Deliverable 5.4

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:	Implementation of a pilot for the security framework	
WP No.	5	
Lead Beneficiary:	1: EMBL	
WP Title	Secure access	
Contractual delivery date:	31 December 2015	
Actual delivery date:	17 December 2015	
WP leader:	5: Christian Ohmann 7: Klaus Kuhn	5: UDUS 7: TUM-MED
Contributing partner(s):	1: EMBL, 5: UDUS, 7: TUM-MED, 18: CSC	

Authors and contributors: Marco Brandizi, Olga Melnichuk, Ugis Sarkans, Raffael Bild, Florian Kohlmayer, Benedicto Rodriguez-Castro, Helmut Spengler, Klaus Kuhn, Wolfgang Kuchinke, Christian Ohmann, Timo Mustonen, Mikael Linden, Tommi Nyrönen

Contents

Figures

1 Executive summary

Managing authorised access to data is crucial in medical research and especially in translational medicine, since in this field open-access information is mixed with restricted-access data sets. In general, all health data has to be regarded as sensitive, subject to special protection. Thus, a sound approach to this problem is particularly critical when we deal with clinical data and patient's personal information.

BioMedBridges addresses these issues in WP5, which collected requirements for data protection and information security from all participating research infrastructures and documented those in D5.1 and D5.2. Report D5.3 applies these data protection requirements when considering theoretical and methodological aspects of secure information exchange, describing the most common security and privacy threat types according to the well-established STRIDE and LINDDUN approaches. Moreover, D5.3 outlines the practicalities of implementing infrastructures and workflows to countermeasure such threats.

Here we describe the pilot infrastructure that we have implemented, where different software applications are coordinated to allow end users appropriate access to restricted-access biomedical data. In such an infrastructure, the user initiates data search in public resources where only general data set descriptors are available. As the next step, the initial search results can be expanded by following cross-links to other, protected resources. Such restricted access is mediated by the integration of well-known identity management software, as well as tools to manage access policies. As we show in the following, our pilot makes use of a well-established software solution, which is already in use in the biomedical field. This minimizes the disruption of the existing IT infrastructures, e.g. by removing the need to create new user accounts and credential management tools. Our solution is modular and makes it possible to integrate components other than the ones we have considered so far. An example of this is shown in section 7.

This work has been done by organisations that have significant experience with the management and exchange of biomedical data, either for clinical studies or other research purposes. The Technische Universität München (TUM) has experience in web applications for managing clinical data and biobanks. In particular, they participate in the BBMRI-ERIC network. The CSC-IT has expertise in developing the Finnish state-owned IT infrastructure, including solutions for IT security. The European Informatics institute (EMBL-EBI) has been providing freely available data for the worldwide research community for more than 20 years, and has developed the BioSamples database as a hub of information on biological samples used in life sciences.

2 Project objectives

With this deliverable, the project has reached or the deliverable has contributed to the following objectives[1]:

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	

3 Background: software components reused

Here we describe the software components and user roles that were involved in the implementation of the secure access pilot.

3.1 EBI Biosamples Database (BioSD)

The European Bioinformatics Institute's Biosamples Database (BioSD [1]) is a public repository focused on biological sample information. Its rationale is to provide a single-point access to uniform annotations about the bio-materials used in biological and/or medical research. Such information can then be linked to assay-specific and/or technology-specific data in other EMBL-EBI databases and external repositories. This way, one can search for biomedical samples of interest, using a single interface and common criteria, and navigate from such search results to external heterogeneous data of other types. This also allows

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

data providers to unify the management of sample information, thus avoiding data duplication and maintenance efforts. For instance, one can search for samples in BioSD having certain phenotypical characteristics, and then navigate from a given result to ArrayExpress, where microarray data, derived from that sample, are available, or to ENA, where sequencing data on the same sample are stored. Summaries about clinical trials and other medical samples are a significant subset of BioSD. As such, BioSD constitutes a tool to aid translational research, since, for instance, one can perform a search for experiments targeting a given disease and obtain information about both clinical research, and basic research on model organisms. BioSD stores only general metadata about clinical information, delegating the access to sensitive details (i.e., anonymised patient records) to external resources.

3.2 BBMRI Hub and biobanks

BBMRI (Biobanks and Biomolecular Resources Research Infrastructure)[2] is a European research infrastructure which had a Preparatory Phase from 2008 to 2011 and is now active in the form of BBMRI-ERIC. It is closely related to the project BBMRI-LPC (BBMRI - Large Prospective Cohorts)[3] which aims to build a network for the large European prospective studies in order to facilitate transnational research about human health and diseases. Using results from the BBMRI Preparatory Phase, a web-based application called “LPC Catalogue”[4] has been developed in the context of BBMRI-LPC which provides a structured overview over the participating cohorts and supports researchers in gaining access to their biomaterials. It comprises a MIABIS-Layer, with metadata establishing compatibility to the standard MIABIS[5] (Minimum Information About Biobank data Sharing), as well as a data cube [2] in order to support more detailed search requests.

[2] http://bbmri-eric.eu/

[3] http://www.bbmri-lpc.org/

[4] http://www.bbmri-lpc-biobanks.eu/catalogue.html

[5] https://github.com/MIABIS/miabis/wiki

For the purpose of the pilot, an adapted instance of the LPC Catalogue called "BBMRI Hub"[6] has been set up by the Technische Universität München (TUM). This new instance provides enhanced functionalities for supporting access to microdata (such as data about individual human samples stored by external biobanks) in a secure and privacy preserving manner. To this end, the Shibboleth system for identity management and the software REMS (see below) have been integrated, and the functionality to detect REMS attributes (see below) in the Shibboleth user session have been implemented. Based on these attributes, user entitlements for accessing requested resources are being verified. Furthermore, the BioMedBridges Legal & Ethical Assessment Tool[7] has been integrated in order to provide additional support regarding ethical and legal questions.

3.3 Resource Entitlement Management System (REMS)

The Resource Entitlement Management System (REMS)[8] is an open source software, developed by the company CSC - IT Center for Science (CSC)[9], which can be used to manage the policies for granting access to resources, including digital data. For example, the data manager may establish that to get access to clinical data an application procedure is required, where certain forms need to be filed to a Data Access Committee (DAC). The application may include applicant's personal details, a research plan or other declarations about the intended use of the data, and the confirmation that documents regarding terms of service, non disclosure agreements and other policies have been read and approved. REMS allows data managers to define, on a per-resource basis, which kind of application must be filled in for getting access to the resource. The application is used to interact with the applicant, and, when (s)he has completed the application, it is sent to a DAC user, who can review the request and grant access to the target resource. REMS digitises, centralises and simplifies procedures that are often dispersed in the bureaucracy of one or

[6] https://shibboleth.imse.med.tum.de

[7] http://www.biomedbridges.eu/sharing-sensitive-data

[8] https://confluence.csc.fi/display/REMS/Home

[9] https://www.csc.fi

more organisations. Moreover, it eases the storage and management of user information and the access rights that have been granted to them. REMS has been successfully used to manage the access to the data in the EMBL-EBI's European Genome-Phenome Archive (EGA, [3]), which provides human genetic data from various sources, in a controlled way. These include consent given by patients only for specific uses, and pre-authorisation of access to the data. REMS can be integrated with Shibboleth (see the next section), both for the delegation of user authentication, and for the distribution of the entitlement attributes used for indicating the DAC's decision to approve the application. The entitlement attributes allow a web application to detect whether the currently authenticated user (as validated by an external IdP that REMS supports) has access to a given digital resource (such as a dataset in a biobank).

3.4 Legal Assessment Tool (LAT)

The LAT tool[10] aims to raise awareness of formal requirements when sharing data with Ethical, Legal and Societal Implications (ELSI). It highlights areas that need further action from the researcher when making the data available (the "data provider"), or issues alerts when further expert advice may be needed. The tool covers the current legal framework in the European Union concerning four areas: data protection, data security, intellectual property and biosample security. General requirements are provided, with hints and solutions such as templates for consent or data sharing agreements. The tool does not provide legal advice, it delivers legal requirements and recommandations. Users are guided through a series of multiple choice questions where they are asked, for example, about the type of data they want to share (metadata, text data, images, genetic data, biosamples or biosample associated data), the form in which the data is provided (data from which individuals can be identified or pseudonymised/anonymised data), or possible limitations on the wider use of the data (e.g. intellectual property requirements). After providing the necessary information, the user is shown the applicable rules and regulations for their

[10] http://www.biomedbridges.eu/sharing-sensitive-data

specific case and the tool recommends possible solutions or necessary further steps to make their data shareable.

LAT was designed to provide researchers with the basic requirements for their data access and sharing needs, in an understandable and non-expert way. In this context, LAT complements other activities in the field of legal requirements for the protection of health data, like The International Policy interoperability and data Access Clearinghouse (IPAC), the BBMRI Legal WIKI, hSERN and the TREAT-NMD / ECRIN Regulatory Affairs Database. All these sources provide the user with general legal information, links to the relevant acts and ordinances, and also provide templates for documents like data sharing agreements. In contrast, our approach is to consider regulations, but to focus on data access and usage rules employed by various data providers, and treat them as requirements for data sharing processes. Thus, LAT is well suited to support, on a high level, data provision or data sharing steps that arise during research processes, when combining databases containing human health data and open-access data. On the other hand, since it is optimised for ease of use by the researchers, LAT is not a tool that can be easily integrated into a data workflow by providing automatic functions usable by other software applications. Therefore, in the pilot, LAT is integrated by the insertion of a link from the BBMRI Hub to LAT, so that the user can consult it during the pilot workflow for legal and ethical advice.

3.5 Identity Management via Shibboleth

In an interconnected world, where multiple providers are able to provide integrated Internet applications and uniform experience across them, standardised solutions to manage digital identities are increasingly important. Shibboleth[11] is an open source software, which can wrap areas of a web application (i.e., URL patterns) so that, before actually serving the respective request, an unauthenticated user can be forwarded to a common login page, where (s)he can select an identity provider (IdP), such as the authentication system managed by the user's institution. After the user has authenticated with the IdP system that already knows about this user, (s)he is forwarded back to

[11] http://shibboleth.net

the initially invoked web address, where the local application (acting as a service provider, or SP) checks session attributes (automatically created by Shibboleth), uses them to decide if the now-authenticated user is authorised to access the protected areas of the application, and behaves according to the level of authorisation the user has (figure 1). Shibboleth is based on SAML[4], a popular XML-based OASIS standard used to exchange identity management-related information. As such, Shibboleth represents a flexible, standards-compliant solution to decouple application logic from application access and permission management, delegating the latter to organisation-wide identity managers. For the users, this has the main practical advantage that they need only one account to access multiple applications, and only one authentication is needed in a single session. For the IT administrators and developers, this is a modular solution, which ensures separation of concerns and allows one to flexibly combine different components. In order to further simplify identity management, Shibboleth allows one to arrange an identity federation, that is, a set of organisations and their identity providers, used to manage applications and users for which mutual trust exist (application's managers trust users coming from the federation, and the way applications work can be trusted by all users and organisations). This is relevant when some of the applications grant access to sensitive data (e.g., patient data, data affected by intellectual property concerns), for which policies in place might require specific forms of identity proofing (e.g., an application form signed by a local officer).

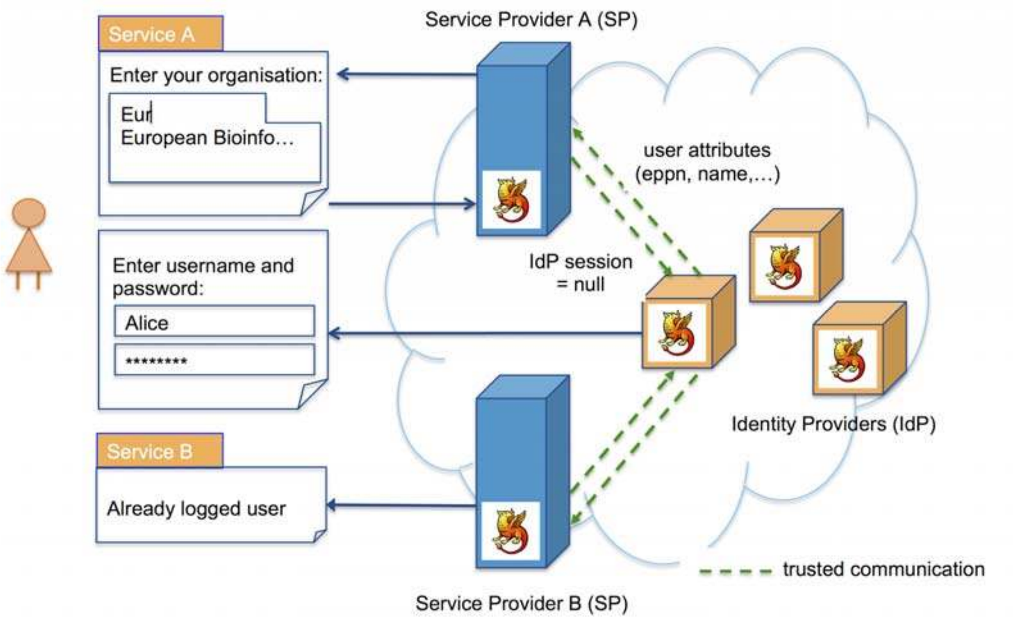


Figure 1: *The Shibboleth functional workflow. Applications (services A and B) delegate identity management to the SP component, which redirects unauthenticated users to login pages (configured with participating organisations). The users can start sessions using the account issued in their own organizations. Initialised sessions are used over multiple services automatically (service B).*

4 The pilot workflow

Figure 2 summarises the workflow that we have implemented in the pilot presented here. The diagram can be considered as a detailed version of figure 6 in D5.3 (outline of the biosamples data integration), while the workflow it represents is an implementation of the activity diagram described in section 8.3 (secure workflow specified by the security architecture).

We consider the case where a researcher is looking for samples of interest, both human and non-human, with an aim to explore experimental data derived from such samples, as well as acquire biological samples for further studies. EBI's Biosamples Database fits a part of this use case - the user will be able to find information in existing -omics repositories. For information on the availability of biobank samples, a link to the BBMRI Hub has been created. We have uploaded summaries about 'demo' data sets, available from the BBMRI

Hub, onto BioSD. For instance, a search initiated from BioSD (step 1) might lead (step 2) to the page about the data set named DE_Biobank7[12].

Such page includes a link to the corresponding information in the TUM BBMRI Hub instance, which stores details about the biobank and mediates the access to it. Similarly to the web service and Semantic Web integration work that BioMedBridges has been carrying on in WP4[13], we have chosen to involve an adapted instance of the LPC Catalogue called BBMRI Hub in the pilot in order to show that an important resource in the biomedical research can be extended to mediate access to protected biobanks. When an unauthenticated user clicks on the biobank details link on a BioSD page, the Shibboleth service provider component (SP) installed on the BBMRI Hub redirects the user to a login page (step 4). At this point the login page is presented, where the user selects an appropriate IdP (e.g., his/her organisation's IdP) and enters his/her credentials for that IdP. Upon successful authentication, the user is forwarded back to the BBMRI Hub (step 5), where the hub application checks the user session attributes that are provided by Shibboleth in a standardised format (i.e., using SAML). Namely, the 'entitlement' attribute contains the list of REMS resources the user has access to. This attribute is automatically passed from Shibboleth to the hub application by the Apache Web Server (via AJP protocol or via Request Headers). In the pilot project, the Shibboleth Attribute Authority (AA) server is used to export entitlements from REMS database. Shibboleth SP on the BBMRI Hub server knows about the AA and requests the 'entitlement' attribute automatically, just after the user is logged in (steps 6 and 7). The BBMRI Hub then presents the more fine-grained aggregate data about the requested resource that is contained in the data cube to the now authenticated user, and offers the option to refine the initial request, by querying the data cube (step 8). If the data cube indicates that a data set that is not already in the list of 'entitlement' attributes is indeed of interest for the user, (s)he can click on a link forwarding to REMS itself (step 9), where (s)he can make an application to gain access to the corresponding microdata stored by the external biobank. For instance, REMS is configured to drive the user through the forms and

[12] http://www.ebi.ac.uk/biosamples/group/SAMEG299071. The list of all the data set records uploaded for the pilot are available at http://tinyurl.com/prtvs4h.

[13] http://www.biomedbridges.eu/deliverables/43, http://www.biomedbridges.eu/deliverables/47

documents needed to complete the data access application required for the aforementioned Biobank_7 data. Such configuration is provided in advance by a user having the role of resource manager in REMS. Moreover, REMS does not authenticate the user again if the user's IdP still has a valid session. Once the application procedure is completed, it is forwarded to the data access committee (DAC) that is responsible for Biobank_7.

The BBMRI Hub provides a link to the BioMedBridges Legal & Ethical Assessment Tool (LAT), so that it can be consulted during the pilot workflow for legal and ethical information. The data managers are guided through a question-based process, and receive an appropriate list of legal and ethical requirements that must be fulfilled when dealing with the data set in question.

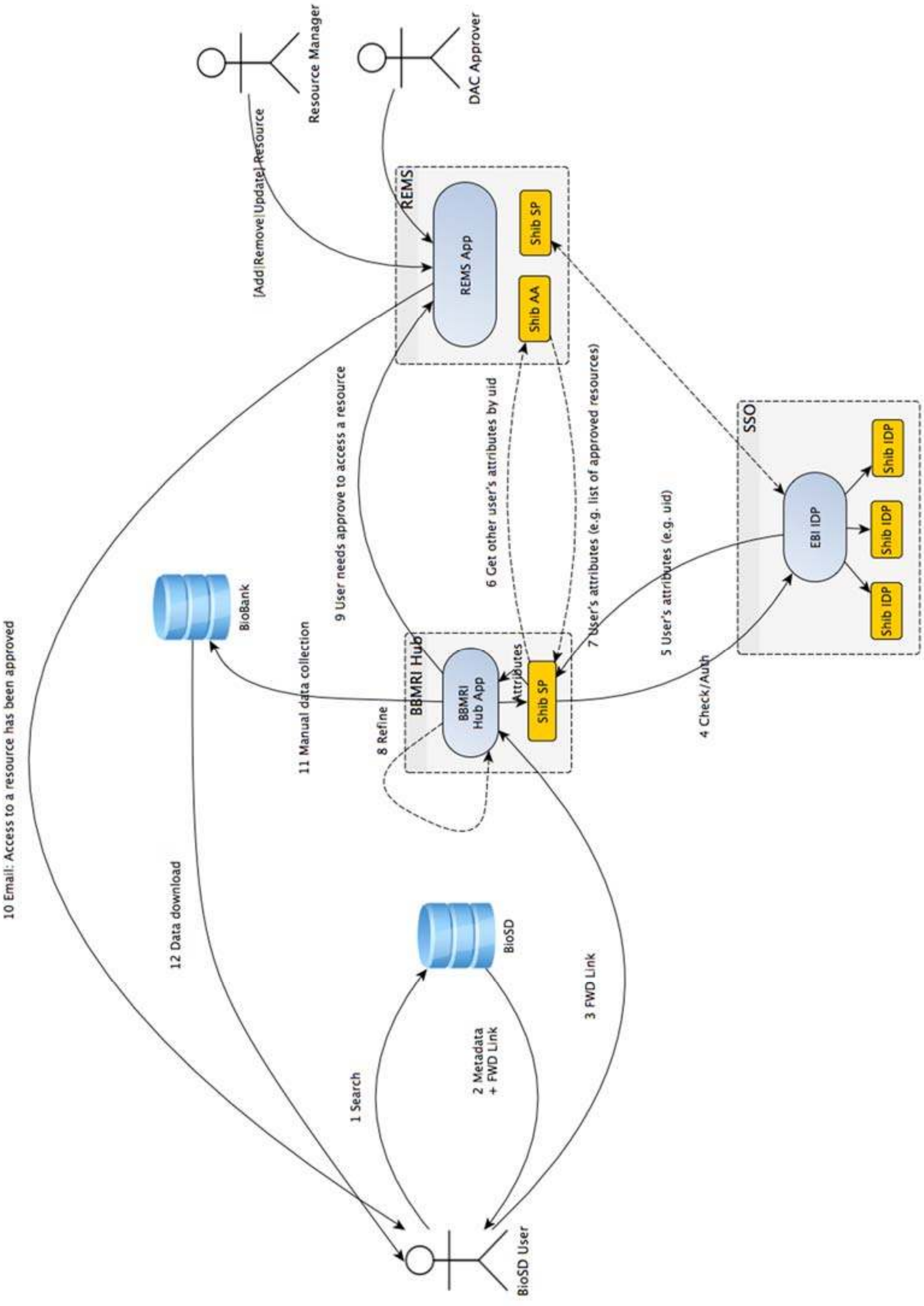

Resource Manager
DAC Approver
[Add|Remove|Update] Resource
REMS
REMS App
Shib AA
Shib SP
SSO
EBI IDP
Shib IDP
Shib IDP
Shib IDP
BBMRI Hub
BBMRI Hub App
Shib SP
Attributes
BioBank
BioSD
BioSD User
1 Search
2 Metadata + FWD Link
3 FWD Link
4 Check/Auth
5 User's attributes (e.g. uid)
6 Get other user's attributes by uid
7 User's attributes (e.g. list of approved resources)
8 Refine
9 User needs approve to access a resource
10 Email: Access to a resource has been approved
11 Manual data collection
12 Data download

Figure 2: *The workflow implemented for the BioMedBridges secure access pilot*

After an approver from the DAC confirms the user authorisation to access the biobank data, REMS notifies the user via email (step 10), where such notification contains the same forward link that has previously triggered the step 3. Now, when the user follows that link, the hub will find the proper REMS entitlement attributes (when repeating steps 6 and 7) and will forward the user to the requested data; in practice, that might be a link to another web application, or to static files, e.g., accessible through an FTP link.

Note that Shibboleth installations define a session lifetime that might be long enough to avoid another re-authentication operation when the workflow is repeated from step 3, after DAC approval (i.e., after the step 10). While this makes user’s life a bit easier, having long-living sessions (longer than few hours) is not recommendable for security reasons.

5 Implementation details and reusing the pilot for other use cases

In this section we describe the technical details concerning how we have implemented the pilot described above, and we report the lessons learnt in the process. In doing so, we focus on offering information useful when dealing with similar use cases, by means of replicating our solution, or parts of it.

5.1 BBMRI demo biobanks

The biobank metadata contained in the BBMRI Hub is fictional ‘demo’ data, providing contact e-mail addresses such as “name7@biobank7.example.com”. Similarly, we have used a BBMRI Hub instance, deployed for the demonstration purposes of the pilot. The data cube of the BBMRI Hub contains, for each Biobank and every value combination of the dimensions Diagnosis (ICD10 code groups), Medical Data available (yes/no), and Material, the corresponding number of donors and samples as facts. These numbers have been generated using a pseudo random number generator. Thereby, in order to produce sensible numbers that mimic the data available in real biobanks closely, every number of samples had to be at least as big as the corresponding number of donors, and numbers greater than zero were only permitted for dimension combinations for which samples are available in real biobanks.

5.2 Uploading biobank metadata into BioSD

The role of BioSD in this workflow is to demonstrate how an open resource can be integrated with a resource requiring authentication, across different infrastructures, i.e., ELIXIR and BBMRI. BioSD contains summary descriptions about biobanks available via the BBMRI Hub. In order to create such summaries, we prepared SampleTab submissions. SampleTab[14] is a simple tabular format that allows one to describe biomedical samples and sample collections, called sample groups in BioSD. This also includes sample and group attributes, such as ‘organism’, ‘disease’, or ‘age’. In order to accommodate the heterogeneity of data that are served, both samples and groups are modelled in a generic way: a sample can range from a single test tube, to a patient, or a lump of collected soil, and a sample attribute can be as simple as a text pair of type + label. The submitters can optionally further constrain this representation by prescribing the use of some controlled vocabulary. They can also enrich it by adding references to terms in an ontology (i.e., their URIs, or accession numbers). Similarly, groups may represent sets of samples used to perform one or more experiments, as well as more general collections, such as clinical trials, or all the samples collected in a research project.

We model individual biobanks in the BBMRI Hub as SampleTab groups, without listing any individual samples in the submission. Submitted files are reflected on the BioSD web pages[15]. This is how BioSD represents data sets where sample details have to remain undisclosed. Each submission contains a single sample group, describing the corresponding biobank. In other words, with this approach BioSD contains only biobank metadata, which are public and present no privacy concerns. Attributes are attached to these sample groups. The meaning of this is simply to indicate that there will be individual samples in the corresponding biobank that possess these characteristics. By using lexical mapping tools (like ZOOMA[16] or Bioportal Annotator [5]), ontology terms have been identified and attached to such attributes. A minimal degree

[14] https://www.ebi.ac.uk/biosamples/help/st.html

[15] For instance, for the case of aforementioned Biobank_7, the link ftp://ftp.ebi.ac.uk/pub/databases/biosamples/GSB/GSB-220/sampletab.txt is reported.

[16] http://www.ebi.ac.uk/spot/zooma/index.html

of data curation experience is necessary to do that. SampleTab 'Database' elements, which are designed to report relevant data cross-references, have been used to report links to the BBMRI Hub (which, as described above, trigger the first step of the pilot workflow). The submission files were automatically built based on the demo biobank data stored in the BBMRI Hub by simply copying organisational information, such as name and details of the data provider, and inserting the ICD10 groups and materials for which the sample counts of the biobank in question are greater than zero.

5.3 Setting up Shibboleth and the federation

There are official pre-built RPM packages of Shibboleth components available for Red Hat Enterprise 6+ and CentOS 6+. These packages are built for, and integrated with only the Apache (httpd) package that is supplied with the OS. Source code for them is available as well, in case the build from source is needed. When building from source or SRPM, it is possible to accommodate any version of Apache (or its derivations) that is compatible, but only Apache installations built using the official Apache sources are supported. Please refer to the official Shibboleth wiki for more details[17].

Once the Shibboleth SP component has been installed, you have to configure it. The default `entityID` should be changed in `shibboleth2.xml file`. New SSL certificates should be generated, in order to ensure more secure communication, and at least one "local" IdP should be configured to work with the SP. A testing service TestShib[18] can be used to test new installations of Shibboleth.

Each Shibboleth entity (SP or IdP) should provide SAML2 metadata (with `entityID`, display names, certificates, endpoint URLs, etc.) so the other Shibboleth network participants can talk to it. After Shibboleth SP and Apache web server are successfully installed, you should be able to see the auto-generated metadata by opening the URL: https://yourdomain/Shibboleth.sso/Metadata

[17] https://wiki.shibboleth.net/confluence/display/SHIB2/NativeSPLinuxInstall

[18] http://www.testshib.org/

Shibboleth trust network relies on the metadata exchange mechanism, i.e., to start talking to an IdP or a SP, you need to exchange metadata first. Here is an example of how an IdP is configured in `shibboleth2.xml`:

```
<MetadataProvider

   type="XML"

   uri="https://some-idp/idp/profile/Metadata/SAML"

   backingFilePath="some-idp-metadata.xml"

   reloadInterval="180000"/>
```

Be sure that the IdP trusts your SP as well (IdPs should have your metadata listed in their configs).

When joining a federation you will be provided with a certificate to use to verify metadata's signature to ensure its validity. Most of the time the federation will provide you with detailed instructions or examples of how to configure the software, and you should follow those instructions.

For more technical details about Shibboleth installation please follow the official Shibboleth wiki pages[19].

5.4 REMS installation and configuration

REMS is based on Java and on Liferay, a framework to build web portals, with modules (‘portlets’) to support functionality typical of Content Management Systems (CMS) and other web applications. A Shibboleth Service provider module (SP) has been integrated into the REMS architecture, as shown in figure 3. Instructions to install and configure Liferay, Shibboleth SP and REMS are available online[20]. REMS is integrated with Shibboleth SP by means of a login module, developed by CSC. This handles the SAML authentication response received from an Identity Provider, identifies the attributes received

[19] https://wiki.shibboleth.net/confluence/display/SHIB2/Installation

[20] https://confluence.csc.fi/display/REMS/Deployment+Guide+1.5

within the authentication response, and integrates them with the attributes in use within Liferay (and REMS). Upon first-time login in REMS, user information is stored to the Liferay Database, which REMS uses as a Liferay component. Configuration instructions to take federated identity into use are available online[21]. REMS is currently taking advantage of the following attributes (which are common in the SAML world): `eppn, mail, cn, sn, displayName, homeOrg, homeOrgType, entitlement, unscoped-affiliation`. Of these, only `eppn` and `email` are mandatory. One relevant thing to notice is that, in order to make REMS to work in a workflow like the pilot, organisation's IdP must be properly configured to release necessary attributes to the REMS's SP and other SPs (e.g. `eduPersonPrincipalName`).

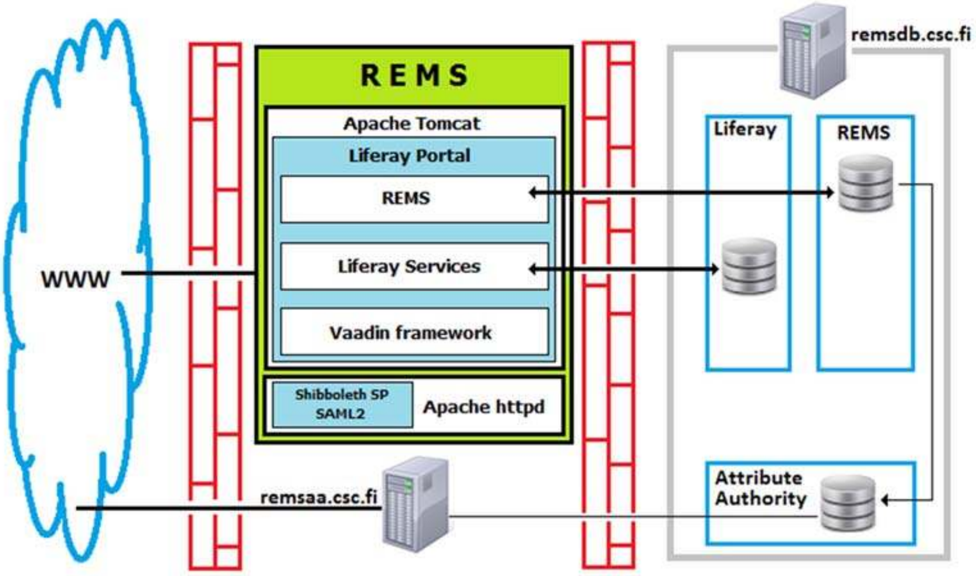


Figure 3: *REMS technical architecture.*

Regarding security concerns (see the 'Discussion' section), the REMS instance used for the pilot is protected by a baseline firewall configuration, included in REMS configuration instructions. Liferay, Shibboleth SP and the Apache httpd server are all generating log files, when configured according to our instructions. In order to improve security and quality of REMS, audit was conducted in March 2015 by an independent organisation (Second Nature

[21] https://confluence.csc.fi/display/REMS/Step+C%3A+Federated+Liferay+1.5

Security, 2NS[22]), which included software testing (based on the OWASP[23] and OSSTMM[24] methodologies) and code-review (done in co-operation with CSC staff).

5.5 Further development of LAT

After being developed in WT4 the Legal Assessment Tool (LAT) has been further improved for easier integration into the pilot described here. The original version used Liferay and PrimeFaces, resulting in a resource hungry and sometimes error prone application. The new version uses HTML and Java, runs faster and without errors. The user interface was improved, too, now displaying the requirements in the same window. The results of the evaluation are shown directly after the question criteria (e.g. data type) are indicated. The next task is the update of the knowledge base; incorporation of national regulations (Germany, UK, Holland, France,…) will provide a more complete guidance. In addition, BBMRI has offered to adopt the tool, integrating it with other tools such as hSERN and the BBMRI Legal WIKI.

5.6 Configuring Shibboleth in BBMRI

The BBMRI Hub is a Java web application running on a Tomcat application server. To implement the workflow from section 3.2 the BBMRI Hub has to be connected to an identity federation. As for other components, in the pilot we used the Shibboleth protocol and three test IdPs (named Haka Test-Idp, DFN Test-IdP 2.x and the ebi-idp). In order to connect the BBMRI Hub to the three IdPs, an Apache web server acts as a proxy (i.e. the module mod_shib) implementing the Shibboleth protocol. The Shibboleth authentication workflow is triggered by an unauthenticated user, causing the BBMRI Hub to redirect the user to a page where the desired IdP can be selected. The selection of the IdP, in turn, triggers a redirection to the IdP's web page where the actual authentication (e.g. the provision of username and credentials) is performed.

[22] http://www.2ns.fi

[23] https://www.owasp.org

[24] http://www.isecom.org/research/osstmm.html

Thereby the Apache web server acts as the service provider (SP) on behalf of the BBMRI Hub which communicates with the IdPs. In this context the SP also retrieves the authorisation statement from the REMS attribute provider. After successful authentication of a user with the IdP, the authorisation statements of this user are retrieved from REMS. As a result the obtained user name (eppn or persistent-id attribute) and the corresponding authorisations from REMS (entitlement attribute) are sent to the Tomcat server, using the AJP protocol. It should be noted that the transfer of these attributes is unprotected by default, and therefore the Apache web server and the Tomcat application server should reside on the same host, or the communication has to be additionally protected.

The obtained user name is used for the login stage and initiate a session in BBMRI. If the username is unknown to the system (e.g. the user logs in the first time) the system displays the “Terms & Conditions” to which the user has to agree. Only after agreeing, a new user account is created in the BBMRI Hub. After the creation of the account the user is redirected to the protected area of the BBMRI Hub, where he can query the data cube. Next to the query results the BBMRI Hub presents links to apply for access to the corresponding microdata. These links redirect the user to the REMS system. The links contain information about the name of the biobank and the name of the sample collection (cf. section 3.2). This allows for a context switch into the REMS system. The access decisions are conveyed via the entitlement attribute obtained from the REMS system during the authentication process. They are then used to allow access to biobank microdata. In the pilot links to dummy microdata to which access has been granted are provided.

In the following, configuration snippets are presented which have been used to implement the Shibboleth process in the BBMRI Hub. As already mentioned, the BBMRI Hub prototype is connected to the DFN-AAI-Test and the HAKA-Test federations, as well as the EBI-IdP. The configuration parameters of the Apache web server for this scenario are shown below. More specifically, the part of the `shibboleth2.xml` configuration file specifying the IdPs for use by the mod_shib module reads as follows:

```
<MetadataProvider type="Chaining">
 <MetadataProvider type="XML"
uri="https://www.aai.dfn.de/fileadmin/metadata/DFN-AAI-Test-metadata.xml"
```

```
       backingFilePath="DFN-AAI-Test-
metadata.xml"  legacyOrgNames="true"  reloadInterval="7200">
       <MetadataFilter type="RequireValidUntil"
maxValidityInterval="2419200"/>
       <MetadataFilter type="Signature" certificate="/etc/shibboleth/dfn-
aai.pem"/>
  </MetadataProvider>
  <MetadataProvider type="XML"
uri="https://haka.funet.fi/metadata/haka_test_metadata_signed.xml"
       backingFilePath="haka-test-idp-metadata.xml" legacyOrgNames="true"
reloadInterval="7200"/>
  <MetadataProvider type="XML"
uri="https://idp.ebi.ac.uk/idp/profile/Metadata/SAML"
       backingFilePath="ebi-idp-metadata.xml" legacyOrgNames="true"
reloadInterval="7200" />

</MetadataProvider>
```

To integrate the REMS system as an attribute provider the following section has to be added in the shibboleth2.xml file:

```
<AttributeResolver type="Chaining">
   <AttributeResolver type="Query"/>
<AttributeResolver type="SimpleAggregation" attributeId="eppn"
                          format="urn:oid:1.3.6.1.4.1.5923.1.1.1.6">
          <Entity>https://remsaa.csc.fi/idp/shibboleth</Entity>
          <Attribute Name="urn:oid:1.3.6.1.4.1.5923.1.1.1.7"
                     NameFormat="urn:oasis:names:tc:SAML:2.0:attrname-
format:uri"
                     FriendlyName="eduPersonEntitlement"/>
   </AttributeResolver>
</AttributeResolver>
```

As all environment variable names which start with “AJP_” are transferred via the ajp proxy the Shibboleth module has to be configured to prepend this string to all its attributes:

```
<ApplicationDefaults
entityID="https://shibboleth.imse.med.tum.de/shibboleth"
attributePrefix="AJP_"
                         REMOTE_USER="eppn persistent-id targeted-id">
```

After configuring the Shibboleth module, the Apache web server has to be configured to invoke the authentication workflow if a directory will be accessed. The following has to be added:

```
<Location /path_to_protected_directory>
```

```
   ProxyPass /application_path ajp://127.0.0.1:8009/application_path
   AuthType shibboleth
   ShibRequireSession On
   require valid-user
   ShibUseEnvironment On
   ShibUseHeaders Off
</Location>
```

It should be noted, that the mod_proxy[25] module should be configured in a way such that no external proxy requests are processed.

To access the attributes, which are passed via the AJP protocol to the Tomcat server, in Java the following code snippet can be used:

```
String attribute = (String) request.getAttribute("attribute_name");
```

It should be noted, that attribute names that are forwarded are not included in the list returned by request.getAttributeNames() call, and that the forwarded attributes has to be accessed without the “AJP_” prefix. To obtain the username from the “eppn” attribute the following code could be used:

```
String username = (String) request.getAttribute("eppn");
```

To use a Shibboleth federation in other settings, the above metadata provider specifications and the REMS instance defined as the attribute resolver have to be adapted accordingly.

6 Discussion

6.1 The pilot and STRIDE/LINDDUN threats

In this section we review how the pilot addresses the issues presented in the BiomedBridges D5.3 report, in particular, the aspects that that report presents

[25] For more details see: https://httpd.apache.org/docs/2.2/mod/mod_proxy_ajp.html

in sections 7.1 and 7.3, including tables 14 and 15. Table 22 in section 13.1.4 is also taken into consideration. This is about security and privacy threats that the STRIDE and LINDUN methodologies identify for a data management infrastructure, and the countermeasures one can adopt to eliminate or mitigate such threats.

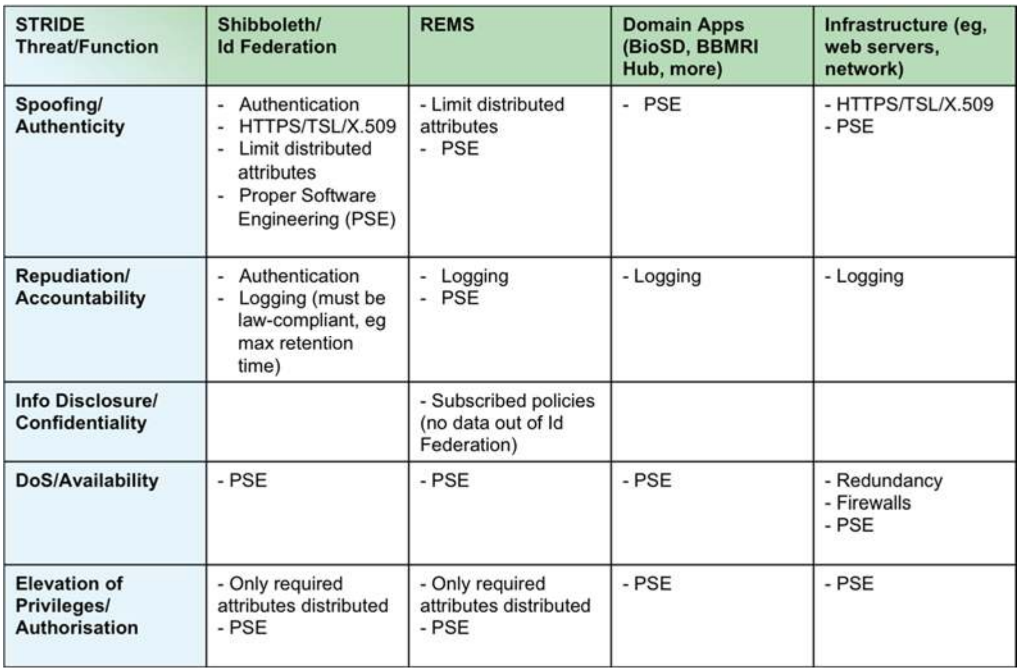
Combining Shibboleth and REMS, to ensure that only authenticated users have access to resources they have been authorised to, limits the risk that confidential information is disclosed to unauthorised persons. REMS simplifies the management of access policies and helps in matching legal requirements on data access with the authorisations that are granted to the users. This includes the management of application details, such as the requirement to specify research purposes for which data are needed, or the requirement to confirm that the relevant policy and regulation documents have been read and will be adhered to. Moreover, REMS records its actions, the data usage terms the users have committed to, and the authorisations that have been given. This form of audit trail offers a basis both to keep the evidence of compliance with law and regulations, and to hold users and data managers accountable for their actions.

Shibboleth wraps the pilot components with a reliable technology to manage digital identities. Authentication operations are often forwarded to organisation-provided, equally reliable systems, which usually are already managed with proper expertise and security policies (additionally, Shibboleth includes an IdP component in its suite). This helps to prevent the spoofing risk (i.e., pretending to be someone else, or using unauthorised credentials). Moreover, the Shibboleth technology limits the access to REMS attributes that are distributed for only those components needing them, which is a feature that mitigates the risk of elevating privileges, for instance by preventing the interception or faking and session-injection of non existing authorisations (i.e., REMS attributes).

The Shibboleth system manages web technology-based communications by enforcing HTTPS connections, on top of otherwise-unencrypted HTTP protocol. This is a standard practice in the World Wide Web, and ensures secure communications between network-distributed applications, based on reliable standards like TLS [6]. Further communication protection is usually set up in Shibboleth to encrypt the SAML messages that the system components exchange (e.g., user credentials), standards like X.509 and XML Encryption[26] are used for that. This kind of protection addresses the risk of exchanged data tampering and information disclosure. While protocols like HTTPS cannot prevent denial of service attacks on their own, most of the systems participating in the pilot are protected by the respective institutional IT policies, which include firewall-based IP filtering and load balancers. Similarly, web servers are configured to generate log files and other access evidence (which, additionally, comply with other legal requirements, such as the maximum time private data can be retained). This, combined with authentication, helps in ensuring non-repudiation and user accountability. Not all the components in the pilot currently enforce HTTPS access (over plain HTTP) at all times, e.g., BioSD can be used through HTTP when accessing public information. Similarly, until the user attempts to gain access to secured information, authentication is not triggered in components like BioSD (i.e., the user remains anonymous, although their IP address is tracked). This scenario, where security wrappers are enabled only for certain components and operations, is typical of complex Internet-based infrastructures. While this is usually not problematic for what concerns the security, one can always adopt the precaution of establishing the same security protocols/technologies, regardless of the component considered and the action being taken. The implementation/management overhead that this policy causes is not likely to be significant.

Other precautions that were made while working on the pilot components and their integration were general best practices for software development and deployment. Namely, application configurations are carefully managed, so that unsafe settings do not compromise security or data protection. Placing clear text passwords in configuration files, or defining too liberal access rights are some examples of bad settings, which we have avoided. Prevention of code injection attacks [7] and thorough testing [8, 9] were adopted in the development of the pilot components, and this is recommended in similar situations.

[26] http://tools.ietf.org/html/rfc2459



STRIDE Threat/Function	Shibboleth/ Id Federation	REMS	Domain Apps (BioSD, BBMRI Hub, more)	Infrastructure (eg, web servers, network)
Spoofing/ Authenticity	- Authentication - HTTPS/TSL/X.509 - Limit distributed attributes - Proper Software Engineering (PSE)	- Limit distributed attributes - PSE	- PSE	- HTTPS/TSL/X.509 - PSE
Repudiation/ Accountability	- Authentication - Logging (must be law-compliant, eg max retention time)	- Logging - PSE	- Logging	- Logging
Info Disclosure/ Confidentiality		- Subscribed policies (no data out of Id Federation)		
DoS/Availability	- PSE	- PSE	- PSE	- Redundancy - Firewalls - PSE
Elevation of Privileges/ Authorisation	- Only required attributes distributed - PSE	- Only required attributes distributed - PSE	- PSE	- PSE

Figure 4: *How STRIDE threats are addressed in the pilot. PSE refers to software design and testing, best practices, established methodologies, techniques and frameworks. As for the life sciences-specific risks identified by the LINDUN methodology, REMS policies, as well as security and reliability of all pilot software components help with mitigating all those risks.*

The aspects of data anonymisation and identifiability were not directly addressed in the pilot. Rather, they are delegated to the data management policies as defined for the local biobanks, which are responsible for collecting informed consent declarations from patients, detailing the kind of data usage that was agreed upon, and managing data access through DACs. Once the data have been anonymised and/or pseudonymised, the transmission over HTTPS from local biobanks to the BBMRI Hub and BioSD users further limits the danger that unauthorised persons intercept the data and attempt person re-identification. Another way to limit the re-identification risk is to ensure that authorised users employ the data only for the kinds of usage they have been granted. Although that cannot be enforced technically by the pilot workflow, rigorous permission management (by the means of REMS) helps in mitigating such a concern. Moreover, relying on IdPs participating in an acknowledged identity federation improves the reliability of user identities, since they are verified by trusted organisations. The pilot infrastructure is compatible with the addition of further data access policies. For example, biobank managers can require (through the REMS application forms) that the data must never leave the network and the IT devices controlled by the organisations participating in the pilot. The technical approach to enforcing such a policy might be encrypting biobank data with systems that require hardware-bound decryption keys [10]. Figure 4 summarizes the above discussion.

6.2 Agreeing on IdP-released attributes

A well-known ‘social’ problem is the difficulty to get organisations participating in an identity federation to agree on their IdPs releasing a sufficient set of attributes, so that different components in the federation can interact to control and grant user authorisations to access services and data. This is often problematic, due to user privacy concerns and the necessity to make ‘n*m’ negotiations to have the minimum set of attributes exchanged between n IdPs and m SPs. Similar problems are posed by the need for consensus on aspects like the strength of the authentication methods in place in each IdP, or the kind of security auditing that has been carried out for IdPs. Dealing with these issues in this pilot and similar scenarios is relatively easy, since we do not have many participating IdPs or organisations, and we require that just a few attributes that are not privacy-critical are exchanged. Apart from that, we have to admit that, should the number of the involved IdPs increase, the problem would get harder, and no easy ultimate solution exists yet.

A possible compromise solution to this problem can be to convert the ‘n*m’ challenge into an ‘n+m’ challenge, which is often a much easier setup. That can be done by introducing a proxy server between the IdPs and SPs; the proxy acts as an SP towards the IdPs (managed by the researchers home institutions), and as an IdP for the SPs (the actual services the researcher is accessing). It may be easier to isolate the attribute release negotiations to a single place (the proxy server), which can also maintain discipline-specific additional user attributes. The ELIXIR research infrastructure is deploying this approach in the ELIXIR EXCELERATE project[27].

[27] https://www.elixir-europe.org/events/introduction-elixir-excelerate

7 Possible future developments

7.1 ELSI tools

In 2014 a workshop about Ethical, Legal and Social Issues (ELSI) implied in the management of personal data was organised by BioMedBridges participants[28]. Several tools developed within or outside the BioMedBridges project to manage those issues were discussed. The workshop showed that researchers need support for their data protection needs, and that several tools exist that provide information, links to regulations, and templates for regulatory documents. In the pilot the LAT has been linked to the BBMRI hub to provide users with legal requirements for specific data sharing situations. In this context, we believe such tools in general may be relevant to the pilot, and could be useful in helping users of similar infrastructures. For instance, the BBMRI Legal Wiki (a wiki about ELSI issues in Europe)[29] and the Human Sample Exchange Regulation Navigator (hSERN, a web resource about legal aspects involved in exchanging human biobanking information)[30] could be used as general references by REMS users having the role of DAC members. Similarly, hSERN might be a significant reference for biobank providers requiring help with defining their data access policies. Links to the documentation available through these two tools could be provided to the users authorised to access biobank data who need to be aware of their responsibilities about ELSI. The International Policy interoperability and data Access Clearinghouse (IPAC, an information tool about policy interoperability and access authorisation)[31] might be useful as a reference for biobanks and REMS users with the DAC role when they need to draft policy documents and forms to be uploaded to REMS and presented to data access applicants (IPAC contains templates that might be used with little or no modification).

[28] http://tinyurl.com/ptlyaya

[29] http://www.bbmri-wp4.eu/wiki/index.php/Main_Page

[30] http://www.hsern.eu/

[31] http://p3g.org/ipac

7.2 Programmatic data access

The pilot focuses on organising data access for human users. For the sake of simplicity and separation of concerns, we have not considered the management of biobank data access from fully automated software components. These obviously have different ways of interacting with data repositories. For example, a long-running web client cannot normally type a password, and proper means are needed to safely automate authentication operations. Here we outline some ideas about how the pilot work could be extended to support such use cases. Prior to dealing with technical aspects, it should be pointed out that this kind of data use must be compatible with legal and policy requirements in place. For instance, it might be the case that a REMS application requires to disclose the fact that data will be analysed offline, by means of a web service. Another general consideration is that the LINDUN/STRIDE methodologies mentioned previously should be applied to components like web services, as they are for all the components of a data access infrastructure[32].

7.2.1 User applications based on web services or similar components

Web services [11] is an important way to realise distributed computing over the web. BioMedBridges widely leverages them to pursue the project objectives[33]. As an example relevant to our pilot scenario, suppose a user accesses data starting from an application like BioSD and following the pilot workflow. Suppose that, in order to carry out the desired computations, the application at the begin of the workflow uses a number of web services that fetch biobank data from the BBMRI Hub, in a way similar to how BioSD interacts with the Hub. Assuming that the user is interacting with the system, we can consider that such a web service plays the same (client) role that a web browser plays in the pilot workflow, as we have described above. The web service must be aware that the BBMRI Hub request might return a Shibboleth link instead of the desired data URL, and that such link must be presented to the user for

[32] https://www.owasp.org/index.php/Web_Service_Security_Cheat_Sheet

[33] http://dx.doi.org/10.5281/zenodo.11891, https://zenodo.org/record/19201

authentication. Sample implementations of such a use case are available[34]. The security of this scenario can be reinforced by implementing a token-based access mechanism at the web service level. In such case, the user is asked both to initiate a session using an IdP, and to authorise (for a limited time) the web service to operate on his/her behalf. SAML (the standard backing Shibboleth) can already support such feature. OAuth[35] is a popular alternative for that, which could be used in combination with SAML[36].

7.2.2 Unattended batch processing

Users might want to download biobank data in an unattended, batch mode, for example, when long-running data analysis computations are needed, or when local (client side) copies of large batches of data are periodically updated. Such use case is more complicated than the previous one, since the user is not there to respond to the authentication challenge, and hence the usual browser forwarding cannot work. A simplistic solution might be to use a user account for the batch component, making it behave like a human who performs the login operation (tools like iMacros[37] or Selenium[38] are available for doing that). This requires to store passwords in program code, or configuration files, and is not recommended due to security considerations. A safer approach is to use time-limited authentication tokens, where the user generates a token for each of the batch components that later needs it; this is token-based authentication, as mentioned above. Using key pairs as tokens ensures even more guarantees against identity spoofing[39]. MyProxy[12] is a tool that works this way; we are not aware of any off-the-shelf solutions that could be integrated into Shibboleth, although there are examples about extending Shibboleth with new identity provider services[40].

[34] http://www.predic8.com/shibboleth-web-services-sso-en.htm

[35] https://tools.ietf.org/html/rfc6749

[36] https://tools.ietf.org/html/rfc7522

[37] http://imacros.net

[38] http://www.seleniumhq.org/

[39] https://help.github.com/articles/generating-ssh-keys/

[40] https://wiki.shibboleth.net/confluence/display/SHIB2/IdPAuthExternal

7.2.3 Semantic Web and SPARQL endpoints

A SPARQL endpoint is a standard way to make linked data available, modelled using the Semantic Web approach [13, 14]. This form of structured data publishing makes data more machine readable, and eases automated data processing, including discovery and integration. In BioMedBridges the implementation of biomedical data exchange by means of such technologies was studied extensively[41]. SPARQL endpoints exist for both BioSD[42] and the catalogue-level data available from the BBMRI Hub[43]. The BioMedBridges deliverable report D4.6 deems these technologies unsuitable to serve confidential clinical data. While this is a reasonable conclusion when the current state of art is considered, we would like to suggest how the experience gained with the pilot might help to protect SPARQL-based data access. Villata et al [15] describe an ontology to define access policies on linked data subsets. They also show a system where this ontology is used to control access to data divided in subsets by means of named graphs [16]. A similar approach is described in [17]. Here, the architecture of the FedX SPARQL engine [18] is extended, so that the initial selection of SPARQL sources (serving clinical data), which are suitable to answer a given query, is further filtered by using the authorisations that the currently authenticated user possesses over the selected sources, to further eliminate the inaccessible ones. This approach is closer to the pilot scenario. In fact, it is natural to think that a SPARQL endpoint for a biobank would be analogous to the sources queried via FedX, and that user authorisations would come from REMS attributes, served in a Shibboleth session. That would clearly imply that the SPARQL endpoint from which the query is started would be wrapped by Shibboleth, in the same way we have described above the case of web services. That is unsurprising, since a SPARQL endpoint is a particular type of web service, hence one can apply the previous considerations to it, in addition to the specific control mechanisms for the contents that they provide.

[41] http://www.biomedbridges.eu/deliverables/44, http://dx.doi.org/10.5281/zenodo.14071

[42] http://www.ebi.ac.uk/rdf/documentation/biosamples

[43] https://www.bbmriportal.eu/bbmri2.0/sparql.html

8 Delivery and schedule

The delivery is delayed: □ Yes √No

9 Adjustments made

No adjustments were made to the deliverable.

10 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

11 References

1. Gostev M, Faulconbridge A, Brandizi M, Fernandez-Banet J, Sarkans U, Brazma A, Parkinson H: **The BioSample Database (BioSD) at the European Bioinformatics Institute**. *Nucleic Acids Res.* 2012, **40**:D64–70.

2. Gray J, Bosworth A, Lyaman A, Pirahesh H: **Data cube: a relational aggregation operator generalizing GROUP-BY, CROSS-TAB, and SUB-TOTALS**. In *Proceedings of the Twelfth International Conference on Data Engineering*. 1996.

3. Lappalainen I, Almeida-King J, Kumanduri V, Senf A, Spalding JD, Ur-Rehman S, Saunders G, Kandasamy J, Caccamo M, Leinonen R, Vaughan B, Laurent T, Rowland F, Marin-Garcia P, Barker J, Jokinen P, Torres AC, de Argila JR, Llobet OM, Medina I, Puy MS, Alberich M, de la Torre S, Navarro A, Paschall J, Flicek P: **The European Genome-phenome Archive of human data consented for biomedical research**. *Nat. Genet.* 2015, **47**:692–695.

4. Bertino E, Takahashi K: *Identity Management: Concepts, Technologies, and Systems*. Artech House; 2011.

5. Jonquet C, Shah NH, Musen MA: **The open biomedical annotator**. *Summit on Translat. Bioinforma.* 2009, **2009**:56–60.

6. *SSL and TLS: Designing and Building Secure Systems*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.; 2001.

7. Su Z, Zhendong S, Gary W: **The essence of command injection attacks in web applications**. In *Conference record of the 33rd ACM SIGPLAN-SIGACT symposium on Principles of programming languages - POPL’06*. 2006.

8. Potter B, McGraw G: **Software security testing**. *IEEE Security & Privacy Magazine* 2004, **2**:81–85.

9. Myers GJ, Sandler C, Badgett T: *The Art of Software Testing*. John Wiley & Sons; 2011.

10. Löhr H, Hans L, Ahmad-Reza S, Marcel W: **Securing the e-health cloud**. In *Proceedings of the ACM international conference on Health informatics - IHI ’10*. 2010.

11. Richardson L, Ruby S: *RESTful Web Services*. “O’Reilly Media, Inc.”; 2008.

12. Novotny J, Tuecke S, Welch V: **An online credential repository for the Grid: MyProxy**. In *Proceedings 10th IEEE International Symposium on High Performance Distributed Computing*. 2001.

13. Berners-Lee T, Hendler J, Lassila O: **The Semantic Web**. *Sci. Am.* 2001.

14. Antoniou G, Groth P, van Harmelen F, Hoekstra R: *A Semantic Web Primer: third edition*. MIT Press; 2012.

15. Villata S, Serena V, Nicolas D, Fabien G, Amelie G: **An Access Control Model for Linked Data**. In *Lecture Notes in Computer Science*. 2011:454–463.

16. Carroll JJ, Bizer C, Hayes P, Stickler P: *Named Graphs, Provenance and Trust*. HP Labs; 2004.

17. Leida M, Marcello L, Andrej C: **Distributed SPARQL Query Answering over RDF Data Streams.** In *2013 IEEE International Congress on Big Data*. 2013.

18. Schwarte A, Andreas S, Peter H, Katja H, Ralf S, Michael S: **FedX: Optimization Techniques for Federated Query Processing on Linked Data**. In *Lecture Notes in Computer Science*. 2011:601–616.