ADOPT BBMRI-ERIC
GRANT AGREEMENT NO 676550

DELIVERABLE REPORT

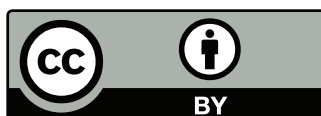| Deliverable no	D3.4
Deliverable Title	Implementation of the BBMRI-ERIC architecture and database
Contractual delivery month	M36 September 2018 (second revision)
Responsible Partner	DKFZ
Author(s)	Lars Ebert, Jori Kern, Martin Lablans, Diogo Alexandre, Frank Ückert, Petr Holub, Michael Hummel, Morris Swertz, Rumyana Proynova, David van Enckevort, Jonathan Jetten

IMPLEMENTATION OF THE BBMRI-ERIC ARCHITECTURE AND DATABASE

Executive Summary

The key goal of the BBMRI-ERIC initiative is to provide access to quality controlled human biological samples and associated medical and biomolecular data. The ADOPT BBMRI-ERIC project is a part of this initiative and has created the infrastructure needed for this goal. This infrastructure includes a software system, which allows search for biomedical data and controlled access to this data, as well as support for the sample request process. This report describes the software architecture of this system. It specifies the software components built or integrated into the system, and provides a brief overview of their function and an explanation of how they support the project's goals.

COPYRIGHT NOTICE

DOCUMENT LOG

Issue	Date	Comment	Author
D3.4rev1	2016-09-30	Revision 1	Diogo Alexandre, Martin Lablans, Frank Ückert, Petr Holub, Michael Hummel, Morris Swertz, Rumyana Proynova, David van Enckevort, Jonathan Jetten
D3.4rev2	2019-04-05	Revision 2	Lars Ebert, Jori Kern, Martin Lablans, Petr Holub

Glossary

AAI An infrastructure comprising one or more services used to authenticate the user and provide sufficient information for applications to make authorization decisions.

API Application Programming Interface

BBMRI-ERIC Biobanking and BioMolecular resources Research Infrastructure - European Research Infrastructure Consortium.

BBMRI-ERIC Directory A service provided by BBMRI-ERIC to enable basic findability of biobanks and their collections of samples/data. `https://directory.bbmri-eric.eu/`

BBMRI-ERIC Locator A service co-developed by BBMRI-ERIC and German Biobank Alliance to enable advanced findability of biobanks and their collections based on sample-level and donor-level data. `https://http://search.germanbiobanknode.de/`

BBMRI-ERIC Negotiator A service provided by BBMRI-ERIC to allow effective communication between researchers, requesting samples/data or other services, and biobanks. Subject of this ADOPT Deliverable D4.4. `https://negotiator.bbmri-eric.eu/`

BIMS Biobank information management system

Common Service A Common Service means a facility of BBMRI-ERIC according to Article 15(1) according to the Statutes.

Common Service IT Common Service on Information Technologies (IT)

CSS Cascading Style Sheets, a W3C standard. `https://www.w3.org/Style/CSS/Overview.en.html`

DICOM® Imaging standard in radiology, `https://www.dicomstandard.org/`

DKTK German Cancer Consortium, `https://dktk.dkfz.de/`

EDC Electronic data capture

ELSI ethical, legal, and societal issues

ETL Extract, transform, and load process

GBA German Biobank Alliance

GBN German Biobank Node

GUI Graphical user interface

HL7 FHIR® HL7 Fast Healthcare Interoperability Resources Specification. `https://www.hl7.org/fhir/`

ICD-10 WHO International Classification of Diseases, 10th revision

JSON JavaScript Object Notation, subject to RFC 8259 and ECMA-404 standards

LDAP Lightweight Directory Access Protocol, subject to X.500 standard

LINDDUN Linkability, Identifiability, Non-repudiation, Detectability, Disclosure of information, Content Unawareness, Policy and consent non-compliance, [1]

MDR Repository of metadata developed in OSSE Project[1] based on ISO/IEC 11179 standard [6].

REST Representational state transfer (REST) or RESTful web services, see `https://en.wikipedia.org/wiki/Representational_state_transfer`

---

[1] `https://www.osse-register.de/en/`

SSO Single sign-on

STRIDE Spoofing, Tampering, Repudiation, Information Disclosure, Denial of service, Elevation of privilege, [5]

UI User interface

Contents

1 Architectural philosophy

Architectural goals, often driven by software requirements, provide the motivation and rationale for decisions. Here it is defined how the system responds to change over time.

- Aggregated data of samples from participating biobanks are accessible in a central repository called BBMRI-ERIC Directory, grouped per collection. Collections in the Directory are based on the definition from MIABIS, with the additional refinements as outlined in the ADOPT Deliverable D3.1 [3].

  - Collections are containers for sample sets and/or datasets, with support for recursive creation of sub-collections. Here properties of the samples and data can be described in aggregated form such as sample counts, diseases, material types, data types, gender, etc. To enable deterministic counts for samples we followed recommendation of MIABIS 2.0 [7] that (sub)collections are strictly based on the concept of set partitioning: for any collection containing countable (discrete) elements (such as samples/aliquots, images), each element must be exactly in one collection (partition) on the given level of recursion, and there must be no empty collections. This allows for straightforward aggregation: content of each parent entity, be it a collection or a biobank, is a sum of child entities – collections, subcollections, etc.

  - The directory is also using a flat data model, which is useful for filtering out biobanks, which are certain to not have the needed samples. A more sophisticated data model is used for the BBMRI-ERIC Locator, which allows fine-granular selection of suitable samples.

- To allow searching data at the non-aggregated sample level, a federated search system called BBMRI-ERIC Locator has been implemented. By default, there are automated replies from the connected biobanks. In the first step, these replies from all biobanks are shown in an aggregated form. After logging in and disclosing some information about the research project, the inquirer gets a view on the number of potential matches per biobank.

- A Metadata Repository (MDR), following a generally accepted international standard, describes all data elements and their relationships within BBMRI-ERIC Common Service IT.

- Data elements in every participating biobanking metadata is mapped to the metadata present in the MDR.

- The actual matchmaking between the requester and the candidate biobanks identified using BBMRI-ERIC Directory or BBMRI-ERIC Locator can be done using BBMRI-ERIC Negotiator, which has been described as a part of ADOPT Deliverable D4.4.

- Shared non-aggregated sample data is pseudonymised locally, to ensure that participants are not identifiable.

- Every service/component developed is designed with privacy in mind and the developments respect the BBMRI-ERIC Common Service IT Security and Privacy Architecture developed as ADOPT Deliverable D3.2 [4].

1.1 Cooperation of BBMRI-ERIC with its German Biobank Node (GBN) and the German Biobank Alliance (GBA)

Biobanking IT at the national level and on a European scale have similar requirements. While BBMRI-ERIC connects European biobanks, the German Biobank Alliance (GBA) applies this concept to Germany. From a researcher's point of view, it would be beneficial to contact as many biobanks as possible using one single entry point.

The collaborative design and development of IT components is key for the success of BBMRI-ERIC, German Biobank Node (GBN) and GBA. Thus, there has been intensive cooperation between these IT teams including members at each of the GBA biobanks to achieve a smooth and productive development. Especially for the development of the BBMRI-ERIC Locator (SearchBroker in GBA) and BBMRI-ERIC Negotiator [2] this was of utter importance.

The existing requirements, workflows and software modules of both networks were analysed. Moreover, a wide group of stakeholders was consulted to collect further requirements. Then a harmonised workflow and software architecture were defined and consented in a collaborative effort. This should enable researchers to search for biomaterial within all participating biobanks.

1.2 Cooperation / work tools

From the start, a communication platform (based on the Confluence software) was established, in which all activities, workshops, protocols, shared working spaces to prepare documents, etc. are filed. All working groups use this transparent information platform extensively.

In order to establish the biobank IT network for BBMRI-ERIC and German biobanks, joint processes and tools for IT cooperation across the involved partners were introduced: For example, the JIRA development tool was introduced, which makes it easy to distribute tasks and responsibilities in the form of tickets. The developer network systems GitHub and Bitbucket were selected, which is used to coordinate parallel IT developments of individual software modules at different locations. To support communication between sites, daily (GBA), weekly (BBMRI-ERIC) and monthly teleconferences and face-to-face meetings were defined, the team was brought together and group-specific responsibilities were distributed.

2 Workflow

Having logged onto the account (ideally of their home institution), researchers can create a sample and data request within the context of their planned research project. This request is distributed to all connected biobanks. The biobanks in turn may configure their replies in accordance with local policies. By default, the number of potentially matching samples is replied. Based on these results, the researcher can then select biobanks to discuss a cooperation for his/her research project. This negotiation is supported by a communication platform tailored to the needs of the participants.

A software suite implements the previously defined workflow. Some of the software modules are central, such as the Authentication, the BBMRI-ERIC Locator, the BBMRI-ERIC Negotiator and the MDR. The local component is called Bridgehead and consists of the Connector and a local data warehouse containing the pseudonymised data. The bridgehead is installed within every participating biobank.

Within the close cooperation with GBA, the current version of those software components have been shared with the IT teams of the German, Austrian, Czech and Finnish BBMRI nodes. Furthermore, BBMRI-ERIC and GBN/GBA collaborated closely on harmonised data sets for basic sample information and extended versions for various diseases. This also includes the latest version of the MIABIS description. For the colon cancer data set, which is part of the H2020 project ADOPT, GBN plays a role in the definition of the respective data set and establishment of the required database. The total number of validated colon cancer cases is currently around 10,000 data sets. Finally, yet importantly, BBMRI.nl and GBN collaboratively developed a German directory instance within the Dutch Molgenis database structure. This German instance is synchronised with the BBMRI-ERIC directory on a daily basis. BBMRI-ERIC and GBA are currently working together to improve the directory data quality and usability.

At the end of 2017, the first German biobanks entered the European biobank registry at `https://directory.bbmri-eric.eu` in agreement with BBMRI-ERIC. One next big step in the activities is to go open source in order to enable other international biobanks as well as German biobanks, which are not funded by the Federal Ministry of Education and Research (BMBF), to install Connectors (Bridgeheads) in their own institutions.

3 Architecturally significant requirements

Architecturally significant requirements are those requirements that played an important role in determining the architecture of the system.

- The integration of the BBMRI-ERIC Directory, BBMRI-ERIC Locator, Connector, BBMRI-ERIC Negotiator and Harmonization services are able to exchange data, but we kept a maximum independence of the components.

    - As a practical example, if the MDR would go down, the other components will not crash because of it and the user will get proper feedback. In the specific case of the MDR being down, as we have the "mdrclient" – a java library, RESTful client that transparently caches MDR data – the components who reuse it automatically support caching and therefore remain able to work with the metadata. However, if some needed metadata is not cached, then the system will specifically mention (to the user) that the MDR is momentarily not accessible.

- The different components are loosely coupled – all communication is done asynchronously and no requests or data are lost whenever one or more of these components fails

- Communication between the system components happens via RESTful or HTTP/JSON web services

- The Connectors poll the queries from the BBMRI-ERIC Locator, due to the clinical restrictive firewalls

- All distributed data communication is done through HTTP over TLS (HTTPS)

- The user interfaces from the different system components ideally are based on Bootstrap and on a common BBMRI-ERIC Common Service IT Cascading Style Sheets (CSS) template (conforming to Bootstrap framework): `https://github.com/molgenis/molgenis-projects/tree/master/BBMRI_eric/css`

- The MDR offers a REST interface for other system components connected to the BBMRI-ERIC Common Service IT central authentication and authorization component

- The MDR is based on the ISO 11179 definition of metadata items, particularly oriented by the "Registry metamodel and basic attributes".

- Each Connector is running under the biobank's organization control.

4 Decisions and constraints

There is a variety of factors that place constraints on the architecture. These architectural constraints, combined with the requirements, helped defining the system architecture.

- The BBMRI-ERIC Common Service IT components are web applications that are able to communicate through REST and REST-like HTTP/JSON web services.

- For a seamless user experience, the user interfaces of the BBMRI-ERIC Directory and the BBMRI-ERIC Negotiator have the same look and feel matching BBMRI-ERIC corporate identity.

- Due to the distributed nature of the system and the diversity of the teams working on it, the development and database technologies used are not bound to a vendor, but are based on open source and on technologies that each team is familiar with.

- The software produced will be shared as open source on GitHub. The source code itself is prepared; however, some negotiations are still ongoing since the software for central and local software components has been developed in cooperation and in agreement participants at several independent institutions.

- Software code is managed in version control (GitHub and Bitbucket (internally)) and had stable releases with stable external interfaces to ensure interfacing components depending on them to not break on a release.

- In order to increase quality of the delivered product, the following reviews was be implemented:

  - Architectural review with focus on privacy & security by design.

  - Architectural review by ethical, legal, and societal issues (ELSI) group - focusing on overall workflow, privacy, informed consent management, confidentiality of projects and project proposals, etc.

- The following reviews is yet to be implemented:

  - Code review: use of Coverity service free for open-source software, cross-WP code reviews for major releases.

5 Architectural Mechanisms

Availability   The Common Service IT is a distributed system in which different components are located on different networks and countries, and communicate and coordinate their actions by passing messages. Although a specific component might fail temporarily, the system as a whole continues running independently. This has been achieved by asynchronous communication between the components.

Disaster Recovery   Being a distributed system, developed by different organizations and teams, each work package involved independently should provide the facilities to recover systems, applications and data of the central components they developed. The components installed locally, administrated by the biobank institutions, are the institution's responsibility.

Defect Management   An open-source issue tracking system has been provided and is still in use: Request Tracker (RT). Users are able to report defect detection and follow management of their reported issues.

Graphic User Interfaces (GUI)   Multiple software components have user interfaces that run independently. That is the case of the BBMRI-ERIC Directory, BBMRI-ERIC Locator, BBMRI-ERIC Negotiator, Connector and Harmonization Services (MDR). Ideally, for a seamless user experience, these user interfaces should all have the same look and feel. However, due to extensive collaborations with other projects and networks as well as an ongoing usability analysis in GBA, this harmonization of Graphical User Interfaces is currently (March 2019) an ongoing process.

Information Exchange   Information exchange between the different software components is done through REST and a JSON structure has been defined for that purpose. WP8 - Harmonization Services played an important role in the semantic and format translations of both aggregated and non-aggregated sample data interchange. The requirements for this purpose has been defined in a separate document, prepared by WP8. Local software components, such as the Connector, are not directly accessible from outside organizational boundaries. Instead, these local systems apply polling to contact central components for data exchange.

Localization / Internationalization   The software facilities supports English only, at least for this first running version. However, each of the individual components technically support multiple human languages, in particular the Harmonization Services (MDR), where the semantics is a main concern.

Metadata   Subject- and sample-related metadata have been defined in the scope of WP8 and are used by the BBMRI-ERIC Directory, BBMRI-ERIC Locator, Connector and BBMRI-ERIC Negotiator. The metadata items (description and its basic attributes) are accessible through a REST interface from the MDR.

Persistence   Each software component provides its own services to handle the reading and writing of stored data, according to its needs. The BBMRI-ERIC Locator (Searchbroker) uses its own database to store queries and related data.

Transaction Management   All system components rely on their underlying database management software for handling ACID (Atomicity, Consistency, Isolation, Durability) transactions.

6 Key abstractions

6.1 Overview

Key abstractions are the key concepts and abstractions that the system needs to handle. In this document, there is a focus on the BBMRI-ERIC Locator, the BBMRI-ERIC Negotiator and the integration with the BBMRI-ERIC Directory.
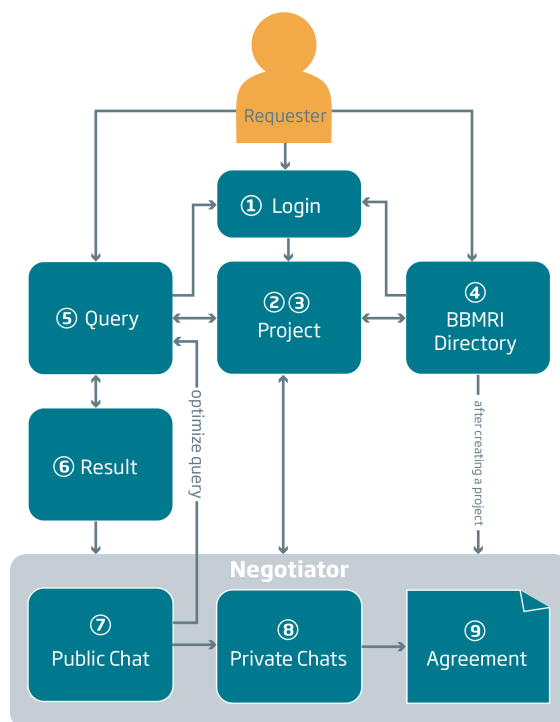
Figure 1: Request workflow
> 1) Login screen; 2) Create/edit a biobank request project; 3) See list of own projects; 4) Browse Directory of biobanks 5) Create (or refine) query (see search interface); 6) See results: biobanks with matching samples or data; 7) Chat with chosen biobanks; 8) Chat with one biobank each; 9) Final request / agreement with biobank(s).

Figure 1 depicts the schematic workflow of a biobank search request. There are different starting points. The requester may start with choosing biobanks worth considering in the BBMRI-ERIC Directory (4). The second possibility is to create a more sophisticated query (5) within the BBMRI-ERIC Locator. In both cases, before getting insight into detailed query results (6) (= possible matches per biobank) or starting the negotiating process (7), the requester has to log in (1) and to disclose contact information and basic data about the planned research project, e.g. organisation, goal, profit/non-profit, etc.

Figures 2 and 3 depict the main abstractions in discussion: BBMRI-ERIC Directory, BBMRI-ERIC Locator, BBMRI-ERIC Negotiator and Connector (and that there are harmonization services and central authentication/authorization).
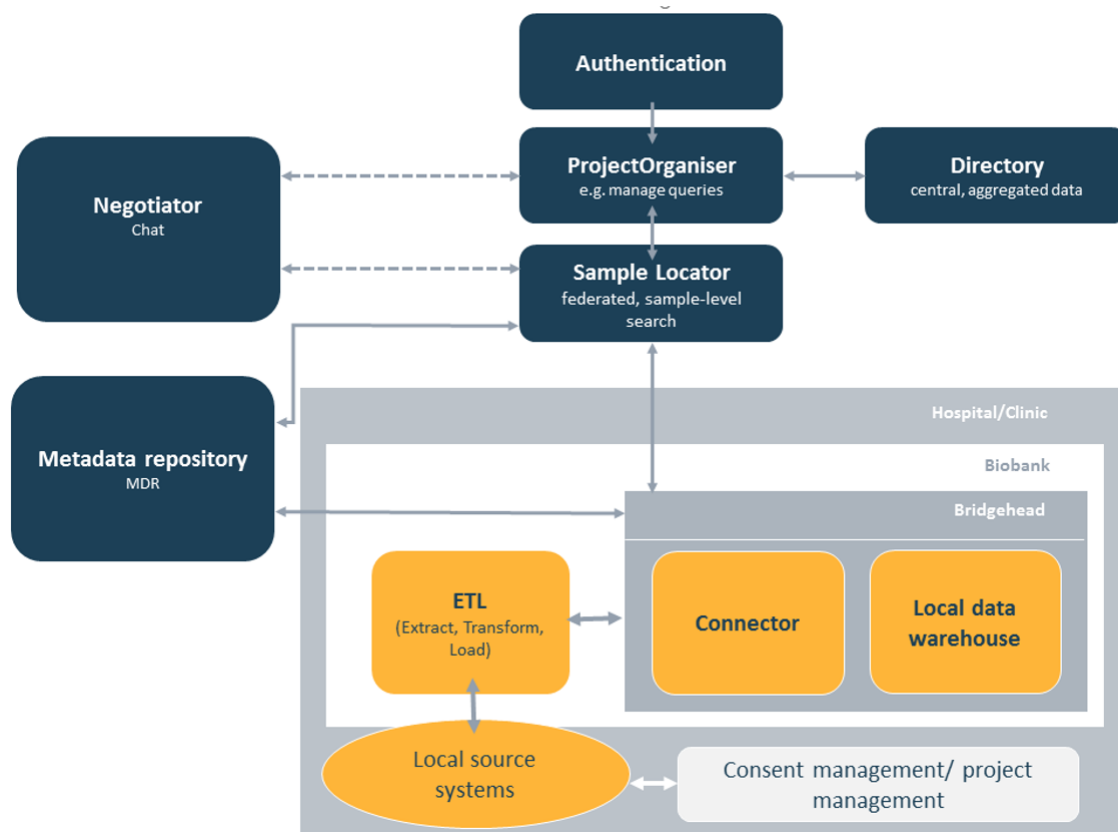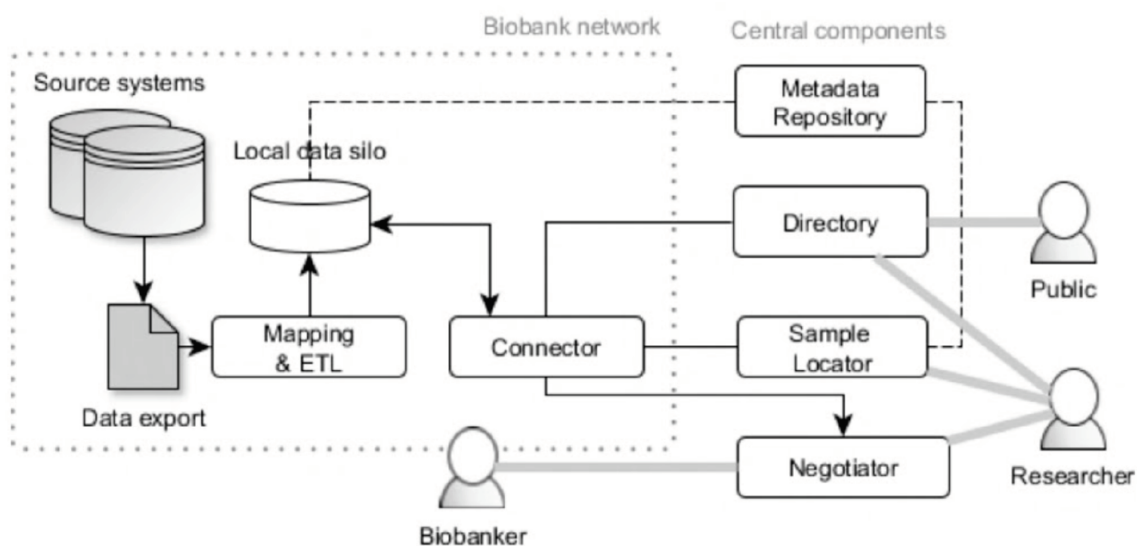
Figure 2: Software architecture.

Figure 3: Software architecture – alternative view.

The direction of the arrows represents the high-level interaction between the components, it is not a detailed message diagram or a technical representation. For instance, although an inquiry is defined at the Sample Locator and will get to the Connector, the Connector is actually polling the Sample Locator for inquiries. Thus the Locator-Connector arrow shows the logical information flow and not the actual implementation. This is done for security reasons and to overcome biobank networks restrictions on communication, to ease biobank integration with BBMRI-ERIC.

A more detailed explanation of the role of each component, and how it communicates with the remaining components, is provided in further sections of this document.

6.2 Metadata Repository (MDR)

The Metadata Repository applications, used for data harmonization, include a Graphical User Interface and REST interface. This component holds data that is commonly known by the term metadata, which can simply be understood as "data about data". It includes the definition, detailed description, validations and types of data used to define samples and correlated information in the scope of BBMRI-ERIC (illustrated in Figure 4). As an example, "Surgery date" is a data element which could be described as "the day of the month and year of a surgical procedure', has the type technically named as "Date" and can be validated by the format "dd.mm.yyyy".
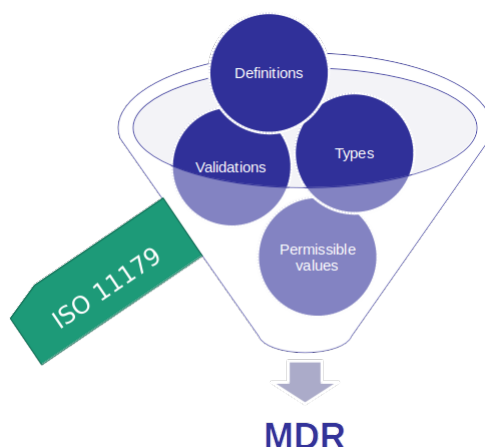
Figure 4: Metadata Repository (MDR).

The MDR follows roughly the standard ISO/IEC 11179, Information Technology – Metadata registries (MDR), especially on the "Registry metamodel and basic attributes", which describes what kind of metadata is needed and the structure of metadata registries. It allows storing metadata and enables consumers to retrieve it. All kinds of data can be defined formally by annotating appropriate metadata and these formal definitions are made broadly available in a central MDR.

As all metadata is defined in the MDR and multiple components access it, implementing cache on the MDR consumers will be crucial. Caches on the Directory, Negotiator, Connector and Sample Locator will highly increase the performance in these components and also in the MDR. It will also enable other systems to

continue running flawlessly even if there is a temporary failure in the MDR. The interfaces between the different components are implemented using the Representational State Transfer (REST) paradigm.

The MDR supports versioning. Every element in every repository is versioned, with full history preserved.

To make it even simpler to integrate with the MDR, a Java library has been created to query for data elements naming, definition and validation information. The MDRClient, as it is called, can be reused in Java projects to get sets of data that describe and give information about other data (i.e. metadata), from the MDR, through REST calls. Applications based on Java Server Faces (JSF) can also use MDRFaces - a JSF library which eases the user interface design of web applications where the respective data model relies on well-defined data elements. An example of the MDRFaces user interface is shown in Figure 5. Especially in case of systems for electronic data capturing, where the necessary data model is not known beforehand and can even vary over time, the user interface has to be easily adjustable. This often means the user instead of the developer designs the various forms for data entry and therefore an easy to use mechanism has to be provided. By using Samply.MDRFaces the developer can focus on the implementation of that mechanism, e.g. some editor component, but does not have to cope with the visualization of every single data element as for that is taken care of automatically.

Figure 5: Rendering a form using MDRFaces and metadata defined in the MDR.

6.3 Semantics and Metadata Repository (MDR)

The MDR is a central IT component within the software architecture and serves to support the data harmonisation process. Use of a common MDR is required to transform the heterogeneous vocabularies among the participating sites into a globally usable terminology. A semantic concept has been developed for the data harmonisation that describes the modules and processes for extraction, transformation and loading (extract, transform, load (ETL)). This will provide IT support for the semi-automatic loading and mapping of biobank data into the local data warehouse. Common data sets were created including existing terminologies. The data sets were subsequently integrated into a central MDR that was set up for this purpose. Local MDR instances are consequently no longer necessary, since locally applied data items can be described with their metadata in respective namespaces of the central MDR.

6.4 BBMRI-ERIC Directory

The BBMRI-ERIC Directory stores data about biobanks and their records in a central location. It enables browsing and querying aggregated information on biobanks and their collections by networking registries of national nodes and networks. It provides a single access point to the European biobank network and lays the basis for national and trans-national research consortia based on samples and data from various sites. Only aggregate-level/anonymous information are be shared via the BBMRI-ERIC Directory to ensure protection of biobank donor privacy and comply with ELSI requirements. This service enables interested parties, such as researchers, biobankers, funding agencies and policy makers, to easily query or browse and aggregate biobank and content information. In particular, the BBMRI-ERIC Directory enables researchers to easily identify biobanks that potentially have samples/data of interest.

The BBMRI-ERIC Directory is a networked service with a data structure in which each node (a biobank, a national node, or other types of entities; e.g., heads of biobank networks, representatives of non-BBMRI-ERIC countries) can share information to be aggregated centrally. Information sharing is based on an open data sharing protocol, which is simple so that new information sources (i.e., national nodes and networks) can easily integrate it into their software. In the initial BBMRI-ERIC Directory 1.0 in 2015, an LDAP protocol has been employed to share data between the existing catalogues of the national nodes and networks and the BBMRI-ERIC central Directory. Current versions of the BBMRI-ERIC Directory have transitioned to JSON data transferred over REST API. In addition, a toolkit will be provided to aid nodes and consortia to add their data to the Directory Service, if the data is not available via Connector.

In the 2016–2019 period, the BBMRI-ERIC Directory has received numerous updates namely in the web-based user interface, to make it more approachable to non-technical users. Currently it sports two different web interfaces: the default simplified one (shown in Figure 6) focusing on browsing and querying the most commonly sought for parameters, and an advanced once allowing to query all parts of the data model. The data is also available via REST/JSON API for public automated access. The data model has been also continuously improved and enchriched (e.g., inclusion of quality-related data, certifications, and information on particular available data categories such as DICOM® imaging), with particular attention to keeping backward compatibility.

6.5 BBMRI-ERIC Locator

The BBMRI-ERIC Locator provides a federated search mechanism for samples and sample-related data in the connected biobanks – including, for instance, clinical data such as diagnoses or therapies. The BBMRI-ERIC Locator is integrated with information model and terminology mapping tools, developed by WP8, in order to support the heterogeneity of the European biobank data structures.

This approach complements the BBMRI-ERIC Directory, which is a centralized, catalogue-like solution with aggregate-level data and flat data model (preventing answering full AND questions at the sample/individual level), by enabling the processing of requests on a sample-based level and giving the biobanks full insight into individual research projects and data requests.

The federated search concept supports specifying search queries based on items from the MDR and the asynchronous interaction with the participating biobanks. Along with the query, the request includes a

Figure 6: Screenshot of BBMRI-ERIC Directory user interface.

description of the research (or clarifying why the user is searching for specific samples) and the contact information of the inquiring partners, which is made available to all participating biobanks.

There will be a two-step approach: In the beginning of the search process, the requester has to fill in the most important data of the research project, like contact information, goal and short description of the project and the information whether the project is profit or non-profit. For specific negotiations, the requester has to add further information.

The request interface of the local biobanks, the Connector, retrieves the query locally and returns the results. Sample-level results, together with the research description and the inquirer's contact information, will be presented to the local data owner before the start of the negotiation process.

Figure 7: BBMRI-ERIC Locator (Searchbroker) screenshot

6.6 Connector

The Connector is a software component, jointly developed by GBA, BBMRI-ERIC and the German Cancer Consortium (DKTK), which runs on the biobank network (at national, regional or local level) and is under the control of the network administrator. It includes a software interface, which is able to retrieve research inquiries from the BBMRI-ERIC Locator, inquires the local data warehouse (pseudonymised data persisted on an accessible database with a BBMRI-ERIC predefined data structure) and enables data owners to manage research requests.

An included graphical user interface (GUI) allows the data owner to visualize the research description and the inquirer's contact information. Moreover, the Connector can automatically display the results of the search request, so that the data owner can see exactly what sample and/or data is being requested. The response process per default is automated. In the future, a so-called "reply-agent" might allow biobanks to configure which requests to answer automatically (with the number of potential matches), which ones have to be checked manually and which ones to blocked altogether.

The GUI also allows to initiate the negotiation (see BBMRI-ERIC Negotiator) with the researcher.

All data managed by the Connector is already pseudonymised: No identifiable data is managed or shared. The Connector has a REST interface with the local biobanks that supplies queries (based on metadata) and receives the results (sample data). The REST interface description has been made available to ease the integration with biobank management systems.

6.7 BBMRI-ERIC Negotiator

The BBMRI-ERIC Negotiator is a component that moderates the data requesting and access process. It is a communication platform between researchers and biobankers regarding sample requests.

Considering the sample search from the BBMRI-ERIC Locator, a request triggers a data sharing negotiation process between the researcher and the participating biobank(s). The BBMRI-ERIC Negotiator enables the refinement, based on discussion, of the search requests for each specific biobank. Biobank sample-level data is shared only through the BBMRI-ERIC Negotiator.

The BBMRI-ERIC Negotiator appears similar to an Internet chat room, or message board – an online discussion site where researchers and biobankers can hold conversations in the form of posted messages. There is a public chat room, shared among all biobanks chosen by the researcher, to discuss questions regarding the query itself. Additionally, there is a private chat room for the researcher and the biobanker only to support bilateral negotiation of the terms of sample and data exchange.

A researcher, through the BBMRI-ERIC Directory or the BBMRI-ERIC Locator, is able to pre-filter a list of biobanks based on the type of samples they provide. Afterwards, the researcher can proceed to the negotiation with these biobanks, on the BBMRI-ERIC Negotiator. This way the researchers who search for aggregated data in the BBMRI-ERIC Directory can easily discover, for instance, which participating biobanks have the samples or data needed for a particular research project.

Figure 8: Screenshot of BBMRI-ERIC Negotiator user interface.

7 Layers or Architectural Framework

There is an ETL process that will take place when the data goes from the Biobank to the local data warehouse, and that involves data pseudonymisation. The Connector, although developed and distributed centrally, is made to run on each biobank's own network (local environment) and it is offered to be operated under the control of the biobanks. This will ensure that no data is shared without the permission of the data owner.

A central user/group authentication and authorisation has been implemented to centrally manage permissions and allow users to login to all applications using a single sign-on (SSO). As a result, the navigation through the different independent systems is transparent for the users and the system integration is secure.

The Harmonization Services play an important role between the central components and the local components, namely on the requests created on the BBMRI-ERIC Locator and the sharing of data and communication through the BBMRI-ERIC Negotiator. The data harmonization tools will enable, through the Metadata Repository, the structured communication between the different systems and heterogeneous biobank data structures. The BBMRI-ERIC metadata definition has to be mapped to the local data structures, so that the integration between the connector and the biobank information management system (BIMS) can be automated. The section MDR contains a description of the MDR and the section Federated Search (BBMRI-ERIC Locator centric) includes an overview of the data flow on the federated search.

8 Architectural views

The Architecture can be represented from a variety of viewpoints, all of which can be combined to create a holistic view of the system. Each architectural view addresses some specific set of concerns.

8.1 Federated Search (BBMRI-ERIC Locator centric)

There are multiple database management systems and tools that biobanks use and it is also common that biobanks create proprietary software for their activities. Despite the development of standards such as ICD-10 or SNOMED, biobanks run on different data structures and systems. Even when dealing with the same kind of data, biobanks persist their clinical data on different structures, naming and units. For successful research across data in multiple biobanks, an efficient harmonization platform is needed for the biobanks to be able to communicate on the same semantic platform.

On the other hand, biobanks are often intransigent, incapable or not allowed to give away their data to be managed by another entity that persists clinical data, centrally (not in the biobank private network), from different biobanks under the same platform and data structure. In these cases, the biobanks demand ownership on the data and control over privacy, although they are usually willing to share pseudonymised data on specific samples and for individual research projects. A federated architecture is therefore recommended as it allows interoperability, harmonization and information sharing between biobanks while maintaining their autonomy. Figure 9 illustrates the architecture needed to implement a federated search.
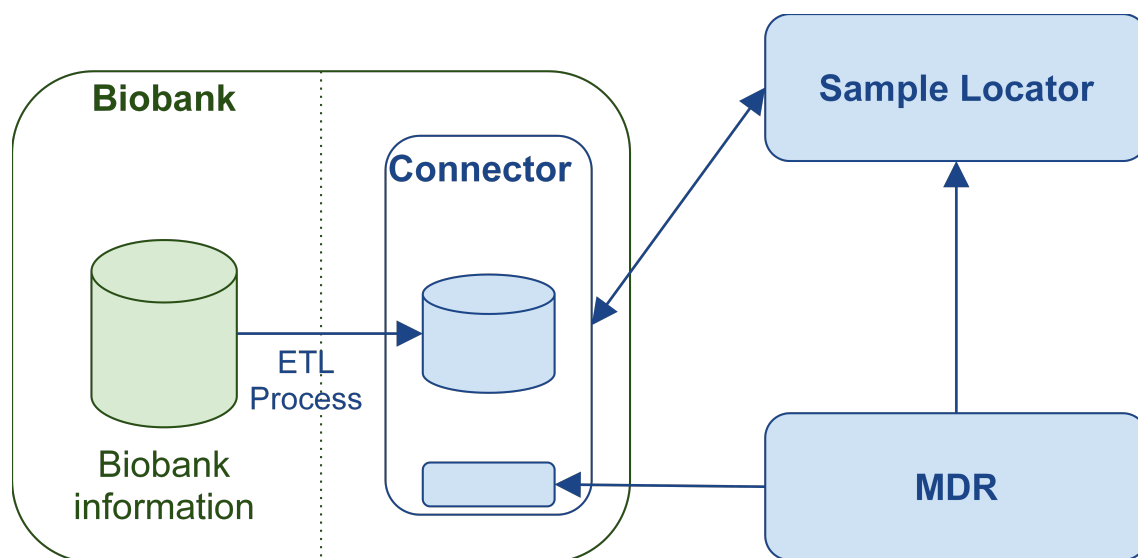
Figure 9: Integration of a Biobank with the Connector, BBMRI-ERIC Locator and MDR.

In this section we represent the BBMRI-ERIC Common Service IT federated biobank architecture.

Figures 10 and 11 depict the process view, which describes how the system is structured as a set of elements that have behavior and interactions. In particular, the interaction between the Researcher, BBMRI-ERIC Locator, Authentication, BBMRI-ERIC Negotiator and the local Biobank (Data owner) is illustrated.
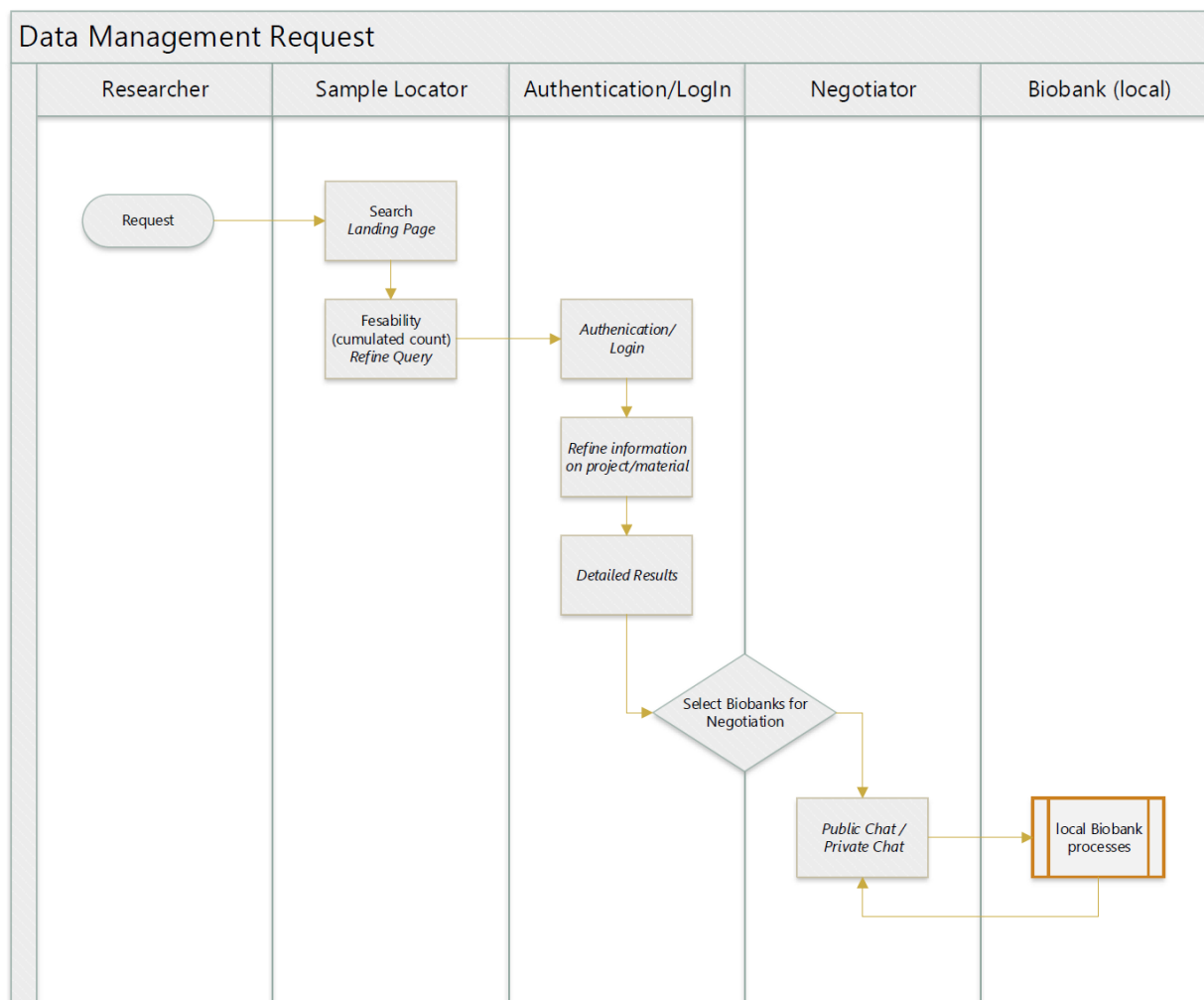
Figure 10: Overview of the federated research request.

8.2 BBMRI-ERIC Directory Centric Negotiation

The directory centric negotiation process starts with a researcher identifying potentially useful samples in the BBMRI-ERIC Directory. From the BBMRI-ERIC Directory user interface, the researcher starts a new request and is redirected to the BBMRI-ERIC Negotiator, where he or she can see the used query and add a free text description of the request. All biobankers with potentially useful samples are notified and can join the negotiating process in the BBMRI-ERIC Negotiator. Subsequent changes to the query or free text are possible, as we expect the negotiation process to include some clarification of the exact nature of the request. The process concludes when the researcher and one or more biobankers decide to collaborate and move on to the next steps needed for physical access to the samples.
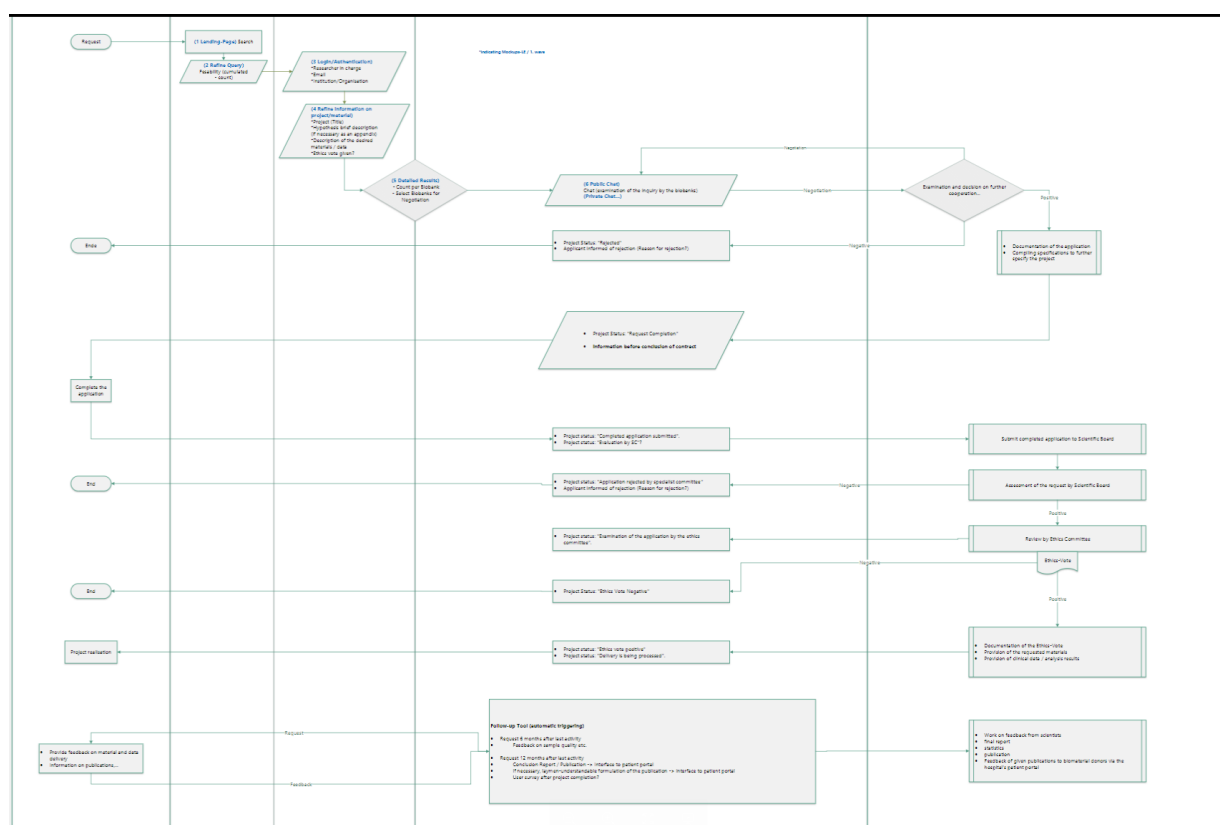
Figure 11: Detailed view of the federated research request.
Columns from left to right: Researcher, BBMRI-ERIC Locator, Authentication, ProjectOrganiser/BBMRI-ERIC Negotiator, Local Biobank A scalable version is made available electronically to project participants.

The whole process has been modelled and implemented in detail, and the resulting sequence diagram is available in the project documentation. The behaviour of the BBMRI-ERIC Negotiator in this process is described in the ADOPT Deliverable D4.4.

8.3 Colon Cancer Data Gathering

In the scope of ADOPT WP2, there the aim was to collect Europe-wide cohort of at least 10 000 cases of colorecatal cancer. This goal has been achieved and the results have been reported as ADOPT Deliverables D2.4 and D2.7. As a technical solution for this use case, we delivered the architecture defined below, based on already existing, tested and running components of the OSSE registry software.[2]
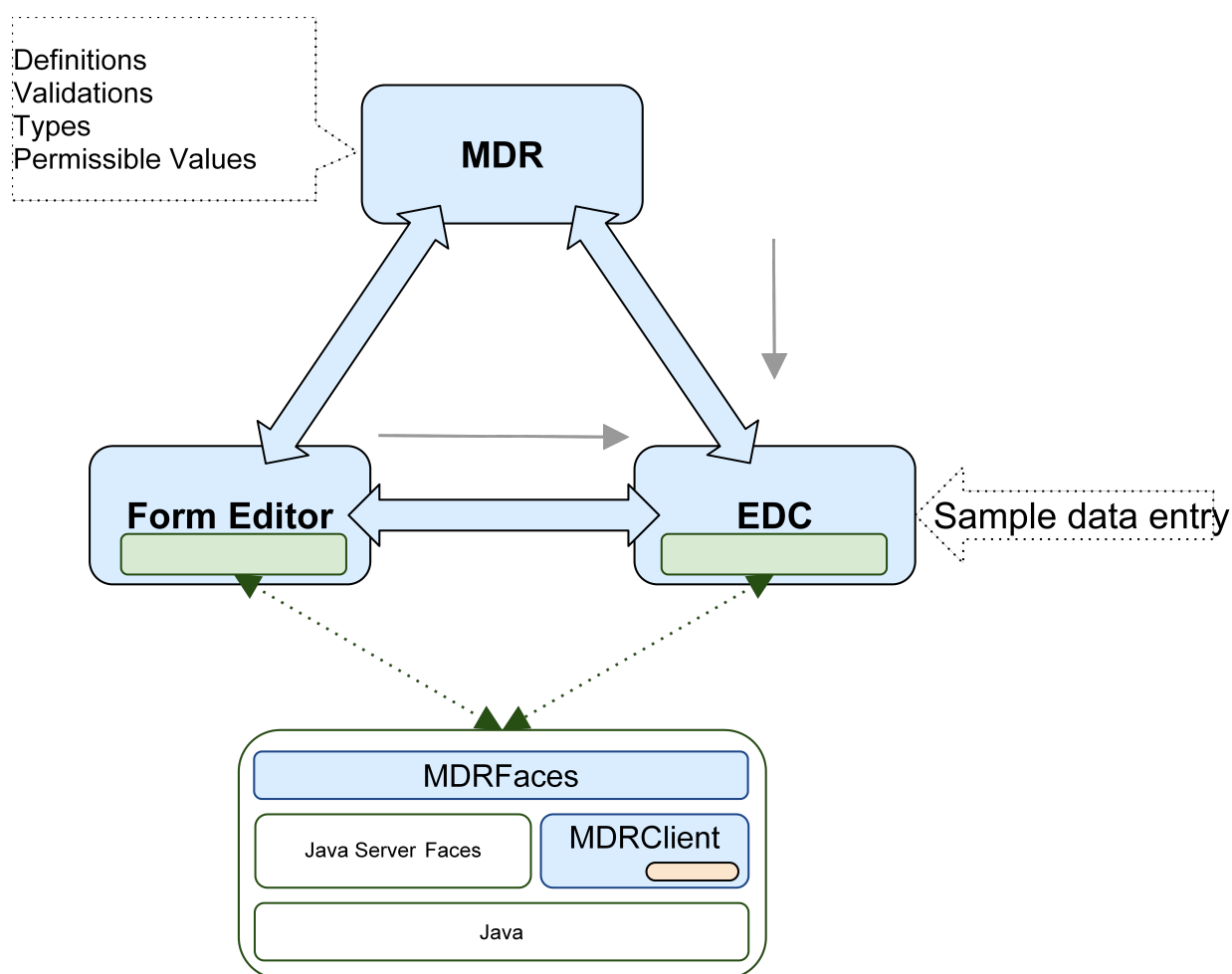
Figure 12: Architecture of the Colon Cancer Data Collection infrastructure.

The architecture is visualized in Figure 12 Architecture of the Colon Cancer Data Collection infrastructure. The MDR, the MDRFaces and MDRClient libraries are described in the Section 6.2. An example view of the MDR content is shown in Figure 13. Additionally, a form editor has been created, where users can

---

[2] https://www.osse-register.de

create and editor form based on pre-defined data elements (from the MDR). This enables reusability of both forms and data entities and assures data harmonisation throughout the electronic data capture systems that rely on forms built with this tool. The easy-to-use user interfaces and the integration with the metadata repository make it possible to effortlessly create, edit and use complex clinical forms without software development knowledge.
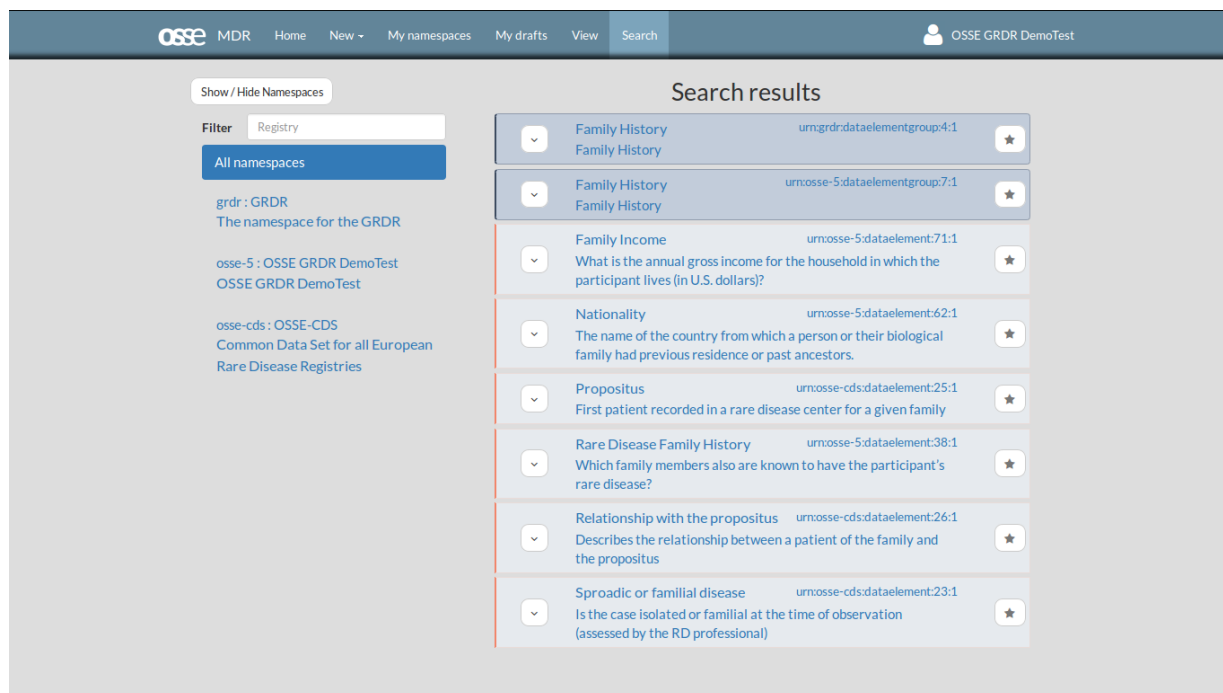
Figure 13: The MDR UI – Searching for a data element.

For the colon cancer data gathering use case, the metadata entered in the MDR are based on the ADOPT Deliverable D2.4.
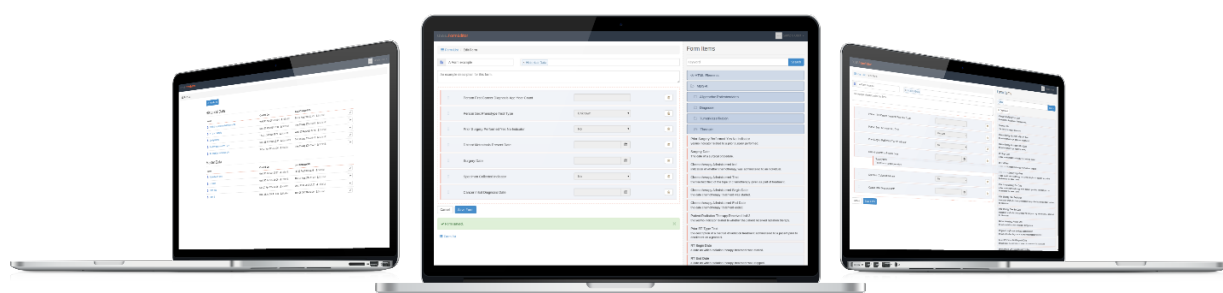
Figure 14: Form Editor – Form list, form detail and form metadata editing.

electronic data capture (EDC) (impression shown in Figure 14) is an electronic data capture component with a user interface where data is entered manually in forms. These forms are designed on the "Form Editor", based on the metadata from the MDR. For the semi-automated data import, a REST interface has been developed. The structure of the POST messages include the MDR data item IDs (metadata keys) and the values, which are the content of the colon cancer cases.

8.4 Authentication and Authorization Infrastructure (AAI)

The BBMRI-ERIC Locator, BBMRI-ERIC Negotiator and MDR provide services that need user authentication. For this purpose, BBMRI-ERIC Common Service IT provides single sign-on services, abstracting various login and identity services into a single API including public APIs like Edugain and Google's OAuth 2.0 API.

OpenID Connect, which is a simple identity layer on top of the OAuth 2.0 protocol, is used. In this way, an integration with the Samply.Auth service within GBA has been realised. It allows clients to verify the identity of the end-user based on the authentication performed by an Authorization Server, as well as to obtain basic profile information about the end user in an interoperable and REST-like manner. A sequence diagram of the authentication process is shown in Figure 15.

Figure 15: Authentication and Authorization Infrastructure

9 Security and privacy considerations

Security and privacy are a major concern for both ADOPT and BBMRI-ERIC. An extensive risk analysis and system design including relevant security requirements and a list of measures to address them is published in a separate document - ADOPT Deliverable D3.2, Security and privacy architecture.

ADOPT uses a combination of the Spoofing, Tampering, Repudiation, Information Disclosure, Denial of service, Elevation of privilege (STRIDE) and Linkability, Identif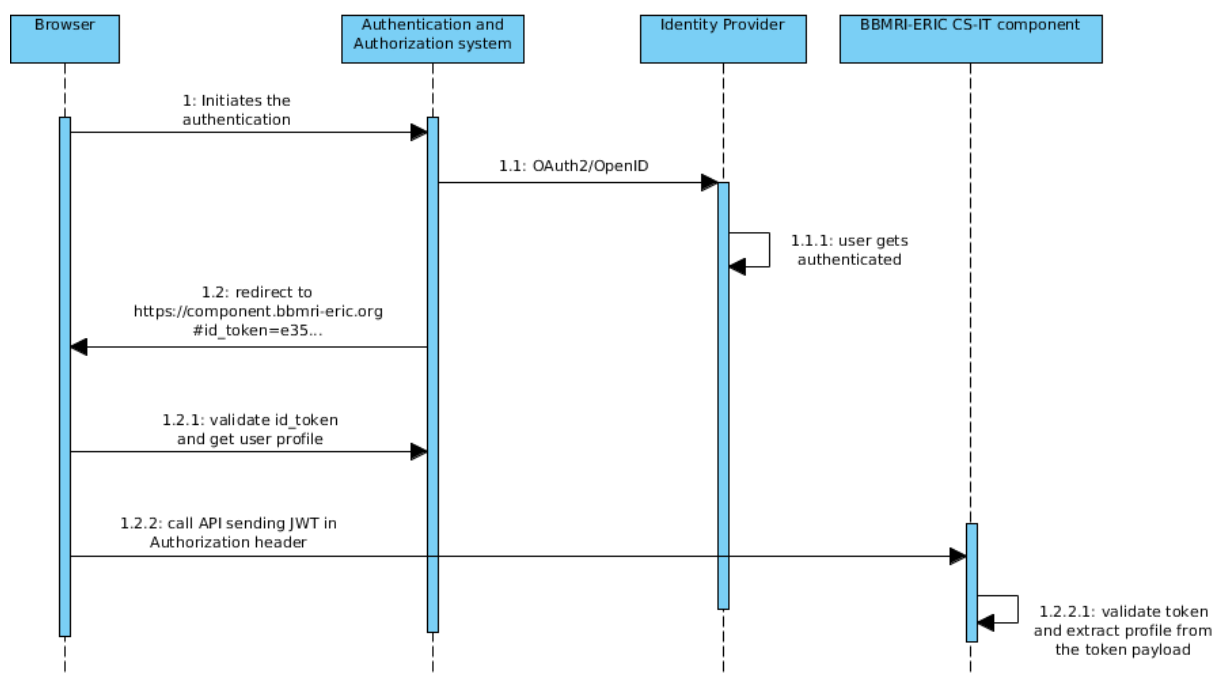iability, Non-repudiation, Detectability, Disclosure of information, Content Unawareness, Policy and consent non-compliance (LINDDUN) methodologies to ensure a secure IT platform. It identifies a number of potential risks and implements adequate mechanisms for mitigation and protection.

Implementing the architecture described in this document and the measures from the Security and privacy architecture document ensures that all components run by BBMRI-ERIC are protected at a state of the art level. As the resulting platform has a decentralized architecture, there are software components, which are run by the biobanks (e.g. the Connectors), over which BBMRI-ERIC has no direct control. However, we consider the associated risk to be low, and it is further mitigated by several security measures.

The most pertinent threat related to low client-side security would be a compromised Connector, due to insufficient security at the biobank at which it is installed. We see the probability for this scenario as low to moderate. Biobanks are usually part of clinical organizations and thus subject to stringent security requirements. Also, there is no automated process for a biobank to join, and the manual process allows BBMRI-ERIC to ascertain, that the new members are bona fide biobanks and provides an opportunity to deny participation to a biobank, should there be doubt that said biobank is capable of protecting its systems against unauthorized access, as specified in measures Me-28-2, 35-5 and 41-4 . Additionally, the probability of account abuse at the biobank is prevented by implementing a LoA (Level of Assurance) ≥ 2 for authentication of the biobanker, as defined in measure Me-41-2.

The impact of a hypothetical compromised Connector is low for the platform as a whole. A Connector only communicates to the central components, not to other Connectors. Since the central components do not store privacy-sensitive data, and cannot request privacy-sensitive data from other Connectors, there is no attack scenario in which a compromised Connector can be used to gain unauthorized access to privacy-sensitive data from other biobanks. Nevertheless, the Connectors and central components use best practices for protection against typical attack vectors such as SQL injections and buffer overflow attacks.

In the realm of non-malicious security risks, the most likely scenario includes biobankers implementing insufficient pseudonymization or anonymization procedures, which would result in privacy sensitive data being shared over the BBMRI-ERIC Common Service IT platform. This is only possible in some scenarios, in which a biobank has configured their Connector to send non-aggregated responses to queries from the Sample Locator. BBMRI-ERIC mitigates this risk by monitoring the data being exchanged and alarming the biobank if the data does not exhibit the appropriate level of anonymization or pseudonymization. If applicable, automatic filters can be implemented that recognize and replace commonly used insecure pseudonyms such as state-level personal identification numbers.

The security of the central components is dependent on proper maintenance. BBMRI-ERIC is operating on a best-effort principle for the duration of ADOPT. This means that security relevant events are handled with priority, but no guarantee can be made for the speed of response, especially outside of business hours.

9.1 German and EU data protection

In Germany, GBA has created a data protection concept that has received a positive vote by the TMF working group for data protection. This verdict is recognized by all data protection officers of the German federal states and thus greatly facilitates the review by the data protection officers at the participating biobank sites. Currently, all German biobank sites have received approval to fill their bridgeheads with clinical data.

10 Recent Achievements

The agile development process, the work tools and communication rules have been successfully established.

At the end of 2017, the first German biobanks entered the European biobank registry at https://directory.bbmri-eric.eu in agreement with BBMRI-ERIC.

Discussions led to a refined version of the biobanking IT framework as well as to a peer-reviewed journal publication in "Der Pathologe" [8]. The framework concept was also published (Ebert et al., 2018). At the end of September 2018, the first live demonstration of the federated search function was presented across 8 of the 11 locations at several meetings, inter alia the Scientific and Ethical Advisory Board Meeting in Berlin (09/2018) and BBMRI-ERIC's Management Committee meeting (10/2018) in London. At this date, all eleven GBA biobank sites and one Czech biobank (Masaryk Memorial Cancer Institute[3] in Brno) provide real data via their bridgeheads, which are linked to the pilot search function of a preliminary GUI. Further bridgeheads across Europe are going to be added continuously as part of a (yet to be signed) Common Service IT contract between BBMRI-ERIC headquarters and DKFZ. Latest steps have taken the central IT components towards open source development to enable further dissemination of the product. The software is available as Docker containers, as the previously developed Windows Installer has proven to be a much more complicated option for the deployment.

The stakeholder and demand analyses for sample and data requests as well as the definition of the sample and data request processes have been carried out and published. The Sample Locator (SearchBroker) has been successfully set up. The local Connectors (Bridgeheads) are also set up in all above mentioned biobanks and are continuously synchronized with current data. The IT solution for the ProjectOrganiser tool and the follow-up management are closely coordinated not only between BBMRI-ERIC and GBA, but prospectively also with the German Medical Informatics Initiative and the German Centers for Health Research.

The German data protection concept was positively evaluated by the TMF Working Group on Data Protection. In the meantime, all biobank sites have also received approval from the local data protectors to fill their bridgeheads with clinical data.

The central IT component of the MDR within the GBA architecture supports the data harmonization process. A semantic concept was developed that describes the modules and processes for ETL. Using existing terminologies, common data sets were created. The data sets are integrated in a central MDR set up for this purpose. For the current version of the software components and the creation of harmonized datasets, the IT teams of the Austrian and Finnish BBMRI-ERIC nodes were specifically involved. This includes the latest version of the MIABIS description. In addition, GBN developed in cooperation with BBMRI.nl a German Biobank Directory within the Dutch Molgenis database structure. This German instance is synchronized daily with the BBMRI-ERIC directory. This cooperation will be continued in order to improve the quality of the directory data and the user-friendliness of the interface.

---

[3] https://www.mou.cz/

11 Comparison of planned and actual project status

The advances in the definition of HL7 FHIR® resources and cross-project discussions led to a new evaluation of HL7 FHIR® as a potentially better store. This milestone has been achieved in January 2019. We decided to go only for the BBMRI-ERIC directory to ensure Europe-wide visibility of the German biobanks and against a separate national registry. We have set up the German Biobank Directory based on the Molgenis solution which is synchronised daily with BBMRI-ERIC's directory. Because of technical delays in the collaboration with Molgenis, we could only start promoting the directory in summer 2018. To date, 24 German biobanks have registered their collections in the directory including all GBA biobanks.

The German and EU data protection concept for the project was finalised on time. The 'Tool for distributed sample search and project mediation' has been developed and implemented. Some sites still need to upload data since local data protection authorities caused delays. A first prototype of the Negotiator (for project mediation) is part of this tool and has been developed and is already in use. The Negotiator has been connected to the Sample Locator to enable private or public chat functions for mediation of the researcher with the chosen biobanks. As soon as all modules are integrated we will undergo a thorough evaluation process.

12 Future

12.1 BBMRI-ERIC after ADOPT

After ADOPT, BBMRI-ERIC should take care to operate the platform properly and provide the necessary maintenance and user support. The existing security policies should be strictly followed, to ensure that no new risks are introduced. An emphasis should be placed on the proper authentication of biobanks and their employees, so the risk of unauthorized access and abuse of user accounts is minimized. Additionally, a training for biobankers can increase the chance that they will operate the Connectors and follow the sample request process in a secure and efficient manner. The provision of such training is conditional on sufficient resources for creating appropriate training material and providing adequate user support.

BBMRI-ERIC should evaluate the need for guaranteed reaction times in the case of a recognized security threat after ADOPT is finalized, and decide whether it should pursue the acquisition of resources to establish a service with guarantees beyond best effort.

12.2 Expand the IT infrastructure

(Networked) IT systems for biobanks will be of central importance to two developments: (1) due to the rapid advances in the molecular subclassification of diseases, the compilation of sufficient case collectives will in future only be possible across sites, and (2) biobanks will increasingly be consulted as sources of (molecular and diagnostic) data on certain diseases. The latter point also includes the generation of research data from existing biomaterials as well as the return of research data to the biobanks. The outcome is a pool of very high-quality data that can sustainably accelerate biomedical research. As of 2019 the DKTK (German Consortium of Translational Oncology) uses a similar IT Network. One of the goals for the future is full interoperability between BBMRI-ERIC, GBA and DKTK. Work on this has already started by defining a HL7 FHIR® dataset that is suitable to be used in all projects.

12.3 Refine tools for biosamples and data searches

For biomedical research and the further development of precision medicine, it is essential for researchers to quickly be able to locate and use well-characterised, high-quality biomaterials and associated data. Systematic searches and queries to individual biobanks are time-consuming and often unsuccessful. Searches across all biobanks on the feasibility of a project can help in both academic and pharmaceutical or diagnostic research contexts. The first stage of the federated biobank search provides the basis for this: the tool developed and installed enables real-time searches using a combination of different parameters (currently age, gender, demographic data, ICD-10 code, material types, etc.). Initially, only the number of biomaterials available at different biobanks matching the search criteria are displayed within a matter of seconds. The second step in the search allows the researcher to contact the relevant biobanks, refine their query and ultimately gain access to samples and data. A tool ("Negotiator") has been developed to support and accelerate this process that can be used after a successful search to contact the biobanks.

The biobanks will form a well-organised governance structure which allows rapid access to biomaterials and their data. This will take the support of biomedical research and precision medicine to the next level.

Possible measures:

- Empower more biobanks to become members of the European IT network for federated sample and data search.
- Facilitate free access to the federated biobank search (first step; feasibility query) for all academic and non-academic researchers
- Integrate German biobanks into BBMRI-ERIC's European federated search

12.4 Position biobanks as data sources

High-quality biomaterials in combination with comprehensive clinical and phenotype information are extremely valuable and very limited in number. Hence biobanks will in future increasingly offer data rather than biosamples. This is already common practice in some European countries (Netherlands, England, Finland, Norway). A number of German biobanks have also begun pursuing this approach. The data available to biobanks can originate from one of two sources: (1) from the biobanks themselves if it is their own or an institutional collection of (molecular) data, or (2) from research projects for which biomaterials have been made available for trials. The biobanks' ability to provide information on the research data that is available must be facilitated. This can be achieved through the appropriate identification in the biobank information system, linking to the relevant data sources or, where appropriate, local storage in the biobanks. In the case of research queries, biobanks must not only be able to provide information on existing biomaterials, but also on clinical data available outside the biobank – depending on the respective query. Biobanks of population-based trials often collect both biomaterials and the data generated from these. It must be obligatory to report data (e.g. from genetic and molecular analyses conducted as part of the trials or by biobank users) back to the biobank. The implementing institution retains ownership of the data and can, with its consent, facilitate a new research project. The biobank thus acts as the data's "trustee". This in turn allows the biobank to link data sets (e.g. phenotype data and analysis data) from several data owners in a pseudonymised system. This is hugely advantageous for potential evaluation projects. A similar system for the recording of genetic and molecular data in biobanks as well as for the surrender of associated data to biobank users is desirable. In order to link clinical data and pure research data with biosamples and to process and make these available for research projects, biobanks must be readied accordingly. A future-oriented approach is needed in this respect so that biomedical research in Europe can keep pace with international developments. Particularly the funding bodies are called upon to facilitate the sustainable use of samples and data after funding ends by providing the corresponding guidelines for the use of funding.

Possible measures:

- Enable the linkage and/or return of research data to biobanks.
- Adapt funding guidelines to allow research data to be used once funding ends.
- Ready biobanks to link sample and research data and develop standards for the disclosure of data.

Bibliography

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

[2] L. Ebert, J. Habermann, P. Duhm-Harbeck, J. Ingenerf, H. Ulrich, B. Kroll, J. Kern, D. Tas, F. Stampe, S. Maqsood, F. Ückert, M. Ataian, R. Proynova, P. Holub, C. Schüttler, H. Prokosch, C. Knell, S. Sahr, T. Illig, M. Breu, R. Jahns, J. Linde, S. Nussbeck, C. Engels, M. Hummel, and M. Lablans. *Harmonizing European and German Biobank Request Workflows: BBMRI-ERIC and GBA – a Synergy!* Europe Biobank Week 2018. Antwerp.

[3] P. Holub, D. van Enckevort, R. Reihs, M. Swertz, H. Müller, A. DiezFraile, M. Hummel, and J.-E. Litton. *Consolidated Registry of BBMRI-ERIC Biobanks*. Sept. 2016. DOI: 10.5281/zenodo.273550. URL: https://doi.org/10.5281/zenodo.273550.

[4] P. Holub and C. S. IT. *Security and Privacy Architecture*. Sept. 2016. DOI: 10.5281/zenodo.159444. URL: https://doi.org/10.5281/zenodo.159444.

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

[6] D. Kadioglu, B. Breil, C. Knell, M. Lablans, S. Mate, D. Schlue, H. Serve, H. Storf, F. Ückert, T. Wagner, et al. "Samply. MDR-A Metadata Repository and Its Application in Various Research Networks." In: *Studies in health technology and informatics* 253 (2018), pp. 50–54.

[7] R. Merino-Martinez, D. v. E. Loreana Norlin, G. Anton, K. S. Simone Schuffenhauer, L. Mook, P. Holub, R. Bild, M. Swertz, and J.-E. Litton. "Towards Global Biobank Integration by Implementation of the Minimum Information About Biobank data Sharing (MIABIS 2.0 Core)". In: *Biopreservation and Biobanking* 14.4 (2016), pp. 298–306.

[8] C. Schüttler, N. Buschhuter, C. Dollinger, L. Ebert, M. Hummel, J. Linde, H. U. Prokosch, R. Proynova, and M. Lablans. "[Requirements for a cross-location biobank IT infrastructure : Survey of stakeholder input on the establishment of a biobank network of the German Biobank Alliance (GBA)]". In: *Pathologe* 39.4 (July 2018), pp. 289–296.