments it, showing how



relevant security and privacy countermeasures can be used to secure data bridges. It also introduces the main components that form the security and privacy architecture and for each component provides (a) a definition; (b) the threats that it addresses; and (c) possible available implementation or deployment solutions. Central elements are the specification of three data access tiers supported by the security framework and a workflow activity diagram that illustrates the main actions involved in a generic data bridge for each access tier.

Section 9 presents an overview of the extensive activities to seek feedback from different kinds of partners as early and as comprehensively as possible. It then looks ahead towards implementation and, illustrating how implementation is based on requirements, suggests specific pilot scenarios.

The remainder of contents in this document, i.e. section 2 and sections 10-16, cover aspects that complement the material previously mentioned (e.g. appendices, references), or that are important to the position of deliverable 5.3 in the overall plan of the BioMedBridges project (e.g. project metrics, goals).

2 Project objectives

With this deliverable, the project has reached or the deliverable has contributed to the following objectives¹:

No.	Objective	Yes	No
1	Report has been completed on regulations, privacy, security, and IP requirements	x	
2	Tool has been realized for assessment of regulatory and ethical requirements	x	
3	Security architecture and framework have been specified, security requirements and risks identified	x	
4	Security framework successfully implemented		x

¹ The project objectives shown correspond to the list of milestones identified for WP5 on the Description of Work document.



3 Background

This deliverable builds upon the work presented in previous deliverables released so far as part of WP5:

- D5.1 [1]: Report on regulations, privacy and security requirements (using preliminary results from WT5). Deliverable 5.1 deals with legal interoperability in the context of data exchange that has become an important central concept for research collaboration. A high-level domain scenario that describes the problem space of the data bridges listing aims, actors, problems and benefits of the bridges in a storyboard has been created, and in a second step concepts of legal interoperability have been applied to five usage scenarios. These scenarios, which are precursors of the corresponding fully developed BioMedBridges use case work packages, have resulted in “requirements clusters” for the data bridges.
- D5.2 [2]: For deliverable 5.2, the Legal Assessment Tool (LAT) for assessment of regulatory and ethical requirements was realized. D5.2 also provided relevant supportive documents to researchers needing to use sensitive data².

The concepts and definitions of these deliverables will be explained in appropriate places throughout this document.

3.1 Related work and projects

This section includes an overview of projects outside BioMedBridges, yet within the same or a similar domain, with data management security or privacy characteristics worth noting. Relevant projects are:

- *Advancing Clinico-Genomic Clinical Trials on Cancer (ACGT)*³. The ACGT project is co-funded by the European Union and ended in July 2010. Its goals involved the development of open-source, semantic and grid-based technologies in support of post genomic clinical trials

² <http://www.biomedbridges.eu/deliverables/52-0>

³ <http://acgt.ercim.eu/>



in cancer research. The project aimed at providing an open platform where new and powerful services can be offered and used by practitioners in the field, which include clinicians and bio-researchers as well as software developers. From a security perspective in the context of BMB, two public deliverables stand out due to their focus on ethical and legal requirements, and security guidelines, respectively:

- Deliverable 10.2: The ACGT ethical and legal requirements⁴, 13/03/2007
- Deliverable 11.4: Requirements and guidelines for developing secured ACGT services⁵, 31/07/2010

— *Electronic Health Records for Clinical Research (EHR4CR)*⁶. As stated on the project website front page, “*the EHR4CR project is -to date- one of the largest public-private partnerships aiming at providing adaptable, reusable and scalable solutions (tools and services) for reusing data from Electronic Health Record systems for Clinical Research.*” EHR4CR also has to manage the security aspects of EHR data, which are addressed in the deliverable:

- D5.1: Requirements and specifications of the security and privacy services⁷

— *Integrative Cancer Research through Innovative Biomedical Infrastructures (INTEGRATE)*⁸. The project seeks to promote large-scale collaboration in biomedical research, developing new infrastructures to enable data and knowledge sharing. The INTEGRATE deliverables that deal with some of the concepts addressed in Deliverable 5.3 of BMB include:

- D1.3 INTEGRATE legal, ethical and regulatory requirements⁹, 06/12/2011

⁴ <http://acgt.ercim.eu/documents/public-deliverables.html#D102>

⁵ <http://acgt.ercim.eu/documents/public-deliverables.html#D114>

⁶ <http://www.ehr4cr.eu/>

⁷ <http://www.ehr4cr.eu/executiveSummaries.cfm>

⁸ <http://www.fp7-integrate.eu/>

⁹ <http://www.fp7-integrate.eu/images/Documents/d1.3%20integrate%20legal%20ethical%20and%20regulatory%20requirements.pdf>

<http://www.fp7-integrate.eu/images/Documents/d1.3%20integrate%20legal%20ethical%20and%20regulatory%20requirements.pdf>



- D2.1 State-of-the-art report on standards¹⁰ (Section 7), 10/31/2011
- *From data sharing and integration via Virtual Physiological Human (VPH) models to personalized medicine (p-medicine)*¹¹. As stated in brief, on the project landing web page, p-medicine is “... *aiming at developing new tools, IT infrastructure and VPH models to accelerate personalized medicine for the benefit of the patient.*” P-medicine also puts forward various public deliverables relevant to the focus of this report, namely:
- D3.1 State of the art report on standards (Section 8), 31/10/2011
 - D5.1: Setting up of the data protection and data security framework, 31/01/2012
 - D5.3 Report on legal and ethical issues regarding access to biobanks, 31/01/2013
- *Translational Research and Patient Safety in Europe (TRANSFoRm)*¹². At the core of TRANSFoRm is the concept of developing a “*rapid learning healthcare system*” leveraging advanced computational infrastructures with the purpose of improving both patient safety and the conduct and volume of clinical research in Europe.
- D3.2: Report on regulatory requirements, confidentiality and data privacy issues¹³, 31/03/2011
- *BiobankCloud*¹⁴. The BiobankCloud project goal is defined on its front web page as: “*BioBankCloud is an EU-funded FP7 project that is developing a cloud-computing platform as a service (PaaS) for the storage, analysis and inter-connection of biobank data. Our platform*

¹⁰ <http://www.fp7-integrate.eu/images/Documents/d2.1%20state-of-the-art%20report%20on%20standards.pdf>

¹¹ <http://p-medicine.eu/>

¹² <http://www.transformproject.eu/>

¹³ Follow the web links: <http://www.transformproject.eu/> > Publications > Deliverables

¹⁴ www.biobankcloud.com/



will provide security, storage, data-intensive computing tools, bioinformatics workflows, and support for allowing biobanks to share data with one another, all within the existing regulatory frameworks for the storage and usage of biobank data.” With these goals in mind, several deliverables¹⁵ are relevant to the scope of BMB. In the context of 5.3, we highlight one in particular because of its focus on security:

- D3.1 Security State of the Art¹⁶, 30/05/2013

- *Biobanking and Biomolecular Resources Research Infrastructure (BBMRI) Preparatory Phase*¹⁷. The goal of the BBMRI preparatory phase project is described in its mission statement: “*To prepare for the construction of a pan-European Biobanking and Biomolecular Resources Research Infrastructure (BBMRI) for biomedical and biological research in Europe and worldwide, building on existing infrastructures, resources and technologies, specifically complemented with innovative components and properly embedded into European ethical, legal and societal frameworks.*” A detailed description of how the project has carried out this mission can be found in its “Final Report”¹⁸ dated 19/04/2011.

- *COordination Of Standards In MetabOlomicS (COSMOS)*¹⁹. The goal of the COSMOS project is to enable free and open sharing of metabolomics data developing new (or promoting existing) community standards. The project calls explicitly for the need of collaboration with the Biological and Medical Sciences (BMS) research infrastructures of BMB, and to that extent includes a specific Work Package:
 - WP6 - Coordination with BioMedBridges and biomedical European Strategy Forum on Research Infrastructures²⁰ (ESFRI)

¹⁵ <http://www.biobankcloud.com/?q=publications>

¹⁶ <http://www.biobankcloud.com/sites/default/files/deliverables/D3.1-final.pdf>

¹⁷ <http://bbmri.eu/>

¹⁸ <http://bbmri.eu/final-and-interim-report>

¹⁹ <http://www.cosmos-fp7.eu/>

²⁰ <http://www.cosmos-fp7.eu/wp6>



- *European Genome-phenome Archive (EGA)*²¹. The EGA repository allows the exploration of datasets from genomic studies, supplied by a range of data providers. The nature of the data in the EGA repository implies that access to datasets is governed by the appropriate Data Access Committee (DAC). The underlying data sharing model of the EGA has been an important reference for the security framework of BMB presented here in D5.3.

Other projects or resources outside the bioinformatics domain worth considering within the scope of this deliverable include:

- *European Grid Infrastructure (EGI)*²². The EGI aims at facilitating the Information and Communications Technology (ICT) needs of the research community across Europe. The vision of the EGI states: “*To support the digital European Research Area through a pan-European research infrastructure, based on an open federation of reliable services, which provide uniform access to computing and data storage resources.*”
- *Future ID*²³. The Future ID project targets to shape the future of electronic identity. More specifically, as stated on the front web page of the project, “*the FutureID project builds a comprehensive, flexible, privacy-aware and ubiquitously usable identity management infrastructure for Europe. It integrates existing electronic ID technology, trust infrastructures, emerging federated identity management services, and modern credential technologies. It creates a user-centric system for the trustworthy and accountable management of identity claims.*”

²¹ <https://www.ebi.ac.uk/ega/>

²² <http://www.egi.eu/>

²³ <http://www.futureid.eu/>



4 Security and privacy concepts and definitions

In order to lay a conceptual foundation for the discussion of the security architecture and framework put forward here, this section introduces some basic concepts from the fields of security and privacy. Some essential definitions used throughout this document are given in section 4.1, including explanations about how the defined terms relate to each other and what role they play with respect to the development of a security architecture. Section 4.2 then briefly presents the threat modelling methodology STRIDE that was applied in order to conduct a security risk assessment for the individual BMB UC WPs. This presentation includes definitions of both the particular threats addressed by STRIDE and the respective security properties they compromise. To analyse the privacy risks for the UC WPs, an extension to STRIDE called LINDDUN was utilized which is introduced in 4.3. The definitions of the threats covered by the LINDDUN methodology and the corresponding privacy properties they imperil are described in 4.3.1.

4.1 Threat, vulnerability, and risk

The goal of any security architecture is to facilitate the protection against threats to the security of a system [3]. To clarify what the notion of a threat precisely means in this context, the term shall be defined as:

“Threat – Any circumstance or event with the potential to adversely impact organizational operations, ..., organizational assets, individuals, ... through an information system via unauthorized access, destruction, disclosure, or modification of information, and/or denial of service [4].”

According to this definition, adverse disclosure of information is considered to be a security threat. Consequently, the protection of personal data from unauthorized access, and thus the preservation of privacy, is a requirement for a security architecture. The aspect of privacy is particularly important for systems dealing with sensitive medical data and thus highly relevant for BMB. Going forward, when we refer to security, the support and awareness of



privacy concerns will be implicitly considered as well, unless mentioned otherwise.

Threats are to be distinguished from vulnerabilities which are properties of a system according to the following definition:

“*Vulnerability* – Weakness in an information system, system security procedures, internal controls, or implementation that could be exploited by a threat source [4].”

The protection of a complex real-world system from any potential threat is of course infeasible. Therefore, in practice, any useful security architecture must balance the benefits of protection against their total costs. This balance can be determined by conducting a so-called *risk analysis* [3] that determines whether an asset should be protected, and to what level, by extending potential threats against that asset to risks which are defined as follows:

“*Risk* – A measure of the extent to which an entity is threatened by a potential circumstance or event, and typically a function of:

- (i) the adverse impacts that would arise if the circumstance or event occurs; and
- (ii) the likelihood of occurrence [4].”

The higher the adverse impacts that would arise if a threat occurs and the higher the likelihood of occurrence, the higher the risk associated to the threat, and the more effort should be put into mitigating it.

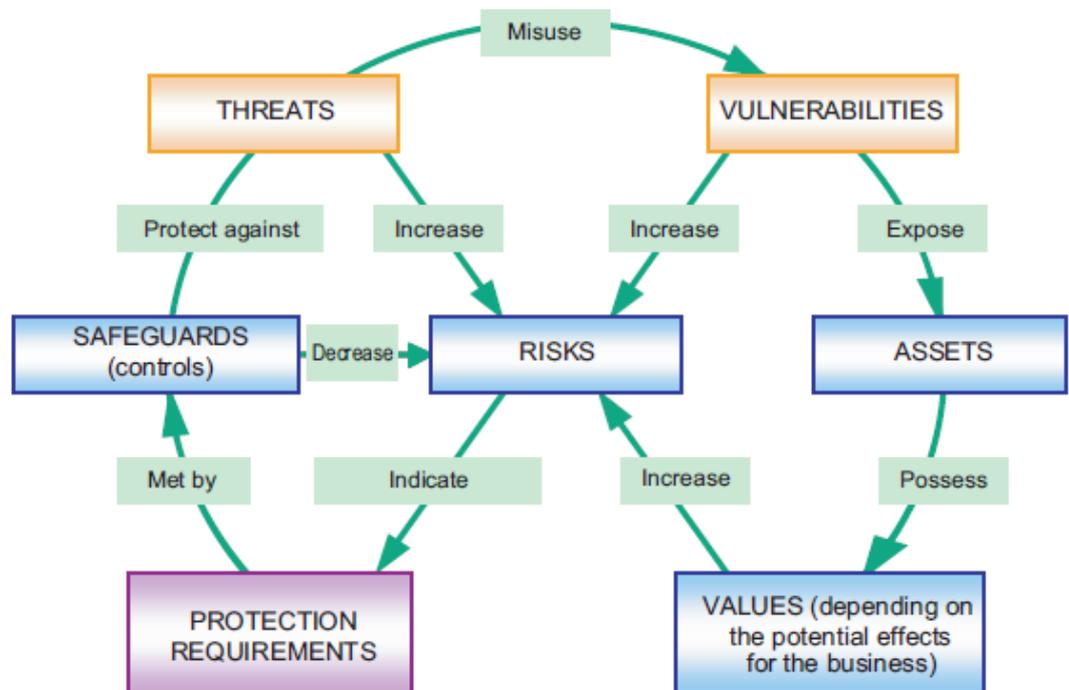
Based on the results of a risk analysis, a reasonable selection of countermeasures (also called *safeguards*) can be derived in order to decrease the highest risks, i.e. mitigate the most likely and harmful threats. In this context, the term countermeasure shall be defined as follows:

“*Countermeasure* – An action, device, procedure, or technique that meets or opposes (i.e., counters) a threat, a vulnerability, or an attack by eliminating or preventing it, by minimizing the harm it can cause, or by discovering and reporting it so that corrective action can be taken [5].”



Figure 1 provides an overview of the security concepts introduced in this section and the relationships between them.

Figure 1. Relationship between vulnerabilities, threats, risks, and safeguards [6]



4.2 Security threat modelling using STRIDE

As explained in the paper [7], STRIDE has been developed by Microsoft as part of the Secure Development Lifecycle (SDL) [8] which is a well-established methodology for building secure software systems. In essence, STRIDE is a systematic approach for security threat modelling that can be used for conducting a risk analysis as introduced in section 4.1. We selected the STRIDE and LINDDUN approaches as they allow for the consistent modelling of privacy and security threats. STRIDE supports the definition of use scenarios, the identification of security requirements, the analysis of threats based on a graphical representation called data flow diagram (DFD) of the system to be assessed, and the derivation of risks from the identified threats, which necessitate appropriate countermeasures. A description of the DFD notation is provided in section 6.2.1 in the context of a risk analysis conducted for the individual BMB use case work packages. In the following,



we explain the individual threats covered by STRIDE and the security properties they compromise.

4.2.1 Threats and security properties addressed by STRIDE

The name STRIDE is an acronym formed from the initial letters of the security threat types addressed by the methodology, namely **S**poofing, **T**ampering, **R**epudiation, **I**nformation disclosure, **D**enial-of-service, and **E**levation-of-privilege. They are defined as follows:

1. “*Spoofing* – Spoofing threats allow an attacker to pose as something or somebody else [8].”
2. “*Tampering* – Tampering threats involve malicious modification of data or code [8].”
3. “*Repudiation* – An attacker makes a repudiation threat by denying to have performed an action that other parties can neither confirm nor contradict [8].”
4. “*Information disclosure* – Information disclosure threats involve the exposure of information to individuals who are not supposed to have access to it [8].”
5. “*Denial-of-service* – Denial-of-service (DoS) attacks deny or degrade service to valid users [8].”
6. “*Elevation of Privilege* – Elevation-of-privilege (EoP) threats often occur when a user gains increased capability [8].”

The security properties these threats imperil are, respectively:

1. “*Authenticity* – property that an entity is what it claims to be [9].”
2. “*Integrity* – property of protecting the accuracy and completeness of assets [9].”
3. “*Accountability* – responsibility of an entity for its actions and decisions [9].”
4. “*Confidentiality* – property that information is not made available or disclosed to unauthorized individuals, entities, or processes [9].”
5. “*Availability* – property of being accessible and usable upon demand by an authorized entity [9].”
6. “*Authorization* – approval that is granted to a system entity to access a system resource [5].”



Information disclosure threats, if left unmitigated, can become privacy violations if the disclosed data is confidential or personally identifiable information [8]. This is, however, the only aspect of STRIDE that involves privacy protection; the other threats rather constitute typical security, but not privacy related issues. Regarding the complex privacy requirements (see also deliverable 5.1) a security architecture for BMB is faced with, a risk analysis based on STRIDE threats alone would thus be clearly insufficient.

Accountability and Auditing are of central relevance to a security architecture. We point out that the related, but somewhat more fundamental concept of “*provenance*” also needs substantial consideration. Citing [10], we can say that “the provenance of a piece of data refers to knowledge about its origin, in terms of entities and actors involved in its creation, e.g. data sources used, operations carried out on them, and users enacting those operations.” In extension of audit trails, provenance traces and provenance-aware systems are needed.

4.3 Privacy threat modelling using LINDDUN

LINDDUN is a complement to STRIDE suggested by Deng et al. that addresses privacy-specific threats [7]. Following the basic STRIDE methodology, LINDDUN supports the mapping of threats to DFD elements in order to derive risks. As a result of their similar approaches, STRIDE and LINDDUN can be combined in order to perform a closely integrated security and privacy analysis, but can also be applied independently.

Deliverable 5.1 dealt with the applicable rules for data protection within BMB. Personal data played an important role for the developed requirements clusters of D5.1. For this reason, definitions for human and non-human data, identifying data, pseudonymized data, anonymized data, informed consent, purpose limitation, and intellectual property issues were created. These definitions were harmonised with the BMB Ethical Governance Framework²⁴. A definition of legal terms can also be found in the online BMB glossary²⁵. These terms were used for the analysis of the usage scenarios as well as for

²⁴ BioMedBridgesBMB Ethical Governance Framework Version 1.1, April 2013

²⁵ http://www.biomedbridges.eu/dms/page/site/biomedbridges-partners/wiki-page?title=Definition_of_terms_used_in_the_project_-_wiki



the data provider survey. Some of these definitions were taken from national and international guidelines and regulatory documents and some are used as central references for the purpose of this deliverable report. But additional definitions that are more technical and data security focused were necessary for the purpose of the security framework. We present these additional definitions together with the LINDDUN definitions below. In this respect, we minimally deviate from LINDDUN.

4.3.1 Threats and privacy properties addressed by LINDDUN

Similar to STRIDE, the name LINDDUN is an acronym based on the particular threat types it comprises: **L**inkability, **I**dentifiability, **N**on-repudiation, **D**etectability, **D**isclosure of information, **C**ontent unawareness, and **P**olicy and consent **n**on-compliance. In [7], they are defined as follows:

1. “*Linkability* of two or more items of interest (IOIs, e.g. subjects, messages, actions, etc.) allows an attacker to sufficiently distinguish whether these IOIs are related or not within the system.”
2. “*Identifiability* of a subject means that the attacker can sufficiently identify the subject associated to an IOI.”
3. “*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.”
4. “*Detectability* of an IOI means that the attacker can sufficiently distinguish whether such an item exists or not.”
5. “*Information disclosure* threats expose personal information to individuals who are not supposed to have access to it.”
6. “*Content unawareness* indicates that a user is unaware of the information disclosed to the system.”
7. “*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.”

The privacy properties these threats compromise are, respectively:

1. “*Unlinkability* of two or more IOIs ... means that within the system ..., the attacker cannot sufficiently distinguish whether these IOIs are related or not [11].”
2. According to the LINDDUN methodology,



“*anonymity* of a subject ... means that the attacker cannot sufficiently identify the subject within a set of subjects, the anonymity set [11].”

Instead, we use this related definition of anonymised data utilized in deliverable 5.1:

”data, which has been rendered anonymous in such a way that the data subject is no longer identifiable [12].”

Furthermore, LINDDUN defines

“*Pseudonymity* – A subject is pseudonymous if a pseudonym is used as identifier instead of one of its real names [11].”

This definition is closely related to the following definition of pseudonymised data used in deliverable 5.1, p. 53: “Data which has been pseudonymised does not relate information to identifiable subjects for people receive on and holding the data but contains information or codes which would enable others to identify an individual from it [13].”

3. “*Plausible deniability* ... means that an attacker cannot prove a user knows, has done or has said something [7].”
4. “*Undetectability* and *unobservability* ... of an IOI ... means that the attacker cannot sufficiently distinguish whether it exists or not [11].”
5. “*Confidentiality* means preserving authorized restrictions on information access and disclosure, including means for protecting personal privacy and proprietary information [14].”
6. “*Content awareness* – The user needs to be aware of the consequences of sharing information [7].”
7. “*Policy and consent compliance* ensures that the system’s (privacy) policy and the user’s consent ... are indeed implemented and enforced [7].”

According to [7], the anonymity privacy property essentially refers to hiding a link between an identity and an action or a piece of information. In this sense, “anonymity of a subject with respect to an attribute may be defined as unlinkability of this subject and this attribute [11]”.

Pseudonymity, which basically refers to replacing a link between an identity and data with a link between one or more pseudonyms and the respective



data, can also be perceived with respect to linkability. While anonymity corresponds to unlinkability of subjects and attributes, pseudonymity may be regarded as a reduced degree of linkability that typically enables only a restricted audience to link subjects to attributes.

Comparing the threat definitions of STRIDE and LINDDUN, the definition of information disclosure is more general in the STRIDE formulation as it captures the exposure of all kinds of information, while LINDDUN specifically addresses the exposure of personal information. LINDDUN thus focuses explicitly on the privacy related aspect of information disclosure we have pointed out in 4.2.1.

Regarding the security and privacy properties, STRIDE focuses on accountability as a measure to “provide irrefutable evidence concerning the occurrence or non-occurrence of an event or action [15].” LINDDUN, on the other hand, suggests the protection of plausible deniability which refers to the ability to deny having performed an action that other parties can neither confirm nor contradict [7]. Plausible deniability has the opposite effect of accountability: “That there be no irrefutable evidence concerning a disputed event or action [15].”

Depending on the character of the system that is to be protected by a security architecture, sometimes accountability is more desirable than plausible deniability, and sometimes the reverse is true. As [14] states, the collection and communication of audit trail is important in the context of health information exchange. Consequently, the support of accountability is of high relevance for a security architecture for BMB, while plausible deniability is less relevant.

5 Modelling a generic data bridge

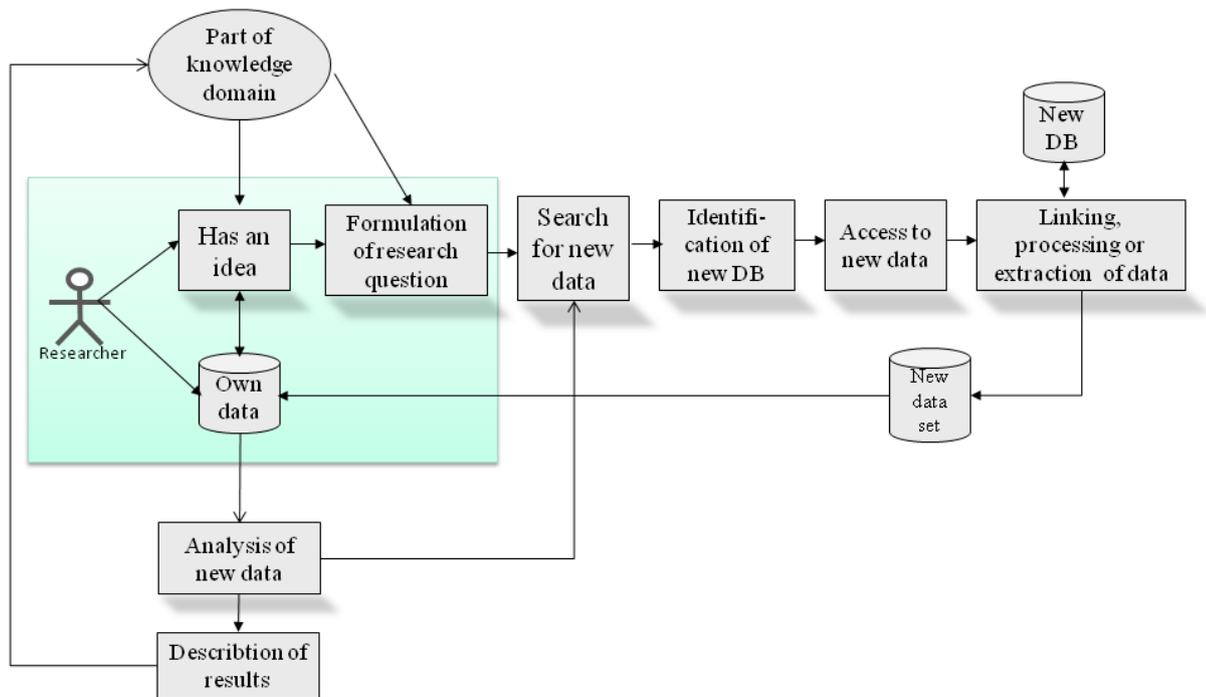
This section discusses the notion of a (generic) data bridge which is a key component of the security architecture and framework. It is a concept used throughout this document that also helps to clarify the scope and application range of the security architecture. A result of the usage scenario analysis in deliverable 5.1 has been that the bridging of research infrastructures by



means of data transport has to consider semantic integration, linking of information, metadata enrichment, etc. While deliverable 5.3 builds upon the two previous deliverables, its main focus is on the security architecture and framework.

In deliverable 5.2, a generic research process for data bridges in BMB has been explained that is illustrated in Figure 2 below. Briefly recapitulating deliverable 5.2, the process starts with a researcher formulating a research question based on their own ideas and own data. This might lead to the need for further data and possibly for building a “data bridge” between their own data and a “new database” (i.e. an external database which the researcher wants to access). If personal data are involved, access to data from such a new database has to be regulated taking legal constraints into account. The data is transferred to the researcher and linked, processed, and analysed in view of the research question. If enough data to answer the research question has been obtained, the process ends with a description of the results. If not, then successive data gathering rounds may be initiated. Deliverable 5.3 addresses the security requirements of this process.

Figure 2. Data flow in BioMedBridges Data Bridges from deliverable 5.2





In deliverable 5.1, it was pointed out that “BioMedBridges will provide data bridges between the individual biological and medical sciences research infrastructures ... clustering them together and linking basic biological research and its data”, so here the aspect of research across RI boundaries is emphasized.

Throughout this deliverable, taking both deliverables 5.1 and 5.2 into account, we will use the term *data bridge* in accordance with the process depicted in Figure 2 and focus on scenarios where the researcher and the “new database” belong to different RIs. Within such a data transfer process, the RI containing the “new database” (or another kind of data source, possibly even a mere text file containing relevant data) acts as a *data provider*, while the RI where the researcher is located acts as a *data consumer*. We note here that in case of personal data involved, the relevant legal representation of a data provider is that of a responsible *data controller*, i.e. “the natural or legal person, public authority, agency or any other body which alone or jointly with others determines the purposes and means of the processing of personal data [12].”

While a data bridge describes a concrete data transfer scenario between RIs, the term *generic data bridge* will denote a workflow model constituting a generalization of several concrete data bridges. From the opposite perspective, the respective data bridges can be seen as instances of the generic data bridge. In this sense, a generic data bridge can serve as a blueprint for data exchange between RIs.

An important prerequisite for the definition of the security architecture and framework was a security requirements assessment for BioMedBridges, including a risk analysis of the UC WPs using STRIDE and LINDDUN. Based on this work, appropriate countermeasures were derived and a generic data bridge was designed that outlines a workflow employing the identified countermeasures in order to facilitate secure and privacy preserving data exchange. Secure data bridges can be created by instantiating this generic data bridge with respect to concrete data transfer scenarios, within the scope of BMB, and hopefully also in biomedical research beyond this project.



6 Security requirements for a e-infrastructure addressing the use cases

One of the tasks required to define the security architecture and framework of BMB involves the identification of the data security (and privacy) requirements for the e-infrastructures that enable the use case Work Packages to meet their goals (i.e. work packages 6, 7, 8, 9, and 10).

According to the Description of Work (DoW) of BioMedBridges, work task 5 of WP5 focussed on “Security requirements for an e-infrastructure addressing the use cases”. In order to identify these requirements, two surveys were carried out. They are referred to here as “*WP5 Survey 1*”, and “*WP5 Survey 2*”:

- ***WP5 Survey 1*** focused on the security and privacy requirements for the BMS research infrastructures that participate in BioMedBridges. Detailed results already been presented in deliverable 5.1. We refer to deliverable 5.1 which has introduced the concept of “Requirements Clusters”. These “Requirements Clusters” have been created with a focus on personal data, and they therefore form the input for the Legal Assessment Tool (D5.2). A summary of WP5 Survey 1 is also given in Section 6.1 below.
- ***WP5 Survey 2*** focused on the security and privacy requirements for the use case WPs of BioMedBridges.

There was an agreement within the project that to fully understand the security and privacy needs of the use case WPs, it was also necessary to understand these needs within the BioMedBridges RIs that ultimately support the use case WPs. This motivated the two surveys and their respective focus. Further reasons for conducting two surveys were project organizational aspects, including the early start of WT5 of WP5 (month 6 of the project). At the time of conducting “WP5 Survey 1” (months 8-9) there were areas of the use case WPs that were not fully defined yet.

The results of both surveys are used in the additional WTs within the scope of deliverable 5.3 (i.e., WT6, and WT7). The remainder of this section examines



in detail the work carried out as a part of both surveys, including their background, aims, methodology, and results.

6.1 Security requirements for the research infrastructures

As stated earlier, WP5 Survey 1 focused on understanding the expected data security and privacy needs of the BMS RIs that are part of the BMB project²⁶ i.e. BBMRI, EATRIS, ECRIN, ELIXIR, EMBRC, ERINHA, EU-OPENSREEN, Euro-Biolmaging, Infrafrontier, and INSTRUMENT.

The target audience of the survey were members of the mentioned BMS RIs, familiar with (a) the types of data that a RI intends to use (consume) or share (provide) with respect to other RIs; (b) the current security measures that the data may be subject to; and (c) the status of implementation or deployment of such measures.

To conduct the survey a series of questions were developed by TUM-MED with the assistance of UDUS. The questions were further discussed and modified by the group during the WP5 workshop on July 12, 2012 in Düsseldorf (Germany). The questionnaire consisted of 13 questions that spanned across 5 different areas of data security and privacy. The questions stressed the key distinction of the use of data versus the sharing of data that the RIs may intend to make. Table 1 shows the 5 areas the questionnaire aimed to cover.

Table 1. Data security areas of “WP5 Survey 1” questionnaire.

Category		Num. of questions
1	Human data on the level of individuals	4
2	Characterization of data, serving and use of biosamples, ethical and contractual situation	3
3	Role of informed consent	2
4	Need for organisational/technical measures	2
5	Open vs anonymous vs pseudonymous	2
Total		13

²⁶ ESFRI Strategy Report and Roadmap Update 2010, http://ec.europa.eu/research/infrastructures/pdf/esfri-strategy_report_and_roadmap.pdf (last access, May 2014)



The survey was distributed by the EMBL project management team in Hinxton (UK) to the BMB partners on August 23, 2012. Not only WP5 members but all RIs and use case WPs were asked to answer by September 15, 2012.

Fifteen answers were received, and nine participants answered on behalf of their entire RIs.

A detailed review of the results gained from “WP5 Survey 1” can be found in section 12 (Annex 1) of deliverable 5.1 [1]. The core results concerning data at the individual level can be summarized as follows:

- 4 RIs intend to share human data on the individual level.
 - BBMRI, EATRIS, ELIXIR, Euro-Biolmaging
- 7 RIs want to consume human data on the individual level.
 - BBMRI, EATRIS, ECRIN, ELIXIR, Euro-Biolmaging, Infrafrontier, Instruct
- Informed Consent, Ethics Committee approval, and Data Use Agreements are considered relevant.

Table 2. WP5 Survey 1. Use/Sharing of individual level-data

Research Infrastructure	Plan for individual level data	
	Use (consume)	Share (provide)
BBMRI	Yes	Yes
EATRIS	Yes	Yes
ECRIN	Yes	-
ELIXIR	Yes	Yes
EMBRC	-	-
ERINHA	-	-
EU-OPENSREEN	-	-
Euro-Biolmaging	Yes	Yes
Infrafrontier	Yes	-
Instruct	Yes	-

6.2 Security requirements for the use case work packages

While WP5 Survey 1 focused on the data security requirements expected by the various BMB RIs, a second survey, WP5 Survey 2, was carried out also as part of WT5, now focusing specifically on the security and privacy



requirements within each one of the use case WPs. The target population in this case consisted of representatives of each of the use case WPs (i.e. WP6, 7, 8, 9, and 10). Feedback from WP9 could not be received by the survey, so answers by WP9 on selected questions were collected later.

Survey 2 of WP5 builds upon the results from WP5 Survey 1 and the security requirements for the use case WPs reported in deliverable 5.1. It extends both efforts in several respects, most notably:

- Assessing the further details of the current or future security and privacy measures desired for the various components of an e-infrastructure concerning each use case WP.
- Going from security requirements for the RIs to security requirements for the use case WPs.
- Attempting to discover aspects of the use case WPs that may not have been covered by the corresponding usage scenarios and activity diagrams presented in section 9 of deliverable 5.1.

Another important factor that shaped the content and goals of WP5 Survey 2 was providing a basis for the threat and risk analysis to be carried out as a part of WT6 of WP5 (see WP5 in BMB DoW document [16] reproduced in section 13). This factor had some implications, mainly:

- The selection of a methodology to perform a threat and risk analysis on the use case WPs, that is STRIDE [8] for security threats, and LINDDUN [7] for privacy threats (as introduced in sections 4.2, and 4.3 above respectively).
- The development of a DFD [17] [18] for every use case WP, given that DFDs are core to the STRIDE and LINDDUN threat and risk analysis methodology.

Therefore, WP5 Survey 2 aimed also at defining DFDs for the use case WPs building upon the activity diagrams included as a part of deliverable 5.1 that illustrated the usage scenarios of the WPs involved. As a result, DFDs became a central topic of WP5 Survey 2, and this was reflected in the questionnaire developed to conduct the survey.

The rest of this section elaborates on the development of WP5 Survey 2, and the discussion of its goals and results.



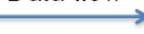
6.2.1 Overview of data flow diagrams

This section provides a basic overview of data flow diagrams focused on those concepts that will be referred to hereupon. DFDs were introduced by Larry Constantine, one of the original developers of Structure Design [17]. As stated in the report by Le Vie [18], “*DFDs are an important technique for modeling a system’s high-level detail by showing how input data is transformed to output results through a sequence of functional transformations*”.

DFDs played a major role in the development process of the security framework for BioMedBridges. They were used to model the use case WPs, which allowed the application of the STRIDE and LINDDUN threat and risk analysis methodology to each use case WP modelled. The outcome of this analysis is reported in section 7.

DFDs consist of four major components. Their graphical notation and definition are given in Table 3 as they are presented in [18].

Table 3. Components of a Data Flow Diagram

Components of a Data Flow Diagram	
Graphical Notation	Definition
 Data store	Data Store: “Data Stores are repositories of data in the system. They are sometimes also referred to as files.”
 Data flow	Data Flow: “Data Flows are pipelines through which packets of information flow. Label the arrows with the name of the data that moves through it.”
 Process	Process: “A process transforms incoming data flow into outgoing data flow.”
 External Entity	External Entity: “External entities are objects outside the system, with which the system communicates. External entities are sources and destinations of the system’s inputs and outputs.”

The sections that follow examine in more detail the use of DFDs in the development process of the security framework and present the specific diagrams of the use case WPs produced as a result of this process.



6.2.2 Survey questionnaire for the Use Case Work Packages

In order to assist the use case WPs to identify their data security and privacy requirements, TUM-MED developed a questionnaire as part of the second survey “WP5 Survey 2”. The questionnaire addressed the relevant aspects that would enable the survey to meet the goals outlined earlier. Table 4 itemizes the questions used in the survey.

The selection of questions corresponded to the following criteria driven by the survey goals:

- To expand the security requirements for usage scenarios (see chapter 9. Usage Scenarios of deliverable 5.1) into more specific use cases, additional technical information regarding types of data, data formats, and types of security measures in use by the WPs was requested (e.g. questions 1, 3, 8, 14, 18, ...). The underlying idea was to elicit as much technical information as possible regarding the intended operations of the use case WPs that could be relevant for security and privacy requirements.
- To develop DFDs that modelled the use case WPs, the questions were grouped by the DFD component they were related to, that is: data stores, processes, data flows, and entities. This grouping is visible in Table 4.

Table 4. Questionnaire of the survey distributed among the use case WPs.

General	
1.	What are the differences between your usage scenario and your Use Case? What is not covered by the usage scenario?
Data Stores	
2.	What kind of data store do you use (e.g. database)? Please give a concrete description (what kind of database or database management system, e.g. MySQL?).
3.	What kind of data do you store in that data store?
4.	Does your data contain personal data? Is the data pseudonymous/anonymous?
5.	What format does this data have?
6.	What security measures already exist (e.g. authentication system via email, authorization system via role-based access control, data validation, encryption, audit trail, k-anonymity etc.)?
Processes	
7.	What program/calculation is executed?
8.	What is the input of the process? Which data format?
9.	What is the output of the process? Which data format?
10.	Is there a process which is executed together with this process (name the abbreviation, e.g. P2)?



11.	What security measures already exist (e.g. authentication system via email, authorization system via role-based access control, data validation, encryption, audit trail, k-anonymity etc.)?
External Entities	
12.	Who is the external entity (user, web service, server etc.) outside your system, you develop for the Use Case?
13.	Do you need an authentication/authorization mechanism for the external entity getting access, to check who the external entity is and what rights it has? Why is it needed?
14.	Do you already have an authentication/authorization mechanism? If so, which?
15.	Does the external entity itself have an authentication/authorization system your process/data flow/data store has to use? If so, which?
Data Flows	
16.	Which data flow do you describe, between what process/data store/entity?
17.	Is there a data flow that exists in parallel?
18.	What data is transferred?
19.	How is the data transferred?
20.	Is the transferred data confidential?
21.	Is there data in the data source that is not allowed to be transferred, e.g. besides anonymous data also personal data?
22.	What security measures already exist (e.g. authentication system via email, authorization system via role-based access control, data validation, encryption, audit trail, k-anonymity etc.)?

A prefilled version of the questionnaire gathered from the usage scenarios of the use case WPs included in section 9 of deliverable 5.1 was made available so that it could be reviewed and modified accordingly beforehand if necessary.

If a particular use case WP dealt with further DFD components that were missing in the preliminary draft (i.e., data stores, processes, data flows, and entities) they could be added to the survey.

6.2.3 Survey results for use case work packages

The survey was conducted and distributed by TUM-MED to members of the use case WPs during July - September 2013 and it included a number of follow-up comments and phone calls to clarify open questions. Out of the 5 use case WPs contacted, 4 participated in the survey (WP6, 7, 8, and 10) while WP9 “Use case: From cells to molecules - integrating structural data”, had to kindly decline. WP9 could not respond to the survey at the time, given that it was still in an early development phase and most of the information to be requested could not be provided.

The following subsections summarize the results of the survey. For each use case WP that participated, the corresponding subsection presents:



- The DFD that “emerged” iteratively as a result of the survey process representing the use case WP. The DFD exhibits some custom extensions to the standard graphical DFD notation that show at a glance relevant security aspects gathered throughout all answers. The extensions to the DFD notation are specified in the subsection that follows.
- A table reporting the most significant information from all the answers to the questionnaire provided by representatives of the use case WP. Regarding the contents of the table:
 - They are organized by the various DFD components following a similar organization to the questions contained in the questionnaire
 - The names of the DFD components in the table refer to the element with the same name in the DFD of the use case WP that resulted from the survey
- A discussion of the most significant aspects of the overall results.

The DFDs obtained were then shared with the corresponding WP representatives to confirm if they reflected the use case WP in question.

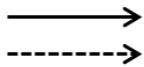
For the WP9 use case that did not take part in the survey, a different approach was chosen. A specific assessment of security and privacy requirements was performed based on the information of WP9 available within the project. The details of this process are given in section 6.2.4 below.

6.2.3.1 Custom extensions to notation of data flow diagrams

An additional contribution of the work carried out for “WP Survey 2” was a custom extension to the standard DFD graphical notation shown in Table 4. The extensions provide visual information (e.g. color code, annotations) of some security properties applicable to the data handled by the components of the DFD (i.e., data stores, processes, data flows, and external entities). The motivation for the additional notation to the DFDs was showing at a glance some of the security requirements expected for the realization of the corresponding use case WP. Table 5 shows the extensions made to the standard DFD notation.



Table 5. Custom extensions to DFD notation to represent security properties.

	DFD components displayed using a green border indicate that the access mode of the data involved is open
	DFD components displayed using a red border indicate that the access mode of the data involved is restricted
	DFD components displayed using a red dashed border indicate that the access mode of the data involved can be restricted or open
	The „User“ of the use case WP is displayed as an „External Entity“. <ul style="list-style-type: none"> • A red border indicates compliance with security requirements (e.g. authentication) • A green border indicates that no security measures are required
	Data Flow components of the DFDs are displayed as: <ul style="list-style-type: none"> • Solid arrows to indicate that secure communication is required • Dashed arrows to indicate that no security measures are required
	A dashed box with a bent corner will be used to attach annotations to a specific DFD component

6.2.3.2 Work package 6 “Imaging Bridge”

This section expands the security requirements for the usage scenario 1: “Imaging Bridge related to WP6” section 9.1 of deliverable 5.1. The WP6 “Use case: Interoperability of large scale image data sets from different biological scales”, focuses on building a data bridge to facilitate the comparison of cellular phenotypes with morphological imaging, accessing the images along with its metadata.

Figure 3 shows the DFD of the use case WP that was derived throughout the survey process.

Table 6 compiles the most significant information gathered from all answers to the survey questionnaire. The main security needs that can be derived from both are highlighted as follows:

- The Data Stores (Phenotator DB, BioMedBridges Webmicroscope), and Processes (Phenotator, Image Browsing), have restricted access (requires user authentication and authorization).
- One Data Store (Webmicroscope) and one Process (Image Browsing) requires the pseudonymity of patient data.

WP6 includes also four entities, out of which two present restricted access (FIMMWebmicroscope, HMGU) and two are open (Ensembl, Mitocheck).

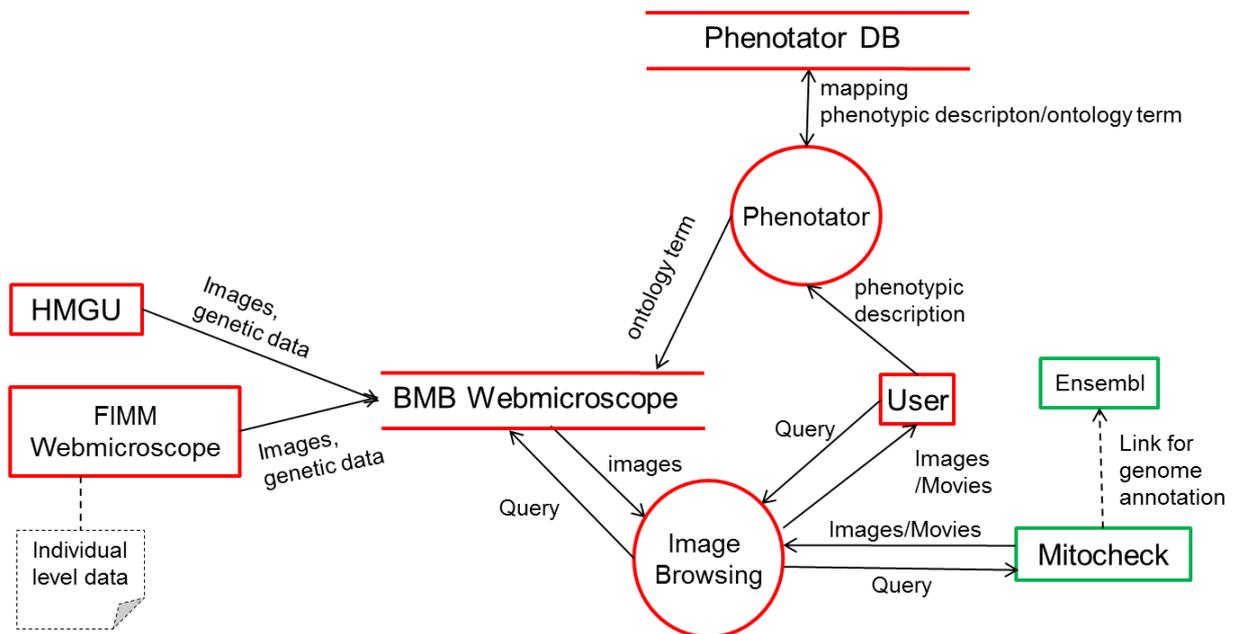


Table 6. Summary of answers to survey questionnaire by WP6.

WP6 DFD Data Store Elements				
Name	Type of Data	Individual Level Data	Access Mode	Security Measures
Phenotator DB	Annotations	None	Restricted	User authentication and authorization system
BMB Web Microscope	Cellular/Tissue images, Genetic data (Human/Mice)	Cellular/Tissue images (not clear if it can be considered individual)	Restricted	User authentication and authorization system
		Usage scenario described that patient related data are stored in pseudonymous form.		Pseudonymous identity management for patient data
WP6 DFD Process Elements				
Name	Input Data	Output Data	Security Measures	
Phenotator	Phenotypic description	Mapping of phenotypic description to ontology term(s)	User authentication and authorization system	
Image Browsing	Query for Images	Images (for Cell/Tissue)	User authentication and authorization system	
			Pseudonymous identity management for patient data	
WP6 DFD Data Flow Elements				
Name	Source	Destination	Security Measures	
Images, genetic data	HMGU	BMB Webmicroscope	Unclear. Considering manual uploading of images and annotations	
Images, genetic data	FIMM Webmicroscope	BMB Webmicroscope	Unclear. Real data transfer may be needed or just the web reference giving access rights to certain users for reading datasets from the source	
Images/ Movies	Image Browsing	User	No security measures planned	
WP6 DFD Entity (External) Elements				
Name			Access Mode	
Mitocheck			Open	
Ensembl			Open	
HMGU Mouse Clinic			Restricted	
FIMM Webmicroscope			Restricted	
Notes:				



Figure 3. DFD of WP6 that resulted from the answers to the survey questionnaire



6.2.3.3 Work package 7 “PhenoBridge”

This section expands the security requirements for the usage scenario 3: "PhenoBridge related to WP7" section 9.2 of deliverable 5.1. The WP7 use case will bridge ontological phenotypic annotation of mouse and human data. The WP plans to provide its services openly to researchers. The DFD that represents the WP is shown in Figure 4, while Table 7 summarizes the most relevant answers collected. Both the DFD and Table 7 have been revised in the time after WP Survey 2 was conducted in order to reflect developments of the use case that occurred in the meantime. In summary, the identified security requirements for WP7 are:

- No security measures for the one Data Store (INFRAFRONTIER portal), and three Processes (“Database browsing”, “Ontology data mapping” and “Data analysis”) that are involved, may be required.
- The WP includes nine external data source entities, out of which only three are restricted (HMGU, CERM, and Univ. Graz), and the rest are open.



Table 7. Summary of answers to survey questionnaire by WP7.

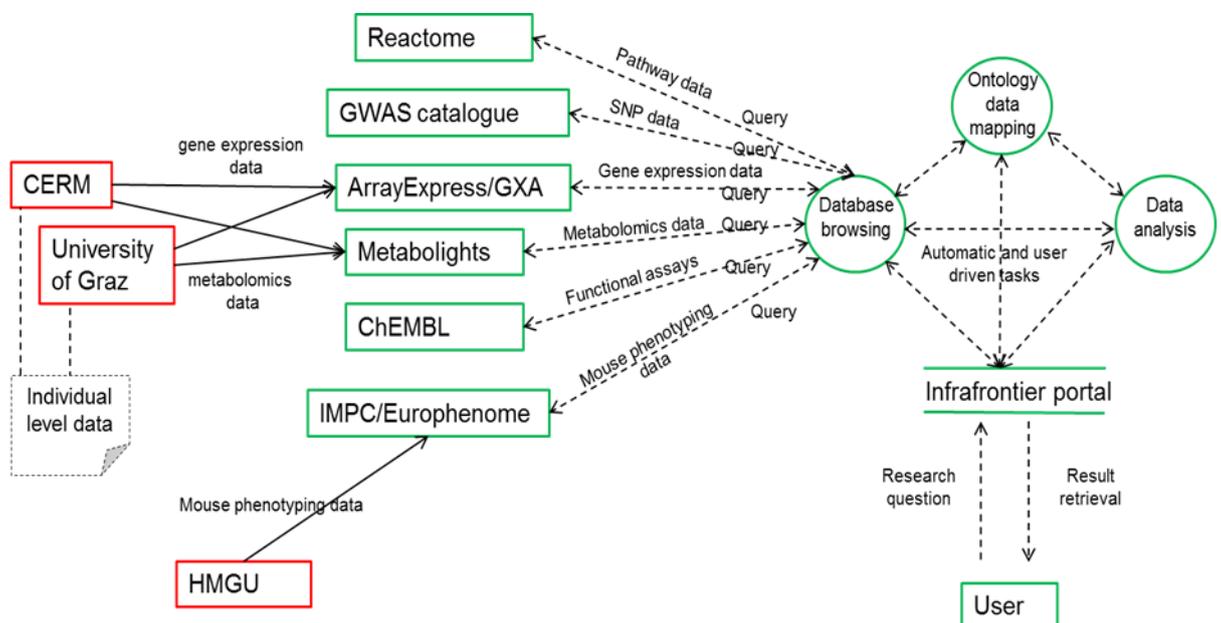
WP7 DFD Data Store Elements				
Name	Type of Data	Individual Level Data	Access Mode	Security Measures
INFRAFRONTIER portal	Intermediate data storage from other open databases in this flow diagram	Human gene expression data, human metabolomics	Open (*)	None
WP7 DFD Process Elements				
Name	Input Data	Output Data	Security Measures	
Data analysis	Datasets (from the two other process elements)	Results of automatic and user-driven analysis by correlation, integration and statistical validation.	None	
Database browsing	Type of identifier (e.g. Gene, Uniprot, Ontology term etc.)	Mouse phenotyping, data, metabolomics/metabolites data, transcriptomic data, GWAS data, pathway data		
Ontology data mapping	Data sets annotated with ontologies			
WP7 DFD Data Flow Elements				
Name	Source	Destination	Security Measures	
Mouse phenotype data, gene expression data, metabolomics data	CERM, University of Graz and HMGU	ArrayExpress/ GXA	None. Data flow is achieved using manual upload of data to respective data sources in the preparatory phase of the use case	
		Metabolights		
		IMPC/EUROPHENOME		



WP7 DFD Entity (External) Elements	
Name	Access Mode
IMPC/EUROPHENOME	Open
Array Express/ Gene expression	Open
Atlas (GXA)	Open
Metabolights	Open
ChEMBL	Open
Reactome	Open
GWAS catalogue	Open
University of Graz	Restricted
CERM	Restricted
HMGU	Restricted

Notes:
 (*) This data originates from open access databases and is not being persisted

Figure 4. DFD for WP7 “Phenobridge”



6.2.3.4 Work package 8 “Personalized Medicine”

This section expands the security requirements for the usage scenario 3: "Personalised Medicine related to WP8" section 9.3 of deliverable 5.1. The WP8 use case aims at integrating complex Personalized Medicine (PM) data sets to understand disease pathogenesis and improve biomarker and treatment selection. In this case, Table 8 reports the overall information derived from the survey questionnaire, and it is supported by the DFD for WP8 in Figure 5. The security requirements identified for WP8 based on both Table 8 and Figure 5 are:



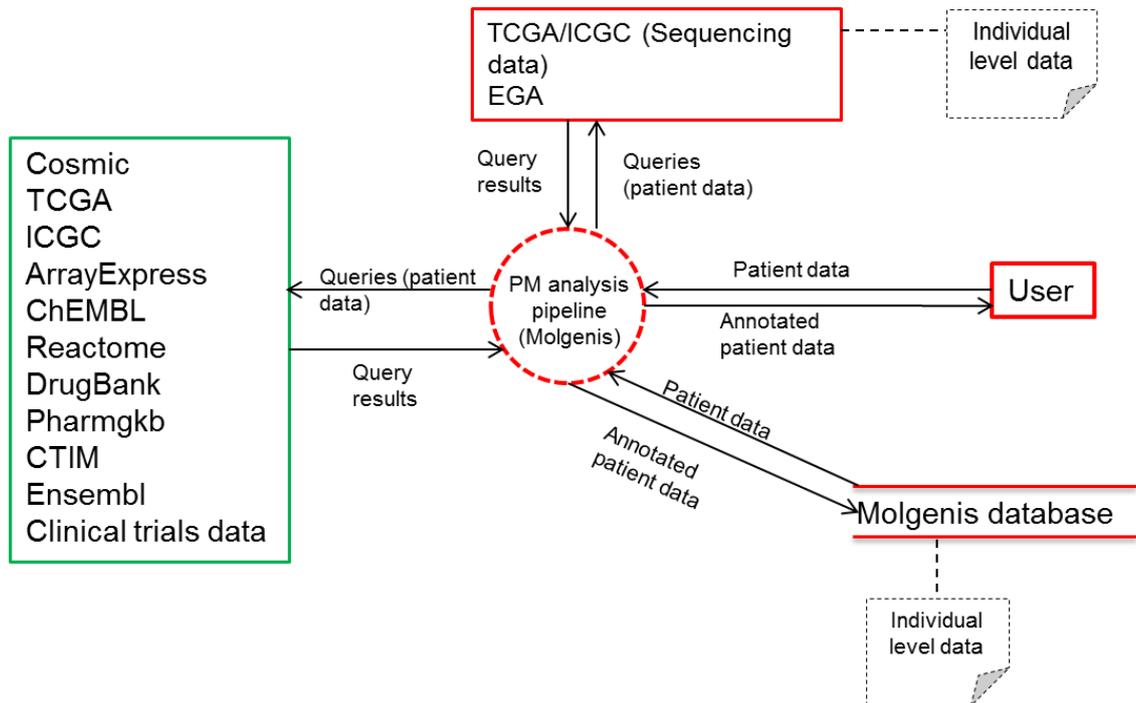
- The one Data Store (PM analysis DB) and the one Process (PM analysis pipeline) are planned to be restricted and require the pseudonymity of PM data.
- The WP utilizes several external data source entities. The majority are available via open access, while 4 are restricted (EGA, ICGC, TCGA, and CTIM).

Table 8. Summary of answers to survey questionnaire by WP8.

WP8 DFD Data Store Elements				
Name	Type of Data	Individual Level Data	Access Mode	Security Measures
PM analysis database	Patient data, Drug data and Lab measurements, Gene expression, DNA Sequence, mutation, diagnosis, etc	Patient demographic, clinical data and genetic data	Open as well as Restricted	User authentication and authorization system (*)
				Pseudonymous identity management for PM data (*)
WP8 DFD Process Elements				
Name	Input Data	Output Data	Security Measures	
PM analysis pipeline	Patient data (e.g. Gene expression, DNA sequence, mutation data, diagnosis, etc.)	Annotation about input data	User authentication and authorization system (*)	
			Pseudonymous identity management for PM data (*)	
WP8 DFD Data Flow Elements				
Name	Source	Destination	Security Measures	
Patient data	User	PM analysis tool	Needs to be secured because it is patient data	
Annotated patient data	PM analysis tool	User		
WP8 DFD Entity (External) Elements				
Name		Access Mode		
Cosmic		Open		
ICGC		Open		
TCGA		Open		
ICGC/TCGA (Sequencing data) (***)		Restricted		
EGA(***)		Restricted		
Array Express		Open		
Clinical Trial Information Mediator (CTIM)		Open		
Drugbank, ChEMBL, Pharmagkb		Open		
Notes:				
(*) refers to security measures to be implemented in the future within the use case.				
(**) the access mode indicates: “Open” if the data access is free; or “Restricted” if the data access is granted only after user registration with the external data source.				
(***) WP8 currently don’t want to access the restricted data from entity namely (EGA, ICGC and TCGA) for the analysis pipeline.				



Figure 5. DFD for WP8 "Personalized Medicine"



6.2.3.5 Work package 10 “Biological Sample Data Integration”

This section expands the security requirements for the usage scenario 5: "Biological Sample Data Integration related to WP10" section 9.5 of deliverable 5.1. The WP10 use case aims at demonstrating the feasibility and providing a prototype for linking Biological Sample database with biobank data. The results of the survey are summarized in Table 9 and the corresponding DFD in Figure 6.

The security characteristics identified for WP10 based on the information from Table 9 and Figure 6 are as follows:

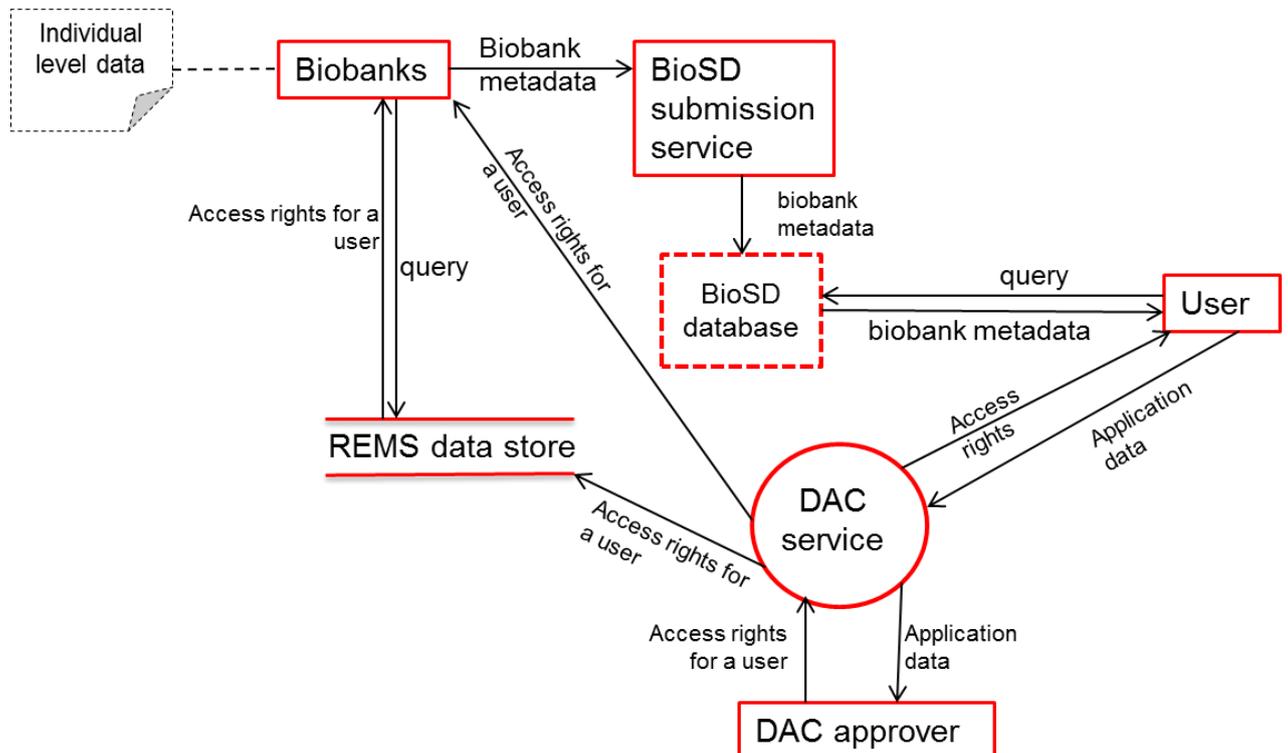
- Use of a Data Store to manage authorization services, e.g. Resource Entitlement Management System (REMS), and a Process consisting of a Data Access Committee (DAC) service (possibly enabled via REMS as well). Both Data Store and Process require user authentication and authorization as security measures.
- Use of three external data source entities of biosample data, two of them restricted, and one (BioSD database), which is both open and restricted.



Table 9. Summary of answers to survey questionnaire by WP10.

WP10 DFD Data Store Elements				
Name	Type of Data	Individual Level Data	Access Mode	Security Measures
REMS	Data access application	Data from biobanks	Restricted	(*) User authentication and authorization system
WP10 DFD Process Elements				
Name	Input Data	Output Data	Security Measures	
DAC service - possibly REMS	Web form based data access application	Access rights	(*) Authentication/ authorization (depends on a DAC application process)	
WP10 DFD Data Flow Elements				
Name	Source	Destination	Security Measures	
Access rights for a user	REMS data store	DAC service	The security measures essential for data flow will be dependent on a concrete DAC application process	
WP10 DFD Entity (External) Elements				
Name		Access Mode		
Biobank BioSD submission service BioSD database		Restricted Restricted Open as well as Restricted		
Notes: (*) refers to security measures <i>to be implemented</i> in the future within the use case.				

Figure 6. DFD for WP10 "Biological Sample Data Integration"





6.2.4 Work package 9 “Structural Data Bridge”

As indicated previously, the use case WP9 “Structural Data Bridge” could not take part in WP5 Survey 2. The causes were mainly organizational and due to early scheduling, given that significant aspects of the intended functionality of WP9 were not yet defined at the time of the survey.

Therefore, a different approach was followed, to attempt to capture the security requirements for WP9 at the same level of detail as that for the rest of the use case WPs involved in the survey. This approach relies on the information available on WP9 within the project. The main contributor in terms of information available concerning security requirements for WP9 was Deliverable 5.1, Section 9.4. “Usage Scenario 4: Structural Data Bridge related to WP9”.

Section 10.2.6 of Deliverable 5.1 already states that the aim of the WP9 usage scenario is to build an “analysis bridge” between the RIs INSTRUCT and ELIXIR, where:

- INSTRUCT will provide access to experimental facilities
- ELIXIR will provide access to information from various publicly available databases
- BMB will enable the link between these infrastructures and facilitate data mining

The security requirements presented here can be found in Section 10 “Requirements Clusters” of deliverable 5.1. The requirements are presented throughout several subsections of deliverable 5.1 devoted to WP9, specifically section 10.1.6, 10.2.6, 10.3.5, and 10.4.6; and they address issues in the areas of “data protection/privacy”, “data security”, “intellectual property and licenses”, and “security of biosamples” respectively.

The rest of this section contains the security requirements referred to the above, grouped by the expected data bridge for which they are relevant. Based on the description of the WP9 usage scenario available on section 9.4.8 of D5.1 and the notion of data bridge explained in section 5, we expect three different data bridges to be required. The three data bridges and the



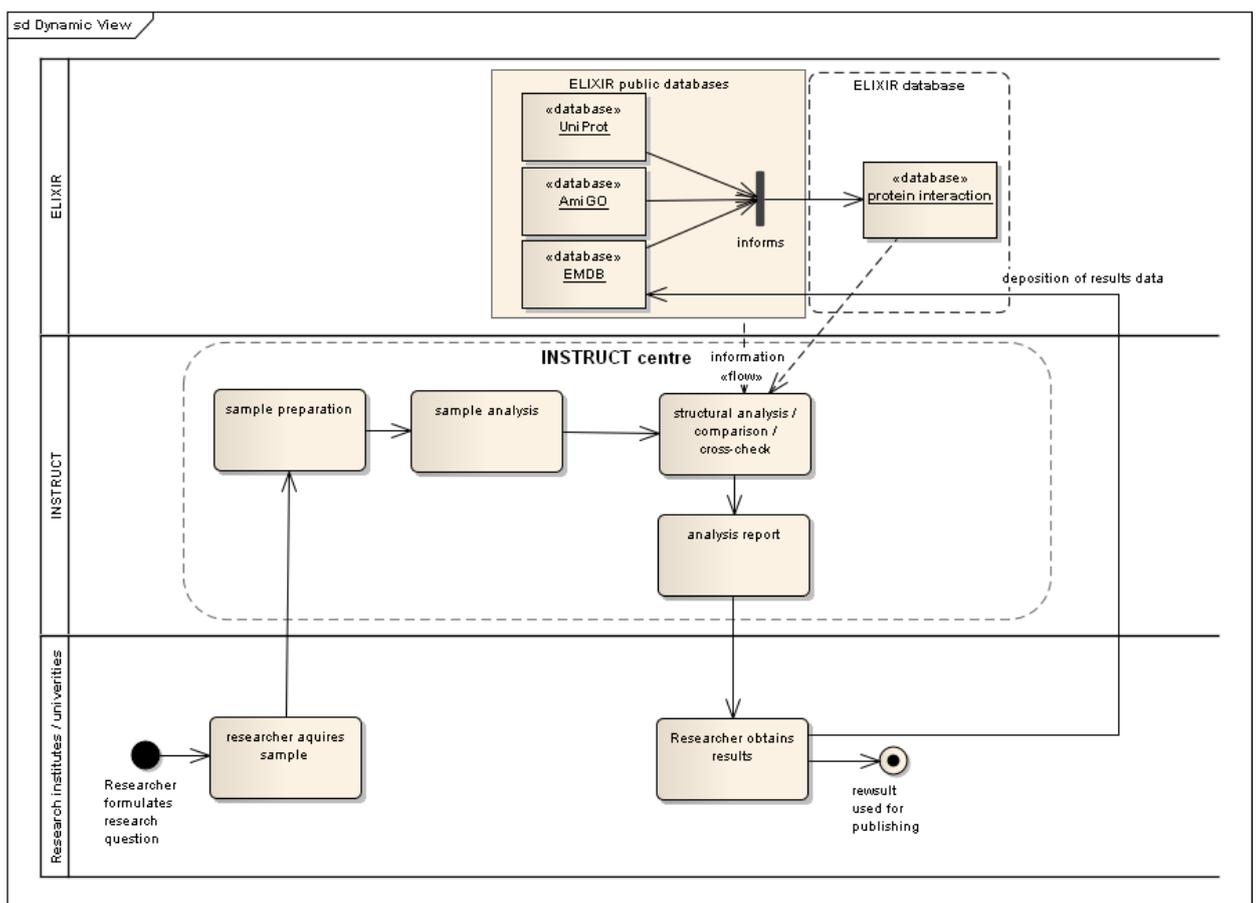
relevant security requirements are described below. The data bridges are listed in the order in which they are expected to be needed to fulfil the complete workflow of a typical WP9 usage scenario.

6.2.4.1 Work package 9 workflow diagram

To help illustrating the three data bridges, the workflow diagram that represents the usage scenario of WP9 is echoed once again in the figure below. The source is Figure 8 in section 9.4.8 of deliverable 5.1. The components and databases that appear in the figure are described in section 9.4 of deliverable 5.1. The usage scenario is summarized as follows in the caption of Figure 8 in deliverable 5.1:



“Report to deliverable 5.1, Figure 8: Data flow of the usage scenario **Structural Data Bridge**. The researcher formulates a research question related to his sample. This sample is prepared and analysed at an **INSTRUCT** centre by the researcher, where initial results on the structure are obtained. This initial structure is then linked to various protein interaction databases at **ELIXIR**, which allow for a more detailed structural resolution. The researcher now holds a final structural report. It is anticipated that the researcher will deposit this data before publishing (arrow back to the **EMDB**²⁷).”



²⁷ EMDB: Electron Microscopy Data Bank (<http://www.ebi.ac.uk/pdbe/emdb/>)



6.2.4.2 Data Bridge 1: From Researcher to INSTRUCT

Data Provider: Research institute/University.

Data Consumer: INSTRUCT.

Security and privacy requirements relevant to the data bridge:

1. The Structural Data Bridge itself does not deal directly with biosamples or biosample related data, and thus no biobanking is involved; but the researchers only use the equipment of an INSTRUCT centre.
2. The researcher has already clarified if consent allows using the sample for structural analysis, including sharing of the sample analysis via the EMDB²⁷ of ELIXIR.
3. Any material (biological sample) that is left by the researcher, no longer required and thus not taken back to his institution shall be disposed of safely, so that neither security of identities nor of research projects is compromised.
4. The researcher who brings a sample to an INSTRUCT centre has to ensure that he is allowed to use this sample and the according data for research.
5. INSTRUCT may control consent/ Ethics Committee approval for human samples.

6.2.4.3 Data Bridge 2: From ELIXIR to INSTRUCT/Researcher

Data Provider: ELIXIR

Data Consumer: INSTRUCT and the research institute/university

Security and privacy requirements relevant to the data bridge:

1. The Structural Data Bridge uses the ELIXIR databases UniProt²⁸, AmiGO²⁹, EMDB²⁷, IntAct³⁰ and GenBank³¹. All these databases are open access.

²⁸ <http://www.uniprot.org/>

²⁹ <http://amigo.geneontology.org/cgi-bin/amigo/go.cgi>

³⁰ <http://www.ebi.ac.uk/intact/>

³¹ <http://www.ncbi.nlm.nih.gov/genbank/>



2. No further data protection requirements have to be considered, nevertheless, the access rules governing the use of these databases have to be respected.
3. Open access data will be transferred to the researcher, while data security mechanisms established by ELIXIR are maintained.
4. These ELIXIR databases involved in the data bridge do not deal with biosamples.

6.2.4.4 Data Bridge 3: From INSTRUMENT/Researcher to ELIXIR (EMDB)

Data Provider: INSTRUMENT and the research institute/university

Data Consumer: EMDB (ELIXIR)

Security and privacy requirements relevant to the data bridge:

1. No data, other than already publicly available through third party databases, will be stored or shared. Thus, the only processes that require attention are the security of data transmission (including experimental raw data and metadata) between INSTRUMENT and ELIXIR, and the disposal of data and the related hard copy material at INSTRUMENT and ELIXIR.
2. The Structural Data Bridge has to ensure that data transfer between INSTRUMENT and ELIXIR is secure and not vulnerable to interception by third parties.
3. This particularly concerns metadata relating to the nature of a substance, the name of the experiment, the researcher's name and any other metadata or data that may identify the nature of the substance and its physiological role. (All these could not only deploy the researcher of his competitive advantage over others but potentially lead to unwanted public disclosure, which may prevent patenting of a substance).
4. The Structural Data Bridge has to ensure that the researcher's identity and the identity of sample providers are safe. The researcher's identity might give away information about the nature of the project (e.g. person X normally works on topic Y, so this information may be related to an innovation Z).



5. For the use of human biosamples it is necessary to get consent and get an Ethics Committee Approval. The approval must include transferring the analysis data to the EMDB database of ELIXIR.
6. Structural data that is stored in the EMDB database of ELIXIR is anonymised. (It may contain sequence information. This is usually the amino acid sequence, which is more general than the DNA sequence).

6.3 Summary

Both the results of WP5 Survey 1 and the results of WP5 Survey 2, including the tables summarizing the answers to the questionnaire of each UC WP, and moreover the DFDs obtained throughout this process, will serve as the basis to perform the threat and risk analysis to be covered in the following section.

In the case of WP9, the information available concerning the security and privacy requirements of the WP outlined above will be used to assess the relevant threats and risks to be considered.

7 Threat and risk analysis for sharing data or biomaterials

This section corresponds to task WT6 of WP5 in the BMB DoW [16] and to emphasize this aspect it uses the same title. The section covers the methodology and results of the threat and risk assessment performed for the UC WPs. The threat and risk assessment leverages the results put forward in section 6 by WP5 Surveys 1 and 2.

7.1 Methodology

This section links together, in the context of each use case WP, the main concepts introduced in Section 4 going from security requirements, to threats, risks, and security countermeasures.

The process adopted to conduct the threat and risk analysis in the BMB project, is based on (i) the STRIDE [8] methodology designed to model



security threats; and (ii) extended with the LINDDUN [7] methodology that focuses on threats to privacy; together with (iii) recommendations to determine the risk of threats defined by the NIST Special Publication 800-30 [4]. This process that in the original specification consists of nine steps (cf. chapter 9 in [8]) has been adapted for the UC WPs of BMB to the seven steps listed below:

1. Define use scenarios
2. Gather a list of external dependencies, as-is state of security measures and security assumptions
3. Create one or more DFDs of the application being modeled
4. Determine threat types
5. Identify the threats to the system
6. Determine risk
7. Plan mitigation

Tasks performed as part of the WP5 Survey 1 and 2 covers steps 1-3 including the DFDs presented in section 6. Step 4 is fulfilled by the security and privacy threats addressed by STRIDE and LINDDUN respectively. Briefly recalling from section 4, these are:

- STRIDE, an acronym for the security threats Spoofing, Tampering, Repudiation, Information Disclosure, Denial of service, Elevation of privilege.
- LINDDUN, an acronym for the privacy threats Linkability, Identifiability, Non-repudiation, Detectability, Disclosure of information, Content Unawareness, Policy and consent non-compliance.

Step 5 implies identifying which STRIDE and LINDDUN threats apply to which components of the DFDs. However, not all STRIDE or LINDDUN threats are applicable to all basic element types of a DFD (i.e., data stores, processes, data flows, external entities)

Table 10 and Table 11 specify which threats may affect each DFD element type according to the STRIDE and LINDDUN methodology respectively. To perform step 5, every threat under both methodologies was mapped to the relevant elements of the DFDs that resulted from WP5 Survey 2 for each UC WP (i.e. WP6, 7, 8, and 10).



Table 10. Mapping STRIDE security threats and countermeasures to DFD element types (see Tables 9-5 and 9-8 in Chapter 9 of [8])

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	
DF: Data flow, DS: Data store, P: Process, EE: External Entity					

Table 11. Mapping LINDDUN privacy threats and objectives to DFD element types (see Tables 4 and 6 in [7])

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 and consent compliance	Policy/consent noncompliance	X	X	X	
DF Data flow, DS Data store, P Process, EE External Entity					

For the risk assessment referred to in step 6, the risk assessment framework put forward by the NIST Special Publication 800-30 [4] has been used as a guideline and tailored for our scope. According to this framework, it is essential to consider the risk factors and the relationships among them. The NIST SP 800-30 considers typical risk factors such as threat, vulnerability, impact, likelihood and predisposing condition.

To present the risk assessment results, a table based on the template introduced in Appendix I of [4] to report “adversarial risk” (see Table I-5) was created for each UC WP involved in WP5 Survey 2. Table 12 indicates the structure of the template used to report the threat and risk assessment results in BMB.



Table 12. Template used to report the risk assessment of threats for BioMedBridges (cf. Table I-5 in Appendix I of [4])

Threat Event	Threat Sources	Vulnerabilities and Predisposing Conditions	Likelihood of Threat	Level of Impact	Risk	Counter-measures (Elements of the Security Architecture)
STRIDE and LINDDUN	(see below)	(see below)	Low, Medium, High	Low, Medium, High	Low, Medium, High	(see below)

Next, the columns that are part of the Table 12 template and a brief explanation of the expected values to populate them are described:

- (i) Threat event (i.e. the threats addressed by STRIDE and LINDDUN)
- (ii) Threat sources: (a) external (individuals or groups that seek to exploit vulnerabilities); (b) processing (data flow or processes); (c) storage (data store); and (d) organization (organizational vulnerabilities, e.g. non-compliance with the regulations, policies etc.)
- (iii) Vulnerabilities and Predisposing Conditions
- (iv) Likelihood of occurrence of threat event. For the purpose of BMB a qualitative scale was adopted consisting of three values: Low(+), Medium(++), High(+++)
- (v) Level of impact. Following the same criteria as in the previous item, it is defined by the three qualitative values: Low(+), Medium(++), High(+++)
- (vi) Level of risk. As defined in Section 4.1 it is determined as a function of the likelihood of occurrence of a threat event, item (iv) above, and its level of impact, item (v).



- (vii) Table 13 specifies all possible values of this risk level function, whose range of values is: Low(+), Medium(++), High(+++)
- (viii) Countermeasures. As reviewed in Section 4, these are the security and privacy measures that are identified to avoid, prevent, mitigate, or minimize the risk of a given threat.



Table 13. A qualitative measure of risk assessment.

Level of Risk			
Likelihood of Threat	Level of Impact		
	Low (+)	Medium (++)	High (+++)
Low (+)	+	+	++
Medium (++)	+	++	+++
High (+++)	+	++	+++

The first six columns are derived from the Table I-5 template in Appendix I of [4]. The seventh and last, was added to capture the relationships across security threat, associated risk, and security countermeasure, and it paves the way to elicit the relevant components of the security architecture and framework for BMB described in Section 8.

7.2 Risk assessment results

This section presents the results of the risk assessment performed for the UC WPs of BMB. For the UC WPs that participated in WP5 Survey 2 (i.e. WP6, 7, 8, and 10) the risk assessment is initially reported via a table based on Table 12 as described in the previous section. These tables (Table 20 to Table 23) are reproduced in Appendix 12.1 of this deliverable 5.3.

As pointed out in step 5 of the risk assessment process adopted, for each element in the DFDs of the UC WPs the relevant threats that may affect it have been identified. Therefore, some of the threats addressed by STRIDE and LINDDUN, may appear several times in the risk assessment results table for a given UC WP.

For example, the STRIDE threat “repudiation” appears three times in Table 19, the risk assessment table of WP6, given that it applies to three different components of the DFD associated to the same WP. Or for example, the case of the LINDDUN threat “identifiability” that appears five times in Table 21, the risk assessment table of UC WP8. The reason is the same; the threat is relevant to five components of the DFD that models the personalized medicine use case, WP8.

In order to provide a concise view of these fine-grained results, the data reported in each of the tables mentioned (i.e., Table 20 to Table 23) have been grouped or aggregated based on the following criteria:



- (i) Threats that affect only the “data flow” element type of the DFD of the UC WPs, whether the threat is addressed by STRIDE or LINDDUN.
- (ii) Threats addressed by STRIDE only that affect the rest of the element types (i.e. “data store”, “process”, “external entity”) of the DFD of the UC WPs.
- (iii) Threat addressed by LINDDUN only that affect the rest of the element types (i.e. “data store”, “process”, “external entity”) of the DFD of the UC WPs.

In all three grouping scenarios above, when the threats involved in a particular grouping have different risk assessment values, the highest risk assessment found is the value assigned to the grouping. This rationale assures to account for the worst case scenario for a given threat.

To illustrate the aggregation or grouping criteria, let us revisit the two examples given earlier, i.e., the “repudiation” threat in Table 19, and the “identifiability” threat in Table 21. In the first instance, the three occurrences of “repudiation” whose risk assessment values are low, medium, and low respectively, are grouped into one occurrence with the highest risk assessment of the three, i.e., medium. Likewise, in the second example, the five cases of “identifiability” are summarized into one whose risk assessment value is high, the highest risk value of the five.

As a result of this grouping, the four original tables (Table 20 to Table 23 in Appendix 12.1) have been condensed into three, i.e., one for each point in the aggregation criteria (i), (ii), (iii) above. The original table structure given in Table 12 has also been adapted for conciseness. The new structure includes: (i) the threat under consideration, (ii) a representative example describing the threat, (iii) the overall value of the risk assessment for each one of the UC WPs, and (iv) the corresponding security or privacy countermeasure(s) required in order to prevent such threat. The three final tables are Table 14, Table 15, and Table 17; and they are presented below.

Note that not all threats addressed by STRIDE and LINDDUN appear in the tables (e.g. spoofing on Table 14, non-repudiation on Table 17). Threats not listed in the tables did not apply to any of the WPs surveyed.



Revisiting the examples given earlier, i.e., the three cases of the “repudiation” threat in Table 19 for WP6, and the five instances of “identifiability” threat in Table 21 for WP8, they appear in the new tables only once respectively as per the grouping criteria followed. Thus, Table 15 shows the “repudiation” threat of WP6 with a risk assessment value of “medium” or “++”, while Table 17 includes the “identifiability” threat of WP8 with a risk assessment value of “high” or “+++”.

The threat and risk assessment of UC WP9 was performed differently, given that the WP could not take part in WP5 Survey 2, and therefore a formal definition of a DFD representing the use cases of WP9 could not be developed collaboratively. In that sense, we attempted to provide an educated estimate of the expected threats and the associated risk levels based on the information available concerning the security aspects of WP9. This estimation was based mainly on two sources:

- (i) The security and privacy requirements, the usage scenario, the workflow diagram, and the expected data bridges, originally reported throughout deliverable 5.1, and summarized here in section 6.2.4.
- (ii) A critical comparison of these security and privacy requirements of WP9 with respect to all the data and information gathered from the rest of UC WPs that participated in WP5 Survey 2 (i.e. WP6, 7, 8, and 10) used to determine the threat and risk assessment of the latter WPs.

The result of this risk assessment estimation for UC WP9 is presented in



Table 14, Table 15, and Table 17 together with the rest of UC WPs in the project.



Table 14. Risk assessment for threats (STRIDE and LINDDUN) to the “Data Flow” element of the DFD.

Threat to „Data Flow“	Example	Risk					Counter-measure
		WP 6	WP 7	WP 8	WP9 ⁽³²⁾	WP 10	
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 15. Risk assessment for security (STRIDE) threats to the “Data Store”, “Process”, and “Entity” elements of the DFD associated to the Use Case WPs.

Security Threat (STRIDE)	Example	Risk					Counter-measure
		WP 6	WP 7	WP 8	WP 9 ⁽³²⁾	WP 10	
Spoofing	Pose as something or somebody else	++	+	++	+	++	- Authentication System - Configuration Management
Tampering	Malicious modification of data or code	++	+++	+++	++	++	- Authorization System
Repudiation	Denial of having received data	++	+	++	+	+	- Auditing and logging
Information disclosure	Exposure of information to unauthorized individuals	++	-	+++	+++	+++	- Authorization System - Input Validation
Denial of service	Resources are not available due to overload or attack	+	+	+	+	+++	- Configuration Management - Input Validation
Elevation of privilege	A user gains unauthorised access to resources	++	-	+++	+	+++	- Authorization System
- (not relevant), + (low), ++ (medium), +++ (high)							

³² The values of the threat and risk analysis for the UC WP9 are an estimation based on the information available of the WP. Further details are provided in this section, and in Section 6.2.4



Table 16. Risk assessment for privacy (LINDDUN) threats to the “Data Store”, “Process”, and “Entity” elements of the DFD associated to the Use Case WPs.

Privacy Threat (LINDDUN)	Example	Risk					Counter-measure
		WP 6	WP 7	WP 8	WP 9 (32)	WP 10	
Linkability	Possibility to detect that different data items are related to the same entity	++	-	++	++ +	++	- Anonymization tool -
Identifiability	Possibility to relate a set of data to a specific entity / person; to recognize a person by characteristics of data	++ +	+	++ +	++ +	++	- Pseudonymization modules - Encryption - Access control system
Content unwareness	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 - Data Use Agreement - Material Transfer Agreement
- (not relevant), + (low), ++ (medium), +++ (high)							

From the survey and the threat and risk analysis results, it could be concluded that:

- Security threats addressed by the STRIDE methodology concern all UC WPs involved in the survey (WP6, 7, 8, and 10).



- Privacy threats addressed by the LINDDUN methodology concern all UC WPs involved in the survey (WP6, 7, 8, and 10).
- Secure data transfer methods are not yet in place.

8 Design of the security architecture and framework

Based on the results of the security requirements identified within WT5 and the risk analysis conducted in the context of WT6, we designed a security architecture and framework for BioMedBridges in the context of WT7 that facilitates the exchange of biomedical data across research infrastructures in a secure and privacy preserving manner. The security architecture involves countermeasures for mitigating the threats identified in the risk analysis of the UC WPs and proposes a generic workflow that incorporates these measures and suggests implementations for realizing them. This generic workflow can be instantiated in order to obtain guidelines for the secure realization of concrete data exchange scenarios within BioMedBridges and hopefully also beyond the scope of this project.

The development of the security architecture has been aided by literature review and an evaluation of existing security solutions in the field of biomedical research (see chapter 3.1). In particular, the security solution employed by the European Genome-phenome Archive (EGA)³³ turned out to constitute a proven, operating system that covers many aspects that are similar to the security and privacy challenges BMB is faced with. Furthermore, since the BMB project partner EMBL-EBI operates EGA, considerable know-how and access to implementations related to the technologies used in EGA is available. Consequently, the EGA highly influenced our work, and many of the suggested implementations for realizing security and privacy measures could be derived from it. Discussions with the European Grid Infrastructure (EGI)³⁴ also gave us valuable inputs for the design of the security architecture. Within BioMedBridges, the data bridge from BBMRI to ELIXIR in

³³ <https://www.ebi.ac.uk/training/online/course/genomics-introduction-ebi-resources/european-genome-phenome-archive-ega>

³⁴ <http://www.egi.eu/>



the context of WP4, which also serves as preparatory work for UC WP10 and for planning the pilot implementation in WT8 summarized in section 9, influenced the design of the security architecture and vice versa. Furthermore, the process specification for secure sharing of and access to personalized medicine data in the context of deliverable 8.1 that is briefly described in section 9.2 can be regarded as an instantiation of the security framework and had strong synergy effects with respect to the design of the security architecture.

In the following, the design of the security architecture and framework is presented. Section 8.1 provides a high-level overview that builds on the concepts introduced in section 5, section 8.2 describes countermeasures for mitigating the threats identified in section 7 as well as implementation suggestions for realizing them, and section 8.3 puts forward a secure generic data bridge that brings everything together in the form of a generic workflow that constitutes a blueprint for implementing secure data exchange.

8.1 Overview of the security architecture and framework

Referring back to the notions introduced in section 5, the scope of the security architecture and framework covers data bridges between different research infrastructures. One of them acts as a data provider, while the other RI acts as a data consumer. Based on the work conducted in WT5, WT6, and WT7, relevant security and privacy preserving measures have been identified and incorporated into the security architecture in order to facilitate inter-RI research in a secure and privacy preserving manner.

Essential concepts for mitigating security and privacy threats in the context of data sharing are:

- Access control
- Secure release
 - Informed Consent, Ethic Committee and Data Access Committee approval
 - Data and Material Transfer Agreements
 - Pseudonymization and Anonymization
 - Encryption



In the context of data use, which includes the analysis and integration of data received from external sources, access control and secure storage have to be considered.

In addition, the actual transfer of data from the provider to the consumer has to be protected using secure communication.

Concrete countermeasures including implementation suggestions that address these aspects and mitigate the threats identified during the risk analysis are presented in detail in section 8.2.

Regarding the division of responsibilities with respect to secure data transfer, we regard the data provider to be responsible for determining and enforcing the appropriate secure transfer workflow depending on the sensitivity of the requested data. All relevant information about laws, regulations, and access rights has been drawn from deliverable 5.2. To cover regulatory and legal aspects comprehensively, the security framework comprises three data access tiers. They form logical layers reflecting the legal/regulatory characteristics of the objects to be accessed and shared:

- *Open/public access tier (access tier 1)*: This access tier contains anonymous data and data which do not contain any other protected (IP related) information. Hence, no protection is required; especially neither authentication nor authorization is needed. Exemplary data for this access tier are metadata like information about the primary purpose of the data collection or schema information, e.g. attribute names.
- *Restricted access tier (access tier 2)*: This access tier contains protected data. These can be anonymous data for which oversight is desired and/or data needing IP protection. This access tier requires authentication (i.e. the user has to login) and agreements to terms and conditions. This agreement can be needed along with account creation or upon each data request. To allow for accountability, authentication is required and access to data is controlled (authorization). Examples of projects or databases following this concept, exist: The Gen2Phen³⁵ project proposes such an access tier

³⁵ <http://www.gen2phen.org/>



to access “quasi-sensitive data” consisting for example of “aggregate data from genome-wide association studies”. ArrayExpress³⁶ and PRIDE³⁷ use the restricted access tier to securely share “private data”, meaning pre-published or unpublished data. The Cancer Genome Atlas (TCGA)³⁸ is an example where no login is required but an agreement to terms & conditions is needed before access is granted. The iDASH data repository³⁹ contains biomedical data that does not contain personal health information. Access to it requires registration and creation of a user account.

- *Committee controlled (access tier 3)*: All security measures of the restricted access tier apply, and additionally a review by a committee (e.g. data access committee, DAC) is needed before the data is released to the requestor. DACs may request the scientific reasons why access is sought. The DACs will also make sure that data access is covered by informed consent and ethics committee approval. Examples of portals having this access tier in place are the European Genome-phenome Archive (EGA)⁴⁰, the database of Genotypes and Phenotypes (dbGaP)⁴¹, the Cancer Genome Atlas (TCGA)⁴², and the International Cancer Genome Consortium (ICGC)⁴³.

The workflows belonging to these data access tiers are further explained in section 8.3.

The data consumer is responsible to comply with all terms, conditions, contracts, and regulations that she/he has accepted and/or signed. Typical examples are Data Use Agreements (DUAs) and Material Transfer Agreements (MTAs) - we refer to the templates available under <http://www.biomedbridges.eu/deliverables/52-0>. Specifically, the data consumer will have to take security measures to comply with these agreements which for their part have to make sure that they are fully considering the original informed consent, ethics committee approval and data access committee decisions. Malin et al. [19] have discussed technical

³⁶ <https://www.ebi.ac.uk/arrayexpress/>

³⁷ <https://www.ebi.ac.uk/pride/archive/>

³⁸ <https://tcga-data.nci.nih.gov/>

³⁹ <http://idash.ucsd.edu/data-repository-0>

⁴⁰ <https://www.ebi.ac.uk/ega/home>

⁴¹ <http://www.ncbi.nlm.nih.gov/gap>

⁴² <https://tcga-data.nci.nih.gov/>

⁴³ <https://icgc.org/>



and policy approaches for data sharing in clinical and translational research. They have made recommendations for the process from data access to sharing and use. We refer to this article and explicitly suggest to follow its recommendations. We further refer to the section 8.2.7 on auditing and provenance.

Concerning open access to aggregate (and anonymous) data, the suggestion [19] is not to post pooled statistical information regarding static, replicable features that are easy to derive from biological information, such as genome-wide SNP scans.

The next suggestion [19] is to establish policies for assessing credentials of data users and committees to institute the policies, together with clear suggestions of potential members of such committees. They suggest to define use agreements, and to describe risks of data aggregation and re-use already during informed consent.

Another relevant suggestion [19] is to formalize liability requirements and procedures for redress. While the principal liability is in the hand of the data controller of the data producer, also the consuming party has to play a role, e.g. when access has to be secured or when an adequate reaction on re-identification on the consumer side is needed.

Auditing practices should be established [19]. Again, we refer to our statements on auditing, accountability, and provenance in the sections 8.2.7 and 4.2.1.

Multiple levels of access are suggested [19]. Compliant with this, this deliverable has presented three tiers. We refer to Malin et al. [19], and we repeat the recommendation that data access committees should be involved for all levels of access.

Focussing on sharing of clinical trial data, Mello et al. [20] have analysed current policies (by EMA, FDA and others) as well as benefits, risks, and legal implications. Benefits include a wider range of analyses and a potential effect on scientific discovery, whereas risks comprise reidentification and loss of intellectual property. From this, they have identified core principles of expanded data sharing and suggested access models, among which are



protection of privacy and intellectual property, accountability, and practicability. They suggest and discuss four possible models for expanded access to participant-level data. Their “Open Access” model still requires the requester to attest that data will not be used inappropriately, which corresponds to the above tier 2. The three further models suggested are variants of the tier 3 approach suggested in this deliverable 5.3: There is a decision maker who may be an independent review board or the trial sponsor. Anonymity, deidentification, and risks of reidentification play a central role, but are not discussed in detail. One of the options presented is that only results of a query, and not the micro data are released.

8.2 Countermeasures of the security architecture and framework

This section describes the countermeasures we derived from the threat and risk analysis of the use cases WPs that are described in section 7. It presents relevant definitions, points out the relations between the countermeasures and the threat types they mitigate, and proposes implementations for realizing them.

8.2.1 Authentication

In [9], authentication is defined as the “provision of assurance that a claimed characteristic of an entity is correct”. It is being frequently applied for user identity verification based on one or more of the following basic “factors”:

- Something the user knows (e.g. password, PIN).
- Something the user has (e.g. ATM card, smart card).
- Something the user is (e.g. biometric characteristic, such as a fingerprint) [21].

Authentication that requires the validation of two or more of these factors is known as multi-factor authentication. Authentication constitutes a countermeasure against spoofing threats.

A service provider can perform authentication locally on its own, e.g. by maintaining a credential store and validating user supplied credentials against it, or delegate this task to a trusted third party called identity provider. Identity



providers can act as trusted common authentication authorities for several different service providers, an approach that enables authentication across organization boundaries and is known as identity federation.

Identity federation can be used to improve user experience by providing single sign-on (SSO) functionality, that is, by allowing a user to provide credentials once per session to the identity provider, and then gain access to multiple service providers without having to authenticate again during that session. As an additional convenience for the user, SSO requires only one initial registration with the identity provider rather than one registration with every service provider involved.

The drawback of identity federation is that identity providers constitute additional points of failure and have to be trusted by all service providers involved. Consequently, for the purpose of the security architecture, we recommend the use of identity federation as an optional possibility for research infrastructures, and suggest that research infrastructures may fall back to perform authentication themselves if no identity provider is available or in case an identity provider is not trusted.

An implementation that can be used for establishing identity federation is Shibboleth⁴⁴, an open source software package for web SSO based on the federated identity standard Security Assertion Markup Language (SAML) [22]. Shibboleth has proven itself in numerous projects, including EGA. Identity federation can also be established using the open standard and protocol OpenID⁴⁵. OpenID is used by several large companies including Google and Microsoft⁴⁶.

8.2.2 Authorization

Authorization is defined as “an approval that is granted to a system entity to access a system resource [5]”. By controlling access to resources, authorization counters the following threat types: tampering, information disclosure, and elevation of privilege.

⁴⁴ <http://shibboleth.net/>

⁴⁵ <http://openid.net/>

⁴⁶ <http://openid.net/2014/02/26/the-openid-foundation-launches-the-openid-connect-standard/>



According to [3], two types of access control can be distinguished, depending on the entity that is able to determine the access rights:

- “If an individual user can set an access control mechanism to allow or deny access to an object, that mechanism is a *discretionary access control (DAC)*.”
- “When a system mechanism controls access to an object and an individual user cannot alter that access, the control is a *mandatory access control (MAC)*.”

In the case of DAC, the owner of an object restricts access to it by allowing only particular subjects to access it. As opposed to this, MACs do not allow the owner of an object to specify access rights, and typically take information about both the subject and the owner of the object into account in order to determine whether access is granted or not. The conditions for allowing or denying access are derived from a set of rules, and thus, MAC is sometimes also called *rule-based access control*. Access control mechanisms that grant subjects access to objects based on roles that are assigned to the subjects rather than based directly on the identities of the subjects are known as *role-based access controls (RBACs)*. In systems with numerous users, RBAC can considerably simplify the management of access rights.

As a suitable implementation to support the authorization workflow, we suggest the Resource Entitlement Management System (REMS) [23] as a general element of the security architecture, i.e. for all Use Case Work Packages. The REMS is an open source tool for managing access rights to research resources that assists both researchers requesting data access and data access committees granting access. While REMS is primarily designed to support electronic workflows, it also allows for paper based agreements as typically needed for *the committee controlled data access tier*. It provides a policy repository storing the authorization information. In order to convey authorization decisions, SAML is suggested. REMS is used for example by the EGA for authorizing users to access datasets that are governed by data access committees.



8.2.3 Secure data communication

Regardless of the access tier, we recommend that all data transmitted between a data provider and consumer should be sent over a secure communication channel. If possible, the provider and consumer internal data communication should be protected as well. That means that the data stream should be encrypted, the identity of the communication participants should be verified, and the integrity of the transferred data should be protected. These measures effectively mitigate information disclosure, spoofing, and tampering threats.

Assuming that standard web technology is used, secure communication channels can be established by using the Hypertext Transfer Protocol (HTTP) over standard Secure Socket Layer (SSL) or Transport Layer Security (TLS) connections, commonly referred to as SSL/TLS [24], based on SSL-Certificates that have been issued by trusted certification authorities.

It is worth noting that standard SSL/TLS protects only the communication channel, but provides no further protection of data that has left the communication channel. In this sense, standard SSL/TLS provides only hop-by-hop protection. For example, a request that is sent to a webserver over a secure SSL/TLS connection might cause the webserver to initiate an unprotected communication with a database server in order to obtain information needed to reply. This unprotected communication between webserver and database server can then lead to information disclosure. In order to avoid such scenarios, additional countermeasures such as certificate pinning or end-to-end encryption (see section 8.2.4) could be desired, but require additional implementation effort.

In addition to secure communication channels, we also suggest filtering of incoming messages as a means to ensure secure data communication. This countermeasure can be used for mitigating denial of service threats and may be realized by properly configured firewalls such as the open source tool iptables⁴⁷. Another method that can be used for addressing denial of service threats is load balancing.

⁴⁷ <http://www.netfilter.org/projects/iptables/>



8.2.4 Encryption of data

In addition to merely securing the communication channel as described in section 8.2.3, sensitive data that is to be transferred through this channel can be encrypted itself before the transfer by the data provider so that only cipher code is sent out. This makes real end-to-end security from the data provider to the consumer possible rather than mere hop-by-hop security.

According to [25], encryption of data for the purpose of data transfer can be performed either symmetrically or asymmetrically. In the former case, the key required in order to encrypt the data is the same as the key needed for decrypting it, and thus needs to be known by both the sender and recipient. In the latter case, two different keys are involved, a public one that is used by the sender for encrypting the data to be transferred, and a private key that is known only by the recipient and used for decrypting the received data. An obvious disadvantage of symmetric cryptography is that the key for en- and decryption has to be exchanged between sender and recipient, and the effort of encryption is in vain if the key is transferred over the same communication channel as the encrypted data. By using asymmetric cryptography, this key exchange problem is avoided. However, other problems are introduced, such as the need for public key management and the verification of public key authenticity. For addressing such problems involved with asymmetric cryptography, Public Key Infrastructures (PKIs) could be installed, but this introduces additional implementation effort. Furthermore, the utilization of PKIs leads to additional trust requirements because they typically involve certificates issued by trusted certification authorities.

As a feasible method for the encryption of data to be transferred, we recommend strong state-of-the-art symmetric encryption methods such as the Advanced Encryption Standard (AES) [26] using a unique key that is randomly generated by the data provider for every data transfer. This key can be forwarded to the data recipient using a communication channel that is different from the one used for the actual data transfer. For example, the key can be communicated by phone. In the future, the application of more sophisticated asymmetric cryptography solutions may be evaluated in case the need arises, involving e.g. elliptic curve cryptography, RSA or Diffie-Hellman key exchange [25].



It is worth noting that besides encrypting data for the purpose of transferring it, data may also be persisted in encrypted form, which is known as encryption of data at rest. This has the benefit that confidentiality is protected also in case an attacker succeeds in getting access to data in the database directly. The keys used for the encryption of data at rest should of course be kept private. As a consequence, a data provider that is going to share data that is encrypted at rest has to decrypt it prior to the data transfer, and, in case it should be transferred in encrypted form, re-encrypt it with the public key of the data consumer respective the key shared with the data consumer in case symmetric cryptography is used for the transfer.

8.2.5 Anonymization

Anonymization is defined as a “process that removes the association between the identifying data set and the data subject [27]”. In a manner similar to the encryption of data described in section 8.2.4, anonymization of data can be performed dynamically as a data release preparation, or data can already be anonymized before persisting it.

As described in [28] and [29], 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 [28].



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 [28].
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 [30].

According to [28], 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*, which means that, regarding the quasi-identifiers, each data item within a given data set cannot be distinguished from at least $k-1$ other data items [31].
- *l-diversity*, which means that for each group of records sharing a combination of quasi-identifiers, there are at least l “well represented” values for each sensitive attribute [32]. l -diversity implies l -anonymity.
- *t-closeness*, which means 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 [28].



- 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 δ_{\min} and δ_{\max} [30].

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 [32]. 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 [28].

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 defined in section 4.3.1, k-anonymity and l-diversity mitigate identifiability and linkability threats according to [7].

An open source tool that implements all of the anonymization metrics described above is the ARX toolkit and software library⁴⁸.

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 [33]. While it has been shown that this measure can be defeated by trackers [34], the susceptibility to tracker attacks can be prevented by only allowing predefined/restricted queries to be issued (as suggested here).

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 [35].

⁴⁸ <http://arx.deidentifier.org/>



8.2.6 Pseudonymization

The pseudonymization of data is defined as a process “[...] 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” in [27]. Compared with anonymization as described in 8.2.5, pseudonymization also mitigates the LINDDUN threat types identifiability and linkability according to [7]. 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.

8.2.7 Auditing and provenance

As we already pointed out in section 4.3.1, accountability is of high relevance for biomedical research in general. Consequently, we strongly advise any entity involved in data transfer scenarios within BMB to audit any relevant actions using appropriate logging and reporting services in order to mitigate repudiation threats. We note that relevant actions can also include transactions performed by the provider internally. If the logged actions



contain confidential information the log has to be kept secure and treated like confidential primary data.

Moreover, and in extension to audit trails, we suggest to keep provenance traces and to make systems provenance-aware as far as possible and feasible. Further information and recommendations for the implementation of provenance-aware systems based on experiences gathered within the biomedical research projects EHR4CR and TRANSFoRm can be found in a paper by Curcin et al [10]. Here, we recapitulate some central points from this article:

According to Curcin et al., the de-facto standard representation model for interoperable provenance data is the Open Provenance Model (OPM) [36]. OPM facilitates the modelling of provenance data in the form of graphs, with edges denoting relationships, and nodes representing the individual occurrences of entities. Nodes may represent

- *Artifacts*, which are pieces of data of fixed value and context, e.g. one version of a data set or document,
- *Processes*, which are actions performed using artifacts that generate other artifacts, or
- *Agents*, which are entities controlling process execution that may be humans or non-mutable pieces of software.

Numerous publicly available libraries and tools related to the Open Provenance Model can be found at the OPM website⁴⁹. As an alternative, the PROV data model (PROV-DM)⁵⁰ developed by the World Wide Web Consortium (W3C)⁵¹ is suggested. PROV is strongly influenced by OPM and differs primarily in improved support for the attribution and evolution of entities over time. A list of PROV implementations can be found on the W3C website⁵². According to Curcin et al., both OPM and PROV contain the basic building blocks for provenance-aware systems and when choosing between them, the available tools and libraries in both systems for potential reuse or adaption should be considered.

⁴⁹ <http://openprovenance.org/>

⁵⁰ http://www.w3.org/2011/prov/wiki/Main_Page

⁵¹ <http://www.w3.org/>

⁵² <http://www.w3.org/2011/prov/wiki/ProvImplementations>



Both OPM and PROV aim at representing provenance data in a domain model independent manner. Consequently, those provenance models and the corresponding data have to be linked with domain knowledge models and data, respectively. In the context of TRANSFoRm, this link was established using so-called *provenance templates* which constitute higher-level abstractions of the provenance graph data. The main difference is that their artifact, process, and agent nodes refer not to concrete instances in the past, but to domain concepts that shall be used for instantiation.

8.2.8 Common data access policies

In biomedical research, common data access processes involve regulatory elements such as compliance to

- Data Use Agreements (DUAs),
- Material Transfer Agreements (MTAs),
- Informed Consent (IC),
- Data Access Committee (DAC) approval and
- Ethics Committee (EC) approval

in order to address the ethical, legal, and social implications (ELSI) of data access. The security architecture put forward takes such regulatory elements into account and regards them as countermeasures against the “policy and consent non-compliance” privacy threat. The aspect of informed consent compliance additionally mitigates content unawareness threats.

These countermeasures have contractual and organizational rather than technical character, but can nonetheless be supported by automated data processing. For this purpose, an online tool for the assessment of regulatory and ethical requirements has been developed in WT 4 that is presented in deliverable 5.2. Further helpful tools are, for example, legal WIKI of BBMRI⁵³, and the human Sample Exchange Regulation Navigator (hSERN) tool⁵⁴. REMS can also be used in this context, e.g. for supporting DAC based authorization workflows as already mentioned in 8.2.2.

⁵³ <http://bbmri.eu/wiki-legal-platform>

⁵⁴ <http://www.hsern.eu/>



The availability of the tool described in deliverable 5.2 is of high complementary relevance to the security architecture. Deliverable 5.2 addresses questions of informed consent and comprises a number of templates⁵⁵, including:

- Data transfer agreement – personal data
- Data transfer agreement – non-personal data
- Material transfer agreement – human biosamples
- Material transfer agreement – non-personal biosamples
- Provider agreement – human biosamples
- Provider agreement – non-personal biosamples
- Data provider agreement – personal data
- Data provider agreement – non-personal data
- Information Sheet and Consent Form

The tool enables quick and easy access for data providers and consumers to these templates as well as to legal and regulatory information relevant to the specific contexts.

Initially, this BMB tool was designed to raise the awareness of the biomedical researcher to legal and ethical issues before consuming and sharing data/material stemming from and destined for different contexts. For this task, it guides the user through a structured query process and provides legal and ethical requirements as well as ways to meet these requirements. Usage of the tool and similar resources within the secure workflow of a data bridge (see section 8.3) is especially helpful if the legal requirements are supplemented by corresponding policies (particularly related to the LINDDUN threats identifiability, content unawareness, and policy and consent non-compliance). Such supplements can also include references and advices, e.g. for implementing k-anonymity, or for being informed about pseudonymity solutions in exemplary scenarios. A workshop on “Personal data in the life sciences: helping researchers handle data protection and ethical requirements” will deal with resources and tools for this task, and aims at providing integrative solutions that can be used in the “secure workflow”.

⁵⁵ <http://www.biomedbridges.eu/deliverables/52-0>



8.3 Secure workflow specified by the security architecture

This section puts forward a generic data bridge workflow that models abstract data transfer and can be instantiated in order to build secure data bridges. It constitutes a generalization of the data bridges occurring in use case WPs and serves as a blueprint for implementing these data exchange processes in a secure and privacy preserving manner. We use this generic data bridge as a basis for pinpointing the security and privacy measures presented in section 8.2, as well as suggestions for concrete implementations.

The generic data bridge is presented in the form of three activity diagrams⁵⁶, one for each access tier supported by the security architecture. Figure 7 shows the activity diagram modelling the workflow for the open data access tier,

⁵⁶ http://en.wikipedia.org/wiki/Activity_diagram (last access April 11, 2014)



Figure 8 depicts the activity diagram corresponding to the restricted data access tier, and Figure 9 displays the activity diagram that outlines the committee controlled data access tier. The most important shape types used in the diagrams are explained in Table 17.

Table 17. Components of the workflow activity diagrams.

	Rounded rectangles representing actions
	Diamonds representing decisions
	Rectangles with a stylized donkey ear representing notes
	A black circle representing the start of the workflow
	An encircled black circle representing the end of the workflow

Directions of arrows represent the order in which activities happen. The horizontal “swim lanes”⁵⁷ visually distinguish responsibilities of different entities involved in the workflow.

The workflow starts with the data consumer sending a data request to the provider. This request may have been initiated by a researcher as explained in section 5. The provider is then responsible for determining the relevant data access tier depending on the sensitivity of the requested data.

Regardless of the data access tier, we recommend that all data transmitted between participating entities should be sent over a secure communication channel. We assume that standard web-technology will be used, i.e. the provider possesses a SSL-Certificate. This allows the consumer to verify the identity of the provider and to establish a secure connection, most likely using HTTP over the SSL/TLS protocol. Furthermore, we assume throughout the process that all relevant actions are logged and audited by the responsible entities.

⁵⁷ http://en.wikipedia.org/wiki/Swim_lane (last access April 11, 2014)



In the following, the workflows corresponding to the three data access tiers are explained.

8.3.1 Open data access tier

As shown in Figure 7, no further measures are required in the case of the open data access tier, so the provider can transfer the requested data right away. In other words, there is no need to authenticate to access this type of data.

8.3.2 Restricted data access tier

The restricted data access tier shown in



Figure 8 applies if access to the requested data is restricted. It requires authentication and acceptance of terms and conditions.

The consumer has to authenticate towards the provider. The authentication process can be managed by the data provider directly or it can be delegated to a federated authentication service using implementations such as Shibboleth or OpenID. A federation of federations of identity providers like eduGAIN⁵⁸ is a further option. ELIXIR and the upcoming EINFRA-7-2014 project consider introducing a Level of Assurance (LoA) for the strength of authentication as an overlay to eduGAIN. Identity providers which qualify to the higher authentication standard could subscribe to the overlay. In the latter case, the data provider can optionally choose to deny authentication requests involving untrusted identity providers and fall back to local authentication instead. We make no assumptions about the type of credentials used, i.e. whether they are certificates, passwords, smartcards, or biometric features such as fingerprints etc., and whether multi-factor authentication is employed, since this will likely differ from provider to provider. In order to facilitate authentication directly by the data provider or by a third party identity provider that is trusted by the data provider, we assume that the consumer has been appropriately registered by the data provider or the identity provider in advance. We refer to similar suggestions by Mello et al [20] and Malin et al. [19].

After authentication, the provider asks the consumer to accept data release requirements, such as terms and conditions regarding the usage of the requested data. If agreements have already been made during account creation, this step will not be necessary. Another option is to allow or deny access based on information originating from the authentication process (e.g. based on the membership attribute “eduPersonAffiliation” in order to take the status of the consumer into account).

The next steps involve the authorization of the requesting user (consumer), data release preparations in accordance to the release requirements by the provider (e.g. anonymization, encryption of the data to be transferred etc.)

⁵⁸ <http://www.geant.net/service/eduGAIN/Pages/home.aspx>



and the actual transfer of the requested data. The consumer may then store and use the data according to the release requirements.

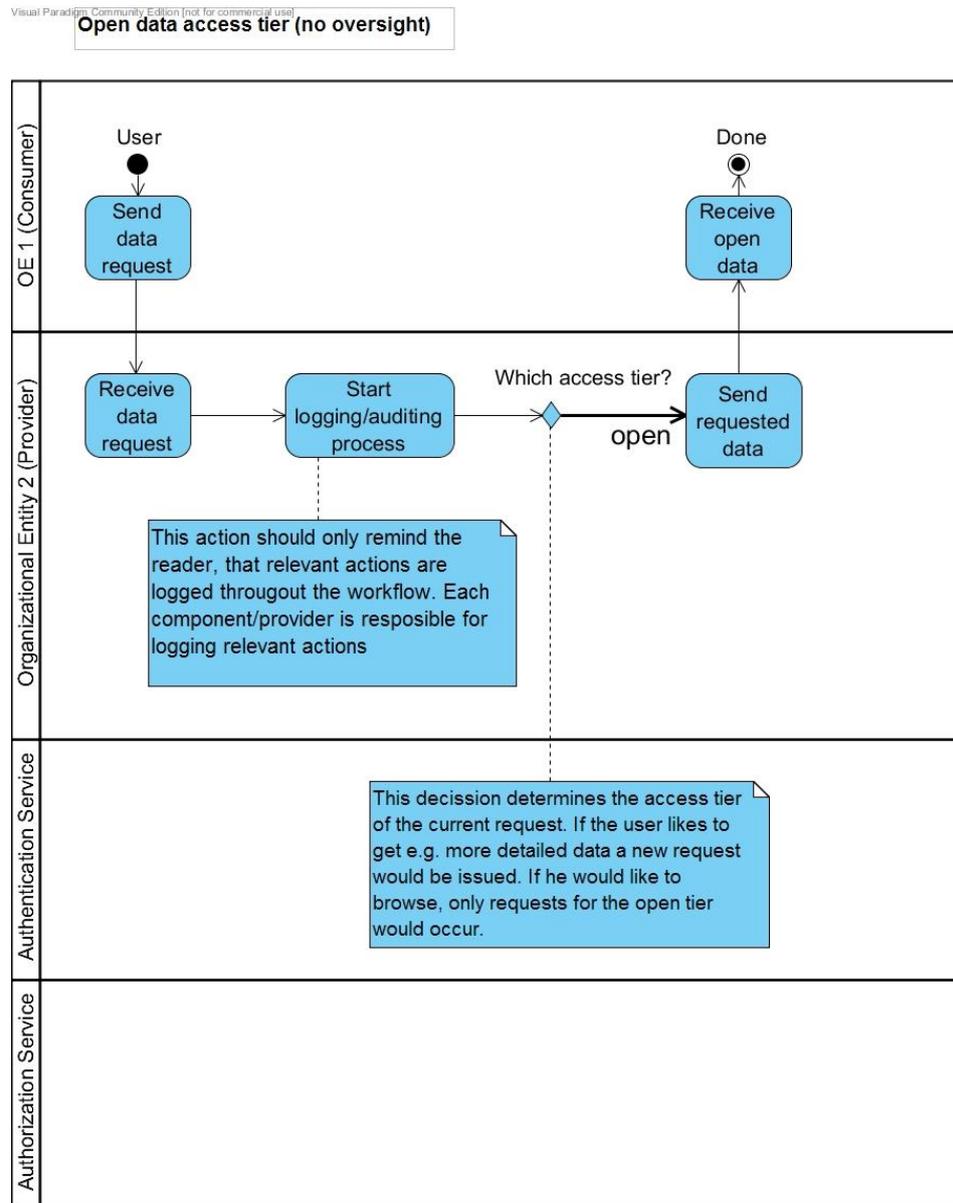
8.3.3 Committee-controlled data access tier

The committee-controlled data access tier shown in Figure 9 Table 9 is the most restrictive data access tier. It involves basically all the steps necessary for the restricted data access tier, and additionally requires the provision of necessary documents (e.g. research project information) from the consumer, committee approval, and the signing of contracts and agreements before the authorization of the consumer. We modelled these additional steps using a loop that allows for several iterations of committee reviews and requests revisions by the consumer until an approval by all involved parties is achieved.

In this data access tier, handling of release requirements, required information provided by the consumer, committee reviews, and the authorization of the consumer may be supported by an authorization service using an implementation such as REMS.



Figure 7. Activity diagram for the Open Data Access Tier.



Requirements:

1. ALL data (on all tiers) transmitted between participating OEs have to be sent over secure communication channels.

It is assumed, that standard web-technologie will be used, i.e. the provider has a SSL-Certificate (process of certificate acquisition omitted here).

This allows the Consumer to verify the identity of the Provider and to establish a secure connection. Most probably using HTTP over SSL/TLS protocol (browsing use case).

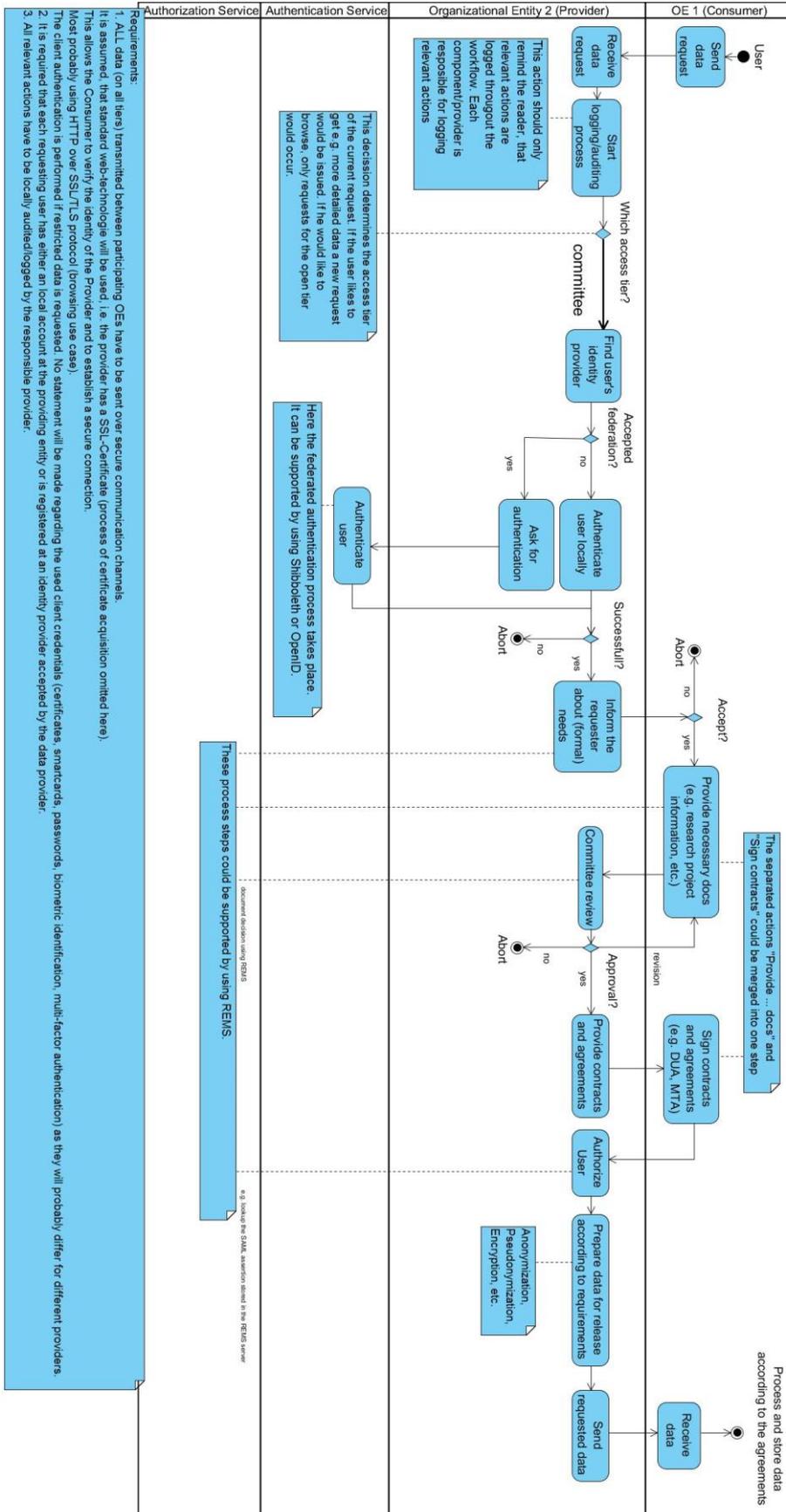
The client authentication is performed if restricted data is requested. No statement will be made regarding the used client credentials (certificates, smartcards, passwords, biometric identification, multi-factor authentication) as they will probably differ for different providers.

2. It is required that each requesting user has either an local account at the providing entity or is registered at an identity provider accepted by the data provider.

3. All relevant actions have to be locally audited/logged by the responsible provider.



Figure 9. Activity diagram for the Committee Controlled Data Access Tier





9 Steps from continuous feedback to implementation

9.1 Reviews and feedback on interim progress

A central element of the construction work package 5 has been its cooperation with other work packages, including the construction work packages and the use case work packages. The contacts to WP11 also established a connection beyond BioMedBridges to external partners, including the European Grid Infrastructure. Some of the most relevant specific activities are described here.

1. *Overview of data bridges and of security requirements of RIs and of use case work packages by means of a survey and of usage scenarios. Workshops to discuss the results.*

Feedback from RIs and use case work packages was sought early. To identify relevant data bridges and to identify security requirements, two surveys have been carried out which are described in this deliverable 5.3. The first survey was conducted in Aug 2012. Its results were of relevance to both deliverables 5.1 and to 5.3, and it has also been described in deliverable 5.1. Moreover, WP5 developed usage scenarios to improve the understanding of the bridges; they have been described in deliverable 5.1 and formed the basis for the development of legal “requirement clusters”. Survey 2 was carried out in September/October 2013, and it was primarily aimed at security focused questions. It resulted in data flow diagrams illustrating security threats. The results of both surveys were discussed in workshops. The workshop following survey 2 was held in Munich in Oct 2013, and it resulted in a first draft of the security architecture.

2. *Discussion of the structural elements of deliverable 5.3 on a Workshop during the GA in Florence.*

Workshop on March 11, 2014, Florence: During the Second Annual General Meeting of BioMedBridges, a Workshop on “Security, regulatory and ethical requirements for the data bridges” was conducted. Deliverable 5.3 is using



results of deliverable 5.1 and 5.2 and it is based on WT 5: Security requirements for an e-infrastructure addressing the use cases, WT6: Threat and risk analysis for sharing data or biomaterials, and particularly WT7: Design of the security architecture and framework. During the workshop, the Legal Assessment Tool (LAT) of deliverable 5.2 was presented and discussed. This was followed by a presentation and discussion of the structure of deliverable 5.3: Its relationship to deliverable 5.1 and 5.2, its basic definitions and concepts, the roles of WTs 5, 6, 7 for deliverable 5.3, all steps from requirements (identified by means of two surveys) data flow diagrams and threats (constructed from results of the second survey) , risks, countermeasures towards architectural elements. The results of the second survey were presented to the audience and discussed.

3. *Feedback from the implementation WT8 of WP5, and from EGA, BioSD, and WP10 during a face to face meeting on April 7 in Hinxton.*

Deliverable 5.3 lays the groundwork for WT8: “*Implementation of a pilot for the security framework*”. In order to make sure that the specification will be usable for the implementation WT8, the partners of WT8 (EMBL-EBI and TUM-MED) met on April 7 in Hinxton to discuss the current work on deliverable 5.3. The meeting was also very fruitful to get feedback from UC WP10 (lead by EMBL-EBI), EGA (with vast experience in data sharing) and BioSD (with whom a data bridge between BBMRI and ELIXIR has been realized in WP4 of BioMedBridges). The meeting also resulted in a concrete plan for a pilot installation.

4. *Distribution of material to WP5 and UC WPs on April 17, asking for feedback by May 10.*

Comments from the Florence workshop were included into a preliminary draft version of deliverable 5.3. This version was sent out to WP5 participants and the UC WPs on April 17, asking for feedback by May 10.

5. *WP5 Telco presenting and discussing work on WTs 5-8 and on D5.3 on May 14.*

On May 14, a WP5 Telco was held. The date was deliberately chosen to be after the feedback deadline of May 10 to have an additional opportunity to



discuss results. WT leaders reported on WTs 5, 6, 7, and 8, and these reports were discussed.

6. Telco with FIMM (Work Package 8) on May 16

During the WP5 Telco on May 14, it was noted that the data sharing process of WP8 can be seen as a highly relevant instance of the secure bridge specified by deliverable 5.3. It was decided to discuss open questions and details, including the DFD of WP8, during a separate Telco on May 16. Here, open question could be clarified, and the work on the security architecture received feedback from one of the important use case work packages. The resulting DFD is part of this deliverable.

7. Discussion of ELSI questions

During the Florence workshop, it was clarified that the security architecture of deliverable 5.3 will build on the previous work on legal and regulatory questions, and particularly on results of deliverable 5.2. Among the important results of deliverable 5.2 are templates, e.g. covering Data Use Agreements and Material Transfer Agreements. Questions going beyond the scope of the Florence workshop were directly asked to the lead beneficiary of deliverable 5.2, TMF.

8. Communication with the Technology Watch Work Package WP11

The security workshop in Florence was attended by members of WP11. In a follow-up activity, security questions were discussed during a Telco on April 28 with EGI as a member of WP11. After this Telco, a technical excerpt of the material distributed on April 17 was created and sent to EGI for further feedback. The same material was sent to further members of WP11 on May 19 (STFC, DANTE, CSC). All feedback was incorporated into this deliverable.

9.2 Implementation of the security architecture by use case work packages

As summarized in section 9.1, process elements and descriptions have been based on feedback from the research infrastructures and the use case work packages. Specifications in the form activity diagrams (deliverable 5.1) and



data flow diagrams (deliverable 5.3) were developed in direct interactions with representatives of the use case work packages. Domain specific (EGA) and domain independent (EGI) experience was sought for the definition of security specific process elements. As already pointed out, the European Genome-phenome Archive (EGA) was selected because of its important role and ample experience as an access portal.

Now, the technical elements described need to be implemented in the pilots of WT8. Pilots will be built in cooperation with WP4 and use case WPs. On the architectural level, deliverable 8.1 can be seen as a specific instantiation. Successful pilots will be critical for WP5. The architecture is designed in a way to allow advanced as well as pragmatic approaches. It is expected that Best Practices will result from WP5. As much will depend on the success of the pilots, close cooperation has been established with UC WPs and with WP4.

Also crucial for success will be quick and easy access to regulatory and legal information, and to templates of relevant forms. Currently, they can be accessed through 5.2, especially by the tool developed in D5.2. A relevant enhancement will be to integrate this tool directly into the workflows of researchers to provide immediate assistance (see section 8.2.8).

WT8 of WP5 “Implementation of a pilot for the security framework” has started in Month 24. The pilot will be based on the architecture presented in this deliverable 5.3; its results will lead to deliverable 5.4. This Implementation will need close collaboration with WP4 and WP3. Parallel to the implementation steps of the services provided by WP4, the relevant elements of the security framework will be implemented in a way oriented towards use case WPs.

A central enhancement will be the implementation of a module integrating the tool of deliverable 5.2 into the workflow of the pilot implementations. This module will allow quick and easy access for data providers and consumers to templates of relevant forms (DUA, MTA, etc..) and to legal and regulatory information relevant to their specific needs. This enhanced version of the tool realized for deliverable 5.2 will be built in a way which allows easy and seamless integration into the workflows of the data bridges. The already existing bridge between ELIXIR and BBMRI, connecting the BioSD database



and the BBMRI.eu catalogue can be extended towards a security filter providing access to tiered data. Previous work of WP4 has realized a REST-service based connectivity already, and it can be extended towards web service based “query” integration. Also for WP8, secure and layered access as well as secure data transfer can be demonstrated by a pilot.

10 Delivery and schedule

The delivery is delayed: Yes No

11 Adjustments made

No adjustments were made to the deliverable.



12 Appendices

12.1 Result Tables of Threat and Risk Analysis

This Appendix section contains the tables that report the detailed results of the risk assessment performed on the DFDs of the UC WPs that participated in WP5 Survey 2 (i.e. WP6, 7, 8, 10). The DFDs can be found in Section 6.2.3 of this Deliverable 5.3. The tables are based on the template for “adversarial risk” defined by Table I-5 in Appendix I of [4].

Table 18. Legend used by tables to report the threat and risk assessment results.

Threat Event	Threat Sources	Vulnerabilities and Predisposing Conditions	LoT: (Likelihood of Threat)	LoI: (Level of Impact)	Risk	Countermeasures (Elements of the Security Architecture)
A row in light blue background color indicates that the threat event is addressed whether by STRIDE or LINDDUN and applies only to the “ data flow ” element type of the DFD under evaluation.						
A row in light red background color indicates that the threat is addressed by STRIDE only and applies to the one of the rest element types of the DFD under evaluation (i.e. “ data store ”, “ process ”, “ external entity ”).						
A row in light green background color indicates that the threat is addressed by LINDDUN only and applies to the one of the rest element types of the DFD under evaluation (i.e. “ data store ”, “ process ”, “ external entity ”).						

12.1.1 Work package 6 threat and risk analysis results

Table 19. Threat and risk assessment results for use case WP6.

Threat Event	Threat Sources	Vulnerabilities and Predisposing Conditions	LoT	LoI	Risk	Countermeasures (Elements of the Security Architecture)
Spoofing: User(Researcher)	External	Weak Authentication system	M	M	M	Authentication system
Spoofing: HMGU user	External	Weak Authentication system	L	M	L	Authentication system
Spoofing:	Processin	Weak	L	M	L	Authentication



Pharming Phenotator Pharming Image Browsing	g	Authentication system/ Exploitation of DNS server				system/ Configuration management
Tampering: BMB Webmicroscope Database (Over capacity failure)	Storage	Missing handling of overcapacity failures	L	M	L	Server configuration
Tampering: Phenotator (Sql Injection, Input validation failure, Over capacity failure)	Processin g	Missing Input validation/ configuration management	M	M	M	Input validation practices/ Server configuration
Tampering: Data flow between user <-> Webmicroscope; user <-> Phenotator (MITM, Replay attacks)	Processin g	Insufficient secure connection	M	M	M	Secure data communication
Repudiation: Annotation changes in Phenotator	Processin g	Weak logging/ Missing audit trail	M	L	L	Auditing and logging
Repudiation: Data changes in the BMB web microscope not traced (Version of the data)	Storage	Weak logging/ Missing audit trail	M	M	M	Auditing and logging
Repudiation: Version of “ External databases” not logged (Mitocheck/Ensembl)	Processin g	Weak logging/ Missing audit trail	L	M	L	Auditing and logging
Information Disclosure: Images uploaded from HMGU mouse clinic	External	No secure transfer	L	L	L	Secure data communication
Information Disclosure: Annotations in Phenotator disclosed (insufficient access control)	External	Insufficient access control	M	M	M	Authorization
Denial of Service of BMB Web- Microscope (resources)	Storage	Insufficient Resources allocated	L	L	L	Configuration management



Denial of Service of Phenotator (missing input validation, resources)	External, Storage	Missing input validation/ not enough resources	L	L	L	Input validation practices/ Server configuration
Elevation of Privilege because of insufficient protection of Web-microscope	External	Insufficient access control	M	M	M	Authorization
Linkability: Annotation is linkable to researchers	External	The annotation has information about the researcher.	M	M	M	Anonymization
Identifiability: Patient can be identified in images based on accompanying meta data.	External	Images insufficiently anonymized/pseudonymized	M	H	H	*Data is stored in the pseudonymous form ⁵⁹ .
Content unawareness: Patient does not know for what his/her data is used.	Organizational	Patient was not informed well enough.	M	M	M	Standard procedure of operation
Policy and consent non-compliance: Consent does not cover images, sharing images and associated data ⁶⁰ .	Organizational	Insufficient consent	H	H	H	Consent Management

⁵⁹In the case of WP6, it is worth noting for the value of "Individual Level Data" that all data related to images have been pseudonymized and all patient identifiers have been replaced by internal codes. The patient identifiers are stored separately on the hospital side and cannot be accessed through the WebMicroscope system. For further details, see Del 5.1 Section 9: Usage Scenario 1: Imaging Bridge related to WP6 (p. 55).

⁶⁰ "Additional comments regarding the use of patient data(human tumor tissue data): Informed consent may not cover the envisioned type of research." Del 5.1 Section 9: Usage Scenario 1: Imaging Bridge related to WP6 (p. 56)



12.1.2 Work package 7 threat and risk analysis results

Table 20. Threat and risk assessment results for use case WP7.

Threat Event	Threat Sources	Vulnerabilities and Predisposing Conditions	L O T	L O I	R i s k	Countermeasures (Elements of the Security Architecture)
Spoofing: Database Browsing process (Pharming) Pathway analysis (Pharming)	External	Weak Authentication system/ Weak configuration management (System administration)	L	L	L	Authentication system/ Configuration management from administration view.
Tampering: Data flow between IMPC/ArrayExpress/Metab olights/ChEMBL and database browsing	Processing	Insecure data transfer	M	L	L	Secure data communication
Tampering: Data flow between Reactome and pathway analysis	Processing	Insecure data transfer	M	L	L	Secure data communication
Tampering: Data flow between Infrafrontier database and database browsing / Pathway browsing	Processing	Insecure data transfer	L	L	L	Secure data communication
Tampering: Data flow between user <-> Database browsing; user <-> Pathway analysis	External	Missing dummy traffic/ no encryption	L	L	L	Secure data communication
Tampering: Infrafrontier database (Overcapacity failure)	Storage	Missing Input validation / Missing handling of overcapacity failures	M	M	M	Input validation practices / Configuration management
Tampering (MITM, Replay attacks): Data flow between User <-> Database browsing; User <-> Pathway analysis	Processing	Insecure data transfer	M	M	M	Secure data communication
Tampering: Database browsing/ Pathway analysis process	Processing	Missing input validation failure	H	H	H	Input validation practices
Repudiation: Data upload from Univ of Graz / CERM to ArrayExpress/ChEMBL not logged (Version of the data)	Processing	Weak logging/ Missing audit trail	L	L	L	Auditing and logging



Repudiation: Data changes in the Infrafrontier database not traced (Version of the data)	Storage	Weak logging/ Missing audit trail	L	L	L	Auditing and logging
Information Disclosure: Dataflow between user <-> Database browsing; user <-> Pathway analysis	Processing	Insecure data transfer	M	M	M	Secure data communication
Denial of Service of Database browsing/Pathway analysis (Not enough resources)	Processing/ Storage	Lack of resources (Processing or storage)	L	M	L	Configuration management
Denial of Service of Database browsing/Pathway analysis (Input validation failure)	Processing	Missing input validation	L	M	L	Input validation practices
Denial of Service: Dataflow between user <-> Database browsing; user <-> Pathway analysis (corrupt message/preplay)	Processing	Insecure data transfer	L	L	L	Secure data communication
Identifiability: Identifiability of a patient based on his/her gene expression data.	External	Insufficiently anonymized/pseudonymized	L	L	L	Anonymization/ Pseudonymization*
Information disclosure: Attribute disclosure of a patient based on his/her SNP data.	External	Insufficiently anonymized/pseudonymized	L	H	M	Anonymization/ Pseudonymization*
Content unawareness: Patient does not know that his data can be published in ArrayExpress.	Organizational	Patient not well informed	M	M	M	Standard procedure of operation
Content unawareness: Patient does not know that his data can be published in Metabolights.	Organizational	Patient not well informed	M	M	M	Standard procedure of operation
Policy and consent non-compliance: Insufficient consent for publishing data in ArrayExpress/Metabolights.	Organizational	Insufficient consent	M	M	M	Consent management

12.1.3 Work package 8 threat and risk analysis results

Table 21. Threat and risk assessment results for use case WP8.

Threat Event	Threat Sources	Vulnerabilities and Predisposing Conditions	L O T	L O I	R i s k	Countermeasures (Elements of the Security Architecture)
Spooing as user to get access to the PM database	External	Weak authentication	L	H	M	Authentication system



Spoofting: Pharming 'PM Analysis tool'	Processing	Weak Authentication system/ Weak configuration management	L	H	M	Authentication system/ Configuration management
Tampering: Modify/ delete data in the PM database	Internal	Weak access control of Molgenis database (accidental)	M	M	M	Authorization
Tampering: Modify data flow from external entity	External	Insecure data transfer	L	L	L	Secure data communication
Tampering of analysis tool (input validation failure)	Processing	Missing Input validation	M	H	H	Input validation practices/ Configuration management
Tampering: Overcapacity failure of PM database,	Storage	Missing handling of overcapacity failures	L	M	L	Configuration Management
Repudiation: Changes in PM database cannot be traced	Processing	No/weak audit trail	L	M	L	Auditing and logging
Repudiation: PM analysis tool activities (e.g. connection to databases) cannot be traced	Processing	No/weak audit trail	L	L	L	Auditing and logging
Repudiation: Version of external data sources used in the analysis not logged	Processing	Weak logging/ Missing audit trail	M	M	M	Auditing and logging
Information disclosure of patient data in PM application	Internal	PM application is not secure, weak access control	H	H	H	Authorization
Information disclosure of query / query results (data flow, process)	External	Insecure data transfer	M	M	M	Secure data communication
Information disclosure of annotated patient data (data flow, process, data store)	External	Insecure data transfer / insufficient access control of data store	M	H	H	Secure data communication/ Authorization
Denial of Service of PM database (input validation, lack of resources)	Processing, Storage	Missing input validation/ insufficient resources handling	L	L	L	Input validation practices/ Configuration management



Denial of Service of PM analysis tool (input validation, lack of resources)	Processing, Storage	Missing input validation/ insufficient resources handling	L	M	L	Input validation practices/ Configuration management
Elevation of Privilege: A user has access to patient data that s/he is not allowed to.	Processing, Storage	Insufficient access control	M	H	H	Authorization
Elevation of Privilege: Unauthorized users can edit/delete patient data.	Processing, Storage	Insufficient access control	H	H	H	Authorization
Linkability of query results for particular patient	External	Query results are in some way connected	L	H	M	Secure data communication or Encryption query result
Linkability of entry in "Pseudonymized/anonymized datastore" to patient	External	Insufficient anonymization/ Pseudonymization	L	M	L	Anonymization/ Pseudonymization
Identifiability of patient with the help of queries/ query results	External	Queries/ results not protected.	L	H	M	Secure data communication
Identifiability of the patient based on visit pattern	External	Insufficient anonymization	L	L	L	Anonymization/ Pseudonymization
Identifiability of patient based on diagnosis codes	External	Insufficient anonymization	M	H	H	Anonymization/ Pseudonymization
Identifiability of patient based on genomic data(Gene expression, SNPs , DNA sequence data)	External	Insufficient anonymization	M	H	H	Anonymization/ Pseudonymization
Identifiability of researcher using analysis tool	External	Researcher is logged	M	M	M	Anonymization/ Authorization
Identifiability: A patient's presence in the PM database can be discovered without having access to	External	Insufficient data protection at PM application.	M	H	H	Anonymization/ Authorization



patient's data. (Membership disclosure)						
Identifiability: Discovering the use of a patient's data in e.g. a published study.	External	Insufficient anonymization	M	M	M	Anonymization
Non-repudiation of patient data in database	External	Data can be somehow linked to patient	L	L	L	Anonymization/ Pseudonymization
Information disclosure: Attribute disclosure -> inferring phenotype from genotype.	External	Insufficient anonymization	M	M	M	Data protection techniques may not exist / Data to be shared with trusted party (Data Access committee)
Content Unawareness: Patient does not know what data s/he provides and how it is processed.	Organizational	Patient is not informed well enough	M	H	H	Standard procedure of operation
Policy and consent non-compliance: Insufficient consent; must cover storing the data in the PM database, annotating/processing the data, research and publishing results.	Organizational	Insufficient consent*	H	H	H	Consent management

12.1.4 Work package 10 threat and risk analysis results

Table 22. Threat and risk assessment results for use case WP10. The REMS is explicitly mentioned here. It is suggested to use it for all UC WPs.

Threat Event	Threat Sources	Vulnerabilities and Predisposing Conditions	L O T	L O I	R i s k	Countermeasures (Elements of the Security Architecture)
Spoofing: Spoofing user	External	Weak authentication	M	M	M	Authentication system
Tampering: Access rights for users are tampered.	External	Missing encryption/ missing access control	L	H	M	Authorization system



Tampering: DAC service (overcapacity).	Storage	Missing handling of overcapacity failures	L	M	L	Configuration management
Tampering: Data flow from DAC service to REMS data store tampered. (MITM, replay)	External	Insecure data transfer/ no encryption	L	L	L	Secure data communication
Tampering: Data flow from DAC service to biobanks tampered. (MITM, replay)	External	Insecure data transfer/ no encryption	L	L	L	Secure data communication
Tampering: Data flow from DAC service to user tampered. (MITM, replay)	External	Insecure data transfer/ no encryption	L	L	L	Secure data communication
Tampering: Data flow from REMS data store to Biobank tampered. (MITM, replay)	External	Insecure data transfer/ no encryption	L	L	L	Secure data communication
Tampering: Data flow from User to DAC service. (MITM, replay)	External	Insecure data transfer/ no encryption	L	L	L	Secure data communication
Tampering: Data flow from DAC service <-> DAC approver. (MITM, replay)	External	Insecure data transfer/ no encryption	L	L	L	Secure data communication
Repudiation: Application procedure from user is not logged (involves Data flow from User to DAC service, DAC service, Data flow from DAC service to DAC approver and back, Data Flow from DAC service to REMS data store, biobank and user).	Processing	No/weak audit trail	L	M	L	Auditing and logging
Repudiation: Access from biobanks to REMS data store is not logged.	Processing	No/weak audit trail	M	L	L	Auditing and logging
Information Disclosure: REMS data store (insufficient access control).	External	Insufficient access control/ missing encryption	H	H	H	Authorization



Information Disclosure: Data flow from User to DAC service and vice versa (MITM etc.).	External	Insecure data transfer/ no encryption	M	M	M	Secure data communication
Information Disclosure: Data flow from DAC service to DAC approver and vice versa (MITM etc.).	External	Insecure data transfer/ no encryption	M	M	M	Secure data communication
Information Disclosure: Data flow from DAC service to biobanks. (MITM etc.)	External	Insecure data transfer/ no encryption	M	M	M	Secure data communication
Information Disclosure: DAC service (input validation failure).	External	Missing input validation	M	H	M	Input validation practices
Denial of Service: DAC service (resources, input validation).	Processing	Missing input validation/ insufficient resources management	L	M	L	Input validation practices/ Configuration management
Denial of Service: REMS data store (overcapacity failure, input validation failure, etc.).	Processing	Missing handling of overcapacity failures	M	H	H	Configuration management
Denial of Service: Data flow User to DAC service (preplay, corrupt message).	External	Insecure data transfer/ no encryption	L	M	L	Secure data communication
Denial of Service: Data flow DAC service to DAC approver (preplay, corrupt message).	External	Insecure data transfer/ no encryption	L	M	L	Secure data communication
Denial of Service: Data flow DAC service to biobank, REMS data store, user (preplay, corrupt message).	External	Insecure data transfer/ no encryption	L	M	L	Secure data communication
Elevation of Privilege: Via tampering of data store or input validation failure at DAC service.	External	Missing input validation, weak access control	M	H	H	Input validation practices/ Authorization



13 Background information

This deliverable relates to WP5; background information on this WP as originally indicated in the description of work (DoW) is included below.

WP5 Title: Secure access
 Lead: Heinrich-Heine-Universitaet Duesseldorf - 5: UDUS
 Participants: EMBL, STFC, UDUS, TUM-MED, ErasmusMC, TMF, HMGU, INSERM

Work package number	WP5	Start date or starting event:			month 1				
Work package title	Secure access								
Activity Type	RTD								
Participant number	1:EMBL	4:STFC	5:UDUS	6:FVB	7:TUM-MED	9:ErasmusMC	10:TMF	11:HMGU	14:INSERM
Person-months per participant	61	15	54	0	58	5	34	10	4

Objectives

Based on an analysis of the complex ethical, legal and regulatory issues resulting from international data and biomaterial sharing between different e-Infrastructures WP5 will develop a security framework that will ensure that services provided by BioMedBridges are compliant with local, national and European regulations and privacy rules. Therefore the developed legal framework will allow the use of data bridges, that consider among other regulations the EU Directive 95/46/EC, EU Directive 2001/20/EC (GCP), national data protection acts, GLP rules, animal protection laws, laws about biobanking, laws concerning genetic data and stem cell research, data access approval rules (by informed consent), rules by Hospital Boards or Research / Ethics Committees as well as regulations for intellectual property and licence rights.

The legal foundation will be applied for the development of a security framework employing security policies, account policies, consent, user agreements of the participating infrastructures and authentication and authorization services. Existing standards and concepts of European e-



infrastructures (e.g. GÉANT / eduGAIN and TERENA) will be considered.

Description of work and role of participants

In WT 1-4 regulations, requirements and design aspects; in WT 5-8 the security implementation are addressed.

In the first part, information collection will require extensive contacting and considerable travelling. In the second part, staff exchange will be an important way to coordinate activities. WT5 will be chaired by UDUS and TUM.

WT 1: Regulations and privacy requirements for using the data bridges (M1-M12)

(Leader: UDUS, Participants: EMBL-EBI, Erasmus MC, HMGU, STFC, TMF, TUM, FVB, INSERM)

This task will analyse the legal and ethical situation concerning the sharing and transfer of data and the access to data in a trans-European context for all e-Infrastructures. The legal implications and corresponding data exchange strategies will be analysed on European, national, regional (e.g. data protection law in Scotland) and local (e.g. hospital law) level. Legal implications for different types of data and the linking of data have to be considered, including biobank data, genetic data, stem cell research data, data of children and vulnerable will be paid to personal data (Directive 95/46/EC) and the roles of data controller and data processor for the data bridges. Subcontracting will be needed for lawyer support and translation of legal documents.

WT 2: Rules and regulations for accessing databases of e-Infrastructures (M6-M18)

(Leader: UDUS, Participants: EMBL-EBI, Erasmus MC, HMGU, STFC, TMF, TUM, FVB, INSERM)

This task will analyse the rules, regulations and associated practices and policies concerning the access to e-Infrastructure databases. A survey will analyse the situation and policies of all e-Infrastructure databases.

Special attention will be paid to the role of different types of informed consent, research exemptions, policies, and approvals by Hospital Biobanks Boards or Research and Ethics Committees.

WT 3.1: Regulations and security issues regarding security of biosamples (M1-M12)

(Leader: TUM, Participants: EMBL-EBI, ErasmusMC, UDUS, HMGU, STFC, TMF, FVB, INSERM)

This task will analyse the rules and regulations that affect data protection and security of bio samples. Especially the physical transfer of samples may be restricted by national legislations.



WT 3.2: Regulations and security issues regarding animal protection (M1-M12)

(Leader: TMF, Participants: EMBL-EBI, Erasmus MC, UDUS, HMGU, TUM, FVB, INSERM)

This task will analyse the rules, practices and regulations concerning data protection and the protection of animal welfare.

WT 3.3: Rules and regulations regarding data connected to intellectual property and licences in e-Infrastructures (M1-M12) (Leader: EMBL-EBI, Participants: Erasmus MC, UDUS, HMGU, STFC, TMF, TUM, FVB, INSERM)

This task will analyse the rules, practices and regulations concerning the access to databases and the sharing of data protected by intellectual property rights.

WT 4: Development of a tool for assessment of ethical and legal requirements and supporting documents (M13-24) (Leader: TMF, Participants: EMBL-EBI, Erasmus MC, UDUS, HMGU, STFC, TUM, FVB, INSERM)

In this WT all results of the previous WTs will be collected, integrated and interdependencies will be developed. The different dimensions of the developed requirements matrix will cover: (1) kind of data (patient data, molecular data, mouse data, phenotype data, etc.), kind of data protection (anonymisation, pseudonymisation, none), regulations and rules for secure access. A priority list of combinations of these dimensions that may happen during cooperation between different e-Infrastructures will be analysed and depicted. In addition, contractual templates and generic texts will be developed to support a legal sound cooperation for data exchange.

WT 5: Security requirements for an e-infrastructure addressing the use cases (M6-30). (Leader: TUM, Participants: EMBL-EBI, Erasmus MC, UDUS, HMGU, STFC, TMF, FVB, INSERM)

Utilizing results from the previous WTs and focussing on a priority list of use cases including WP8, WP7 and WP10, security requirements for aggregated or shared data or biomaterials will be identified, including confidentiality, integrity, and availability. These requirements will consider the different levels of integration (WP4), type and content of integrated data (including the specific risk of re-identification) or shared biomaterials, security policies and consent agreements of the participating infrastructures and European regulations. The use of de-identification and (k-) anonymity will be specified.

Requirements for data access layers will be defined. Suggested tiers are: (1) Public access to meta and coarse grained data, where typical risks need to be considered (e.g. statistical inference of membership); (2) access to k-anonymous derived or summary data based on use agreements and user accounts, (3) access to de-identified microdata integrated / accessible across infrastructures which requires approval of a data access committee. Consent agreements and security policies of the participating infrastructures will be considered in these tiers.



WT 6: Threat and risk analysis for sharing data or biomaterials (M9-30)
(Leader: TUM, Participants: EMBL-EBI, Erasmus MC, UDUS, HMGU, STFC, TMF, FVB, INSERM)

Based on the security requirements, a threat and risk analysis will be performed. Attacker models, origins of threats (e.g. trails), and possible points of attack will be identified, considering results from latest research. Following typical (risk) categories need to be considered: Membership disclosure, attribute disclosure and re-identification. The risk analysis will weigh the different threats, considering the interests of researchers, protection of research-related IP, and privacy of patients.

WT 7: Design of the security architecture and framework (M18-30)
(Leader: EMBL-EBI, Participants: TUM)

Derived from the requirements developed in previous WTs, a security framework will be designed, comprising authentication, authorization, and accounting services. Different security solutions will be evaluated, ranging from decentralized to tightly integrated authentication and authorisation. Access layers and corresponding approval workflows will be specified. Authentication mechanisms for the integrated databases need to be designed, using standards (e.g. OpenID, Shibboleth, Liberty Alliance) and utilizing concepts or solutions from European identity federation initiatives (GÉANT and TERENA). The security policies of BioMedBridges will comprise access policies and use agreements and will consider security policies of participating infrastructures and European laws and regulations (derived from WT 4). The security framework needs access to a repository of authorization rules as part of a metadata repository. These authorization rules will be based on consent and regulations of the participating infrastructures combined with rules and contracts for co-operation. Authorization policies have to be expressed in an appropriate format (e.g. XACML). The policy administration repository will be related to defined access tiers. Logging of user activities is used to ensure accountability.

WT 8: Implementation of a pilot for the security framework (M24-48)
(Leader: EMBL-EBI, Participants: TUM, UDUS, STFC, TMF)

Implementation will need close collaboration with WP4 and WP3. Parallel to the implementation steps of the services provided by WP4, and for the same use cases, the security framework developed in this WP will be implemented. The policy administration repository will be a central part of this implementation.

Subcontracting for legal costs: UDUS (partner 5) for legal costs associated with WP5 - Work Task 1 of WP5 will analyse the legal and ethical situation concerning the sharing and transfer of data and the access to data in a trans-European context for all e-Infrastructures. Subcontracting is required for legal advice, and the translation of legal documents.



Deliverables		
No.	Name	Due month
D5.1	Report on regulations, privacy and security requirements	18
D5.2	Tool for assessment of regulatory and ethical requirements, including supportive documents	24
D5.3	Report describing the security architecture and framework	30
D5.4	Implementation of a pilot for the security framework	48



14 References

- [1] BioMedBridges, „Deliverable 5.1 - Report on regulations, privacy and security requirements,“ <http://www.biomedbridges.eu/deliverables/51-0>, June 2013.
- [2] BioMedBridges, „Deliverable 5.2 - Building data bridges between biological and medical infrastructures in Europe,“ <http://www.biomedbridges.eu/deliverables/52-0>, December 2013.
- [3] M. Bishop, *Computer Security: Art and Science*, Addison-Wesley, 2003.
- [4] „NIST Special Publication 800-30, Guide for Conducting Risk Assessments,“ September 2012.
- [5] R. Shirey, „RFC-4949. Internet Security Glossary, Version 2,“ August 2007.
- [6] „ISO 27799. Health informatics — Information security management in health using ISO/IEC 27002. First ed.,“ July 2008.
- [7] M. Deng, K. Wuyts, R. Scandariato, B. Preneel und W. Joosen, „A privacy threat analysis framework: supporting the elicitation and fulfillment of privacy requirements,“ *Requirements Engineering*, Bd. 16, Nr. 1, pp. 3-32, March 2011.
- [8] M. Howard und S. Lipner, *The security development lifecycle: SDL, a process for developing demonstrably more secure software*, Microsoft Press, 2006.
- [9] „ISO 27000. Information technology - Security techniques - Information security management systems - Overview and vocabulary. First ed.,“ May 2009.
- [10] V. Curcin, S. Miles, R. Danger, Y. Chen, R. Bache und A. Taweel, „Implementing interoperable provenance in biomedical research,“ *Preprint submitted to Future Generation Computer Systems*, 9 December 2013.
- [11] A. Pfitzmann, and M. Hansen, „A terminology for talking about privacy by data minimization: Anonymity, unlinkability, undetectability, unobservability, pseudonymity, and identity management,“ (Version 0.33), tech. rep., TU Dresden and ULD Kiel, April 2010.
- [12] European Parliament and the Council, *Directive 95/46/EC*, 1995.
- [13] B. Riedl, V. Grascher, S. Fenz und T. Neubauer, „Pseudonymization for improving the Privacy in e-Health Applications,“ in *Proceedings of the 41st Hawaii International Conference on System Sciences*, Hawaii, 2008.
- [14] Erika McCallister, Tim Grance, and Karen Scarfone, „NIST Special Publication 800-122. Guide to Protecting the Confidentiality of Personally Identifiable Information (PII),“ April 2010.
- [15] M. Roe, „Cryptography and evidence,“ PhD thesis, University of Cambridge, 1997.
- [16] BioMedBridges, „Description of Work (DoW),“ 2012.
- [17] W. P. Stevens, G. J. Myers und L. Constantine, „Structured design,“ *IBM Systems Journal, Volume 13, Issue 2*, pp. 115-139, <http://dx.doi.org/10.1147/sj.132.0115>, 1974.
- [18] J. Donald S. Le Vie, *Understanding Data Flow Diagrams*, last access April 2014.
- [19] B. Malin, D. Karp und R. H. Scheuermann, „Technical and policy approaches to balancing patient privacy and data sharing in clinical and translational research,“ *J Investig Med*, January 2010, 58(1):11-8.
- [20] M. M. Mello, J. K. Francer, M. Wilenzick, P. Teden, B. E. Bierer und M. Barnes, „Preparing for Responsible Sharing of Clinical Trial Data,“ *N Engl J Med*, 24th October 2013, 369(17):1651-8.
- [21] Federal Financial Institutions Examination Council, „Authentication in an Internet Banking Environment,“ http://www.ffiec.gov/pdf/authentication_guidance.pdf, 2008.
- [22] Security Assertion Markup Language (SAML) TC https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=security.



- [23] M. Linden, T. Nyrönen und I. Lappalainen, Federated authorisation: Resource Entitlement Management System, TERENA Networking Conference (TNC), 2013.
- [24] RFC 5246: The Transport Layer Security (TLS) Protocol Version 1.2, August 2008.
- [25] K. Schmeh, Kryptografie, dpunkt.verlag, 2009.
- [26] NIST, Announcing the ADVANCED ENCRYPTION STANDARD (AES), <http://csrc.nist.gov/publications/fips/fips197/fips-197.pdf>, November 26, 2001.
- [27] ISO, „Health informatics - Pseudonymization (ISO/TS 25237),“ Intec, p. ISO/TS 25237:2008(E).
- [28] N. Li, T. Li und S. Venkatasubramanian, „t-Closeness: Privacy Beyond k-Anonymity and l-Diversity,“ in *IEEE 23rd International Conference on Data Engineering*, Istanbul, 15-20 April 2007.
- [29] T. Li, N. Li, J. Zhang und I. Molloy, Slicing: A New Approach for Privacy Preserving Data Publishing, March 2012.
- [30] M. E. Nergiz, M. Atzori und C. Clifton, „Hiding the presence of individuals from shared databases,“ in *SIGMOD '07 Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, New York, NY, USA, 2007.
- [31] L. Sweeney, „k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY,“ *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, Bd. 10, Nr. 5, October 2002.
- [32] A. Machanavajjhala, J. Gehrke, D. Kifer und M. Venkatasubramanian, „l-diversity: Privacy beyond k-anonymity,“ *ACM Transactions on Knowledge Discovery from Data (TKDD)*, Bd. 1, Nr. 1, March 2007 .
- [33] N. R. Adam und J. C. Worthmann, „Security-control methods for statistical databases: a comparative study,“ *ACM Computing Surveys (CSUR)*, Bd. 21, Nr. 4, pp. 515-556, December 1989.
- [34] D. E. Denning, P. J. Denning und M. D. Schwartz, „The tracker: a threat to statistical database security,“ *ACM Transactions on Database Systems (TODS)*, Bd. 4, Nr. 1, pp. 76-96, March 1979.
- [35] B. Zhou, J. Pei und W.-S. Luk, „A Brief Survey on Anonymization Techniques for Privacy,“ *ACM SIGKDD Explorations Newsletter*, Bd. 10, Nr. 02, pp. 12-22, December 2008.
- [36] L. Moreau, B. Clifford, J. Freire, J. Futrelle, Y. Gil, P. Groth, N. Kwasnikowska, S. Miles, P. Missier, J. Myers, B. Plale, Y. Simmhan, E. Stephan und J. V. d. Bussche, „The Open Provenance Model core specification (v1.1),“ *Future Generation Computer Systems*, Bd. 27, Nr. 6, pp. 743-756, June 2011.