ADOPT BBMRI-ERIC
Grant Agreement no. 676550

DELIVERABLE REPORT

Deliverable no	D 3.4
Deliverable Title	Implementation of the BBMRI-ERIC architecture and database
Contractual delivery month	M12
Responsible Partner	DKFZ
Author(s)	Diogo Alexandre, Martin Lablans, Frank Ückert, Petr Holub, Michael Hummel, Morris Swertz, Rumyana Proynova, David van Enckevort, Jonathan Jetten

Architecture Notebook

BBMRI-ERIC CS-IT Architecture

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 BBMRI-ADOPT project is a part of this initiative and will create 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 which will be built or integrated into the system, and provides a brief overview of their function and an explanation of how they will support the project's goals.

Architectural goals and philosophy

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

- Aggregated data of samples from participating biobanks will be accessible in a central repository (Directory), grouped per collection. Collections in the Directory are based on the definition from MIABIS, with the additional refinements as outlined in the deliverable ADOPT 3.1, "Consolidated Registry of BBMRI-ERIC Biobanks" by D. van Enckevort.
    - Collections are containers for sample sets and/or datasets, with support for recursive creation of sub-collections (of arbitrary nite depth); 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 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. is 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 Sample Locator, which will allow fine-granular selection of suitable samples.
- To allow searching data at the non-aggregated sample level a federated search system (Sample Locator) will be implemented, in which data in answer to queries is only shared upon the data owner approval - "trust relationships", enabled by the data owner, will allow automated replies
- A Metadata Repository (MDR) will be implemented, following a generally accepted international standard to describe all data elements and their relationships within BBMRI CS IT
- It shall be technically possible to map metadata from the Metadata repository to data elements in every participating biobanking metadata
- Shared non-aggregated sample data is pseudonymised to ensure that participants are not identifiable

- Every service/component developed will be designed with privacy in mind and the developments shall respect the BBMRI-ERIC CS-IT Security & Privacy Requirements document

Assumptions and dependencies

In this section there is a list of the assumptions and dependencies that drive architectural decisions.

- Due to the dimension of the project and distribution of the development teams involved, the technologies used for each system component are relatively flexible, but the integration must be transparent to the users for all parts
- The technologies used shall be discussed, reviewed and approved within the CS-IT group and keeping a reasonably number of different technologies is desirable
- There will be periodic meetings between the teams involved in the different components' integration
- The Metadata Repository (MDR), from the Harmonization Services, shall store and enable consumers to retrieve metadata items, which include detailed description, data type and lists of permitted values (if applicable)

Architecturally significant requirements

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

- The integration of the Directory, Sample Locator, Connector, Negotiator and Harmonization services shall be able to exchange data, but none of the components shall become unworkable because of dependencies of another one
    - As a practical example, if the MDR would go down, the other components shall not crash because of it and the user should get proper feedback stating exactly what the issue is. In the specific case of the MDR being down, as we have the "mdrclient" - a java library, RESTful client that automatically caches MDR data - the components who reuse it automatically support cache and therefore it should still able to get the metadata. However, if some needed metadata is not cached, then the system shall specifically mention (to the user) that the MDR is momentarily not accessible.
- The different components shall be loosely coupled - all communication shall be done asynchronously and no requests or data shall be lost whenever one or more of these components fails

- Communication between the system components shall be done through RESTful or HTTP/JSON web services
- The Connectors poll the queries from the Sample Locator, due to the clinical restrictive firewalls
- All distributed data communication shall be done through HTTP over TLS (HTTPS)
- The user interfaces from the different system components shall be based on Bootstrap and on a common BBMRI-ERIC CS-IT Cascading Style Sheets (CSS) template (conforming to Bootstrap framework)
- The Metadata Repository (MDR) shall offer a REST interface for other system components connected to the BBMRI-ERIC CS-IT central authentication and authorization component
- The MDR shall follow the ISO 11179 definition of metadata items, particularly oriented by the “Registry metamodel and basic attributes”.
- The Connector will be running under the biobank’s organization control, in their demilitarized zone (DMZ), as it is known in computer security. It will not run on the same physical network of the biobank itself, i.e. the Connector will not be able to directly query the biobank.

Decisions, constraints, and justifications

There are a variety of factors that place constraints on the architecture being developed. These architectural constraints, combined with the requirements, will help defining the system architecture. Capturing these constraints will ease integration and may reduce risk, cost and duplication of solution elements.

- The BBMRI-ERIC CS-IT components will be web applications that are able to communicate through REST and REST like HTTP/JSON web services
- For a seamless user experience, the user interfaces shall all 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 it should be based on open source and on technologies that each team is familiar with
- Use of third-party software or a particular technology shall be communicated to the CS-IT director and CIO, who decide whether it should be allowed
- The software produced should be shared as open source or the IP should be with BBMRI-ERIC (conform the BBMRI-ERIC CS IT tender)

- Software code should be managed in version control (e.g. github) and have stable releases with stable external interfaces to ensure interfacing components depending on them to not break on a release (phasing out of outdated interfaces will be coordinated with BBMRI CS IT team leadership).
- There has to be a clear/documented hand over from development to operations, see transition between P3 and P4 as requested in the CS IT tender (even if this was implemented within a single team).
- In order to increase quality of the delivered product, the following reviews will be implemented:
    - Architectural review with focus on privacy & security by design.
    - Architectural review by ELSI group - focusing on overall workflow, privacy, informed consent management, confidentiality of projects and project proposals, etc.
    - Code review: use of Coverity service free for open-source software, cross-WP code reviews for major releases.

Architectural Mechanisms

Availability

The CS-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 shall continue running independently. This will be achieved by asynchronous communication between the components.

Disaster Recovery

Being a distributed system, developed by different organizations and teams, each work package involved shall independently 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

Upon defect detection, users shall be able to report it to the CS-IT. As a result, an issue tracking system shall be provided to the users to follow management of their reported issues.

Graphics

Multiple CS-IT components will have user interfaces that run independently. That is the case of the Directory, Sample Locator, Negotiator, Connector and Harmonization

Services. For a seamless user experience, the user interfaces shall all have the same look and feel, based on a Bootstrap template developed for BBMRI-ERIC.

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

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

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

Persistence
Each software component will provide its own services to handle the reading and writing of stored data, according to its needs.

Transaction Management
Each system component shall have mechanisms for handling ACID (Atomicity, Consistency, Isolation, Durability) transactions.

Key abstractions
Key abstractions are the key concepts and abstractions that system needs to handle. In this document there is a focus on the Sample Locator and Negotiator and integration with the Directory.

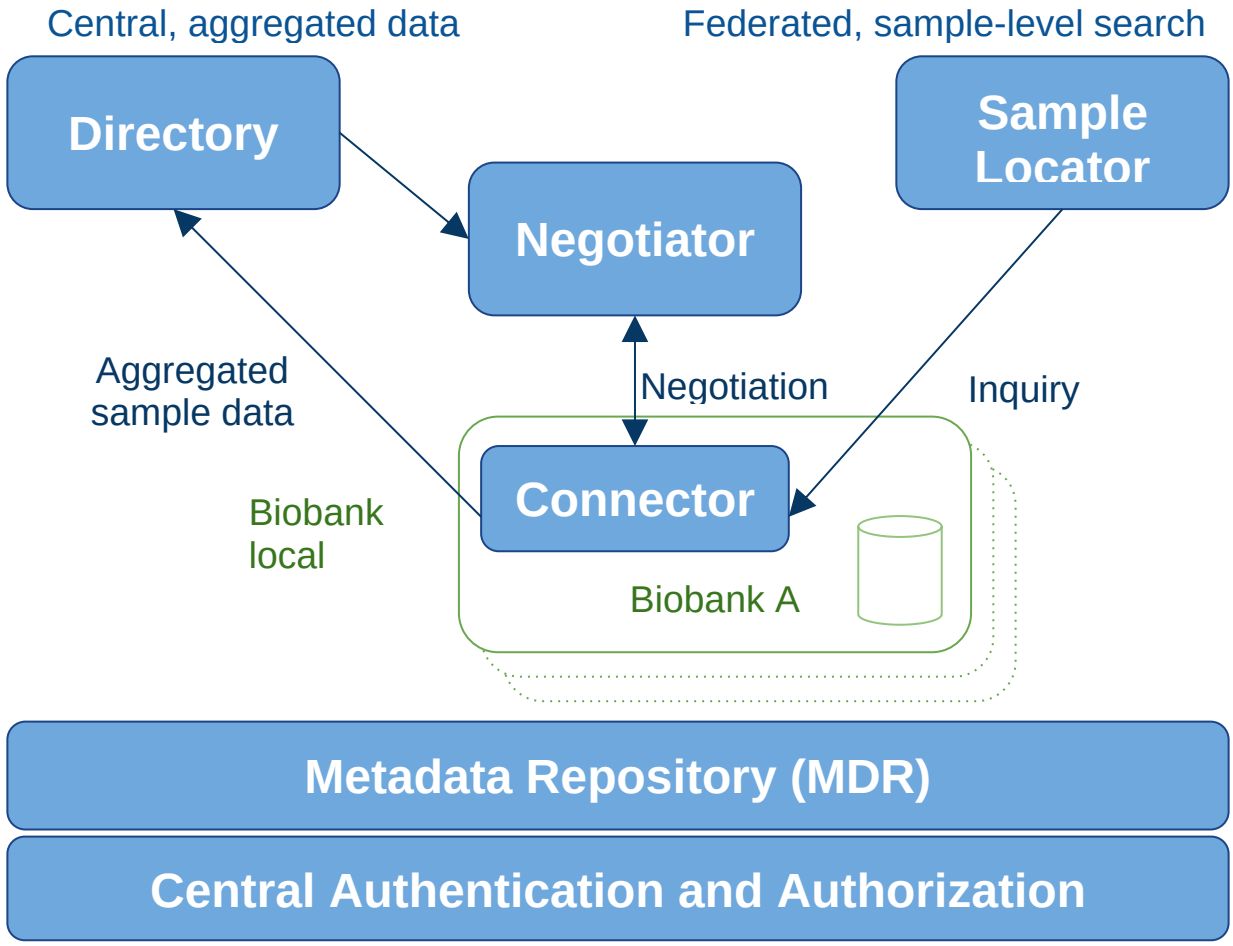

Figure 1Overview of WP1, WP2, WP8 and auxiliary components

Figure 1 depicts the main abstractions in discussion: Directory, Sample Locator, Negotiator and Connector (and that there are harmonization services and central authentication/authorization). In contrast, user interaction and business process diagrams are expressed in the Architectural Views section.

All the components on top (Directory, Sample Locator, Negotiator and Connector) get data from the harmonization services, which is expressed by having the harmonization services as a horizontal layer, supporting the components above it. Due to the intensive communication with the metadata repository (MDR) all the components that integrate with it shall support cache.

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. This is done for security reasons and to overcome biobank networks restrictions on communication, to ease biobank integration with BBMRI.

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 2). 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”.

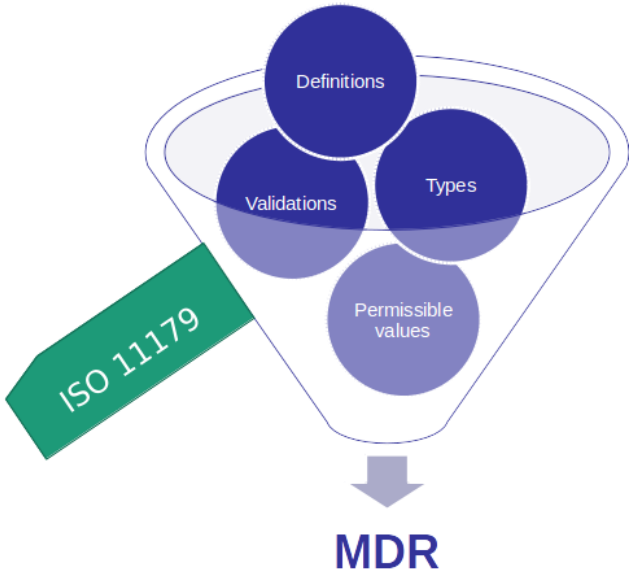

Figure 2 Metadata reposotory

Example
Surgery date
Definition
the day of the month and year of a surgical procedure
Type
Date
Validation
Dd.mm.yyyy

The MDR follows 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. Cache 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.

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[1],

[1] https://code.mitro.dkfz.de/projects/MDR/repos/samply.common.mdrclient/browse

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[2] - 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 3. 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 3 Rendering a form using MDRFaces and metadata defined in the MDR

Directory

The Directory Service (WP1) stores data about biobanks and their records in a central location. It enables the discovery of summarised information on biobanks and their collections by networking registries of national nodes and networks. It aims to provide a single access point to the European biobank network and to lay the basis for national and trans-national research consortia based on samples and data from various sites. Only summary level/anonymous information will be shared via the Directory to ensure

[2] https://code.mitro.dkfz.de/projects/MDR/repos/samply.common.mdrfaces/browse

protection of biobank donor privacy and comply with Ethical, Legal and Social Implications (ELSI). This service enables interested parties, such as researchers, biobankers, funding agencies and policy makers, to easily find, browse and aggregate biobank and content information. In particular, the Directory should enable 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 will be based on an open source data sharing protocol, which will be made simple so that new information sources (i.e. national nodes and networks) can easily integrate it into their software. In the pilot, a LDAP protocol has been employed to share summary level information between the existing catalogues of the national nodes and networks and the BBMRI-ERIC central Directory. This system is based on LDAP or REST, which can be easily implemented by local IT departments. 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.

Sample Locator

The Sample Locator (WP2) aims at finding samples and sample-related data (possibly including, for instance, clinical & imaging data, phenotypic data in the broad sense, etc.) from the biobanks through a federated search mechanism. The Sample Locator will be 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 (WP1), which is a catalogue-like solution with summary-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 control over individual data requests.
The federated search concept supports specifying search queries based on items from the metadata repository (WP8) and the asynchronous interaction with the participating biobanks. Along with the query, the request can include a 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. This can be supplied in the beginning of the search process, or also added later if some other steps (such as signing a confidentiality agreement) are needed first.
The request interface of the local biobanks, the connector, retrieves the query and the results, locally. Sample-level results, together with the research description and the inquirer's contact information, are presented to the data owner, who decides what to reply to each request.

Connector

The Connector is a software component, supplied by BBMRI-ERIC CS-IT, 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 Sample Locator, inquires local data sources (either through its API that can be used by local BIMS or on pseudonymised data persisted on an accessible database with a BBMRI-ERIC predefined data structure) and enables data owners to manage research requests.
An included user interface (UI) 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 data is being requested. The response process can be automated for trusted requesters (e.g. share all request data with Researcher A), if the local data/biobank owner so desires. However, no data is shared without the explicit consent of the data owner. The UI also allows to initiate the negotiation (see 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 gets the results (sample data). The REST interface description will be made publicly available to ease the integration with biobank management systems. There are plans to develop tools to help biobanks map BBMRI-ERIC metadata to local data structures.

Negotiator

The 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 Sample Locator, a request triggers a data sharing negotiation process between the researcher and the participating biobanks. The Negotiator enables the refinement, based on discussion, of the search requests for each specific biobank. Biobank sample-level data is shared only through the Negotiator. In the case the data owner has previously given access on some specific sample level data to the researcher (trust relationship), the data would be shared automatically through the Negotiator.
Additionally, the initial version of the Negotiator will serve as an Internet forum, or message board - an online discussion site where researchers and biobankers can hold conversations in the form of posted messages. A researcher, through the Directory, 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 Negotiator – see Directory Centric Negotiation section. This way the researchers who search for aggregated data in the Directory can easily discover, for instance, which participating biobanks have the samples or data needed for a particular research.

Layers or Architectural Framework

Figure 4 depicts the structure and decomposition of the CS-IT - considering WP1 and WP2 - into layers. It is not meant to suggest a layered technology (which it is not), but rather depicting another perspective from the system to show what are the local and central components, how they are integrated and what components depend on others.

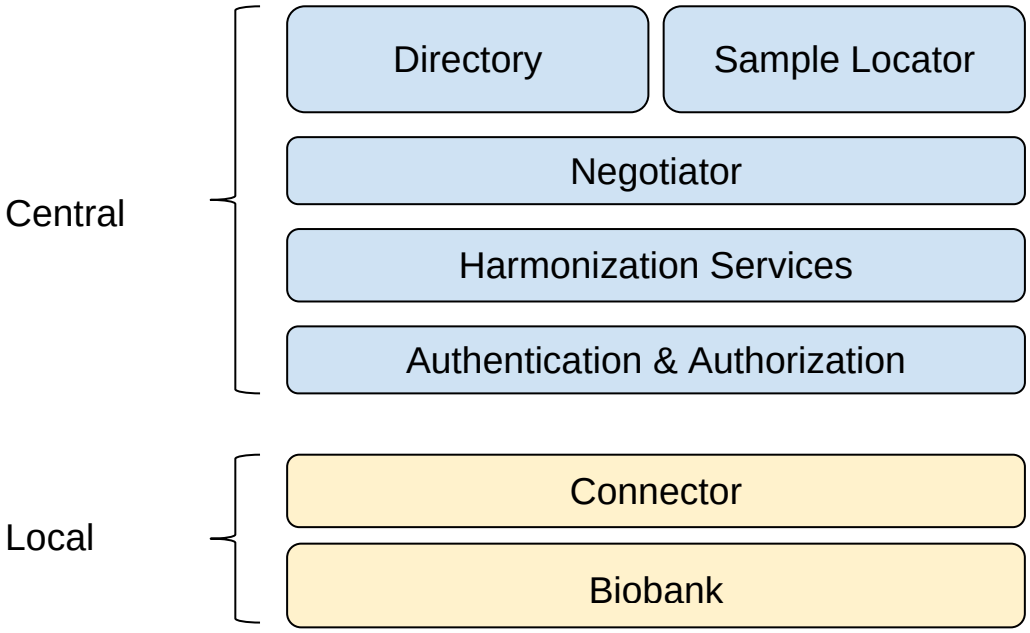

Figure 4 Architectural Framework of CS-IT WP1, WP2, WP8 and auxiliary components

There is an extraction, transformation, and loading (ETL) process that will take place when the data goes from the Biobank to the Connector and it involves data pseudonymisation. The Connector, although developed and distributed by BBMRI-ERIC CS-IT, is made to run on the biobank networks (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 will be implemented so that permissions can be managed centrally and users can 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 Sample Locator and the sharing of data and communication through the Negotiator. The data harmonization tools will enable, through the Metadata Repository and Terminology Mapping Tool, the structured communication between the different systems and heterogenous biobank data structures. The BBMRI-ERIC metadata definition will need to be mapped to the local data structures, so that the integration between the connector and biobank the biobanks information management system (BIMS) can be automated. For this reason, mapping tools will be analysed for reuse and a solution will be designed to help

biobanks map terminologies and thus integrate with BBMRI-ERIC. The section Metadata Repository (MDR) contains a description of the MDR and the section Federated Search (Sample Locator centric) includes an overview of the data flow on the federated search.

Architectural views

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.

Federated Search (Sample 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 1 illustrates the architecture needed to implement a federated search.

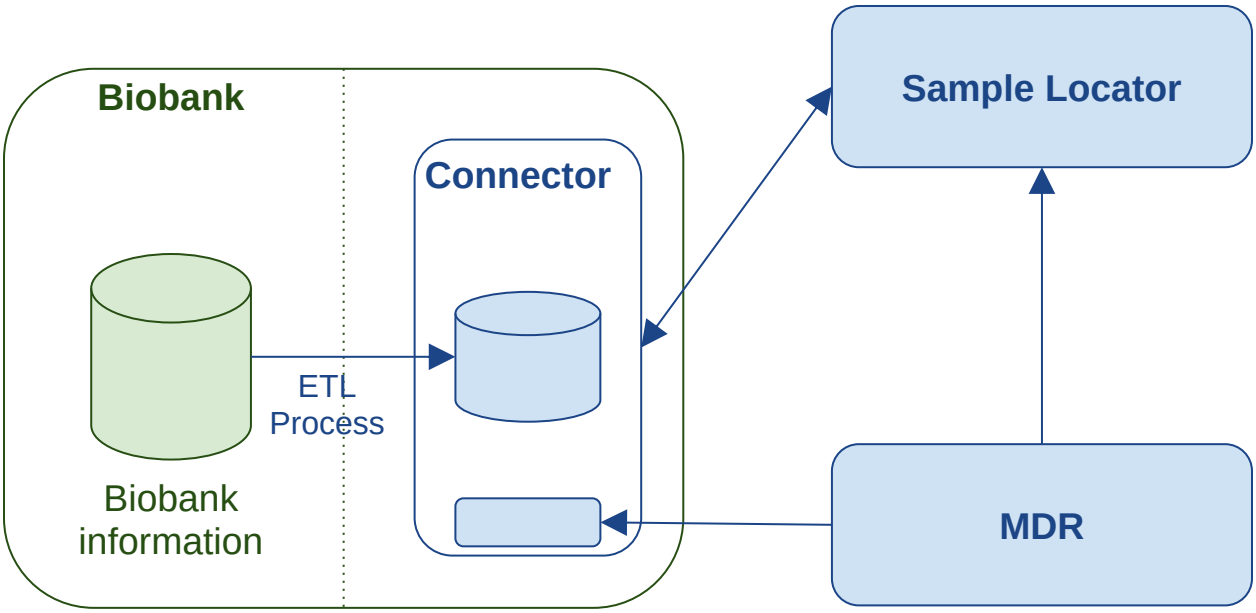

Figure 5 Integration of a Biobank with the Connector, Sample Locator and Metadata Repository (MDR)

In this section we represent the BBMRI-ERIC CS-IT federated biobank architecture. Figure 1Figure 6 depicts 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, Sample Locator, Connector, Metadata Repository (MDR), Biobank and Data Owner is illustrated.

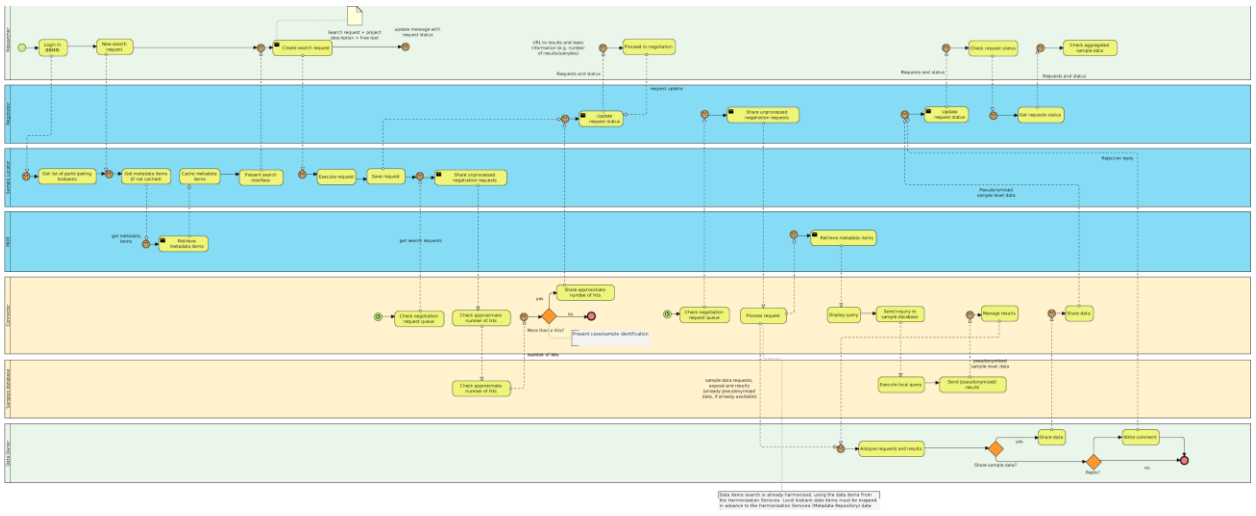

Figure 6 Overview of the federated research request

For the sake of simplicity, the interaction with the Authentication and Authorization Infrastructure (AAI) is not represented in this diagram. This workflow is represented in the Authentication and Authorization Infrastructure (AAI) section.

Directory Centric Negotiation

The directory centric negotiation process starts with a researcher identifying potentially useful samples in the Directory. From the Directory user interface, the research starts a new request and is redirected to the Negotiator, where he or she can see the used query and add a freetext description of the request. All biobankers with potentially useful samples are notified and can join the negotiating process in the Negotiator. Subsequent changes to the query or freetext 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.
The whole process has been modelled in detail, and the resulting sequence diagram is available in the project documentation. The behaviour of the Simple Negotiator in this process is described in the BBMRI-ERIC CS-IT deliverable D2.2 - Implementation and deployment of the sample Negotiator.

Colon Cancer Data Gathering

In the scope of BBMRI-ERIC ADOPT WP2, there is a plan for the manual collection of 3,000 colon cancer cases and semi-automated collection of an additional 7,000 cases. As a technical solution for this use case, we will deliver the architecture defined below, based on already existing, tested and running components of the OSSE registry software[3].

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

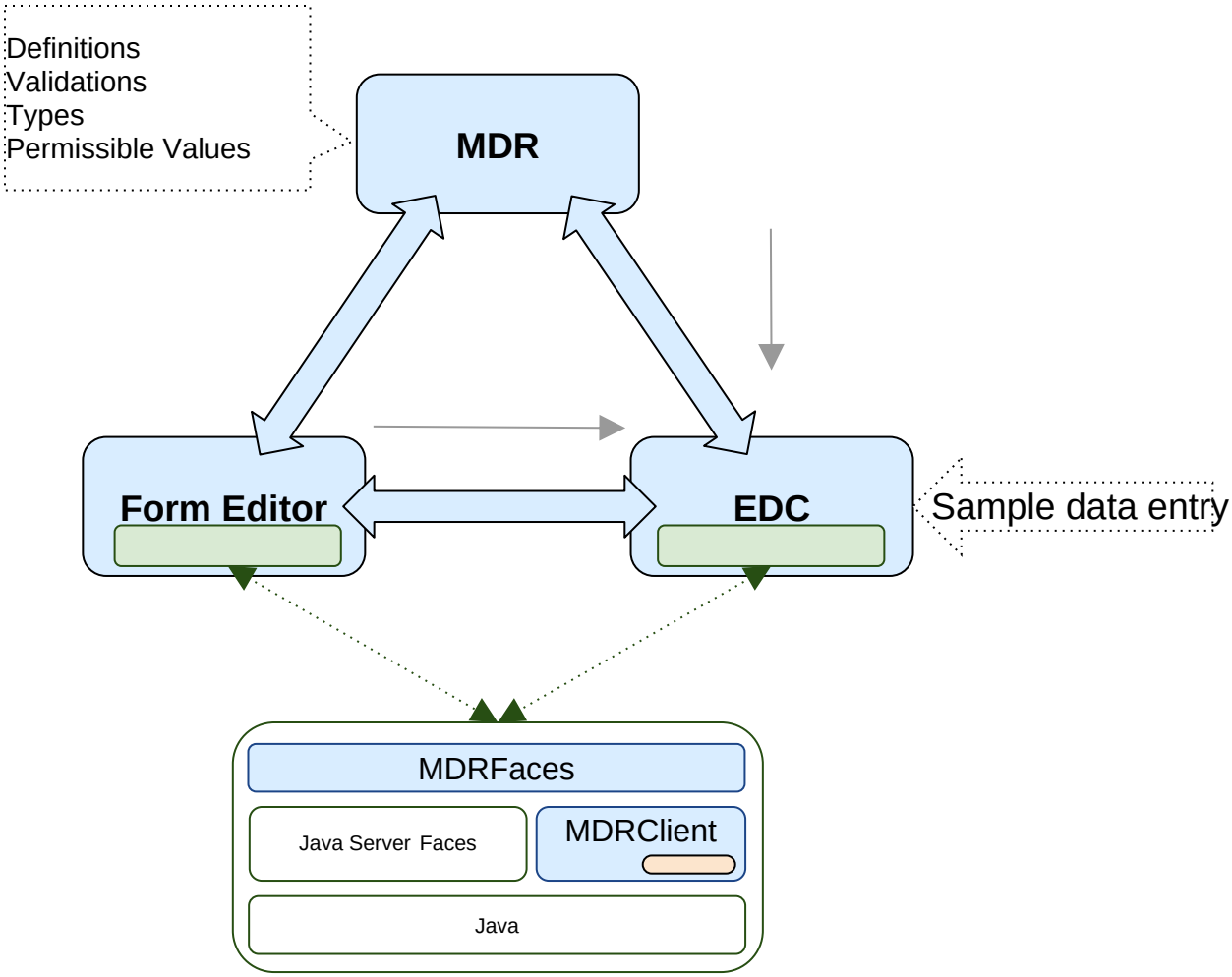

Figure 7 Architecture of the Colol Cancer Data Collection infrastructure

The architecture is visualized in Figure 7 Architecture of the Colol Cancer Data Collection infrastructure. The Metadata Repository (MDR), the MDRFaces and MDRClient libraries are described in the section Metadata Repository (MDR). An example view of the MDR content is shown in Figure 8. Additionally, a form editor has been created, where users can 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 computer software development knowledge.

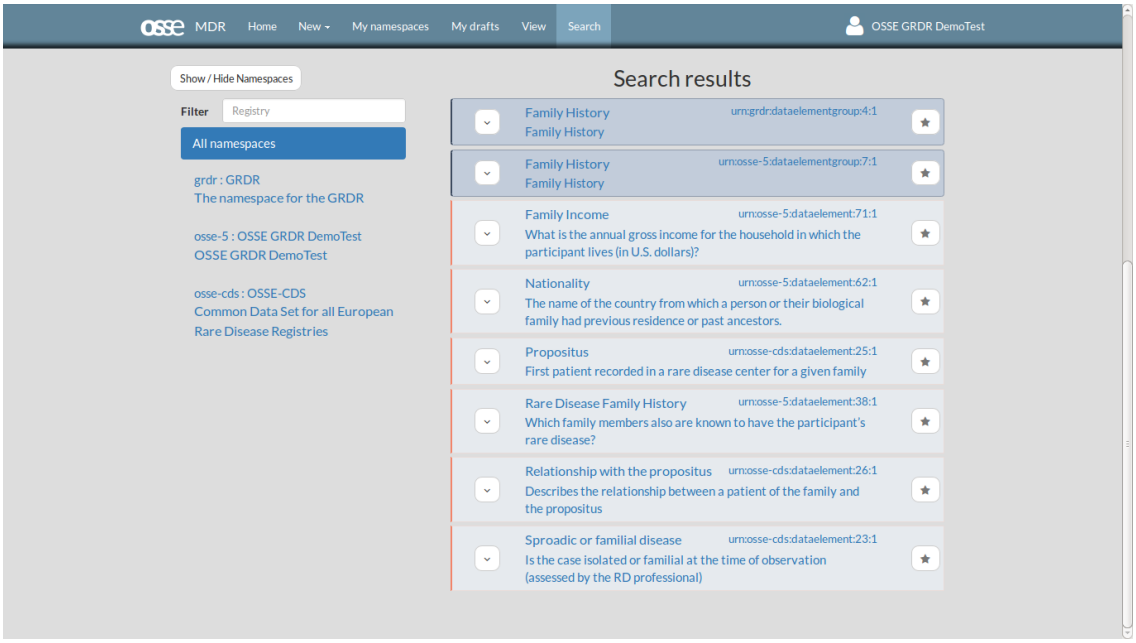

Figure 8 The MDR UI - Searching for a data element

For the colon cancer data gathering use case, the metadata entered in the MDR is based on the "Definition of Data Model for Colorectal Cancer Data Gathering in ADOPT BBMRI-ERIC" document[4].

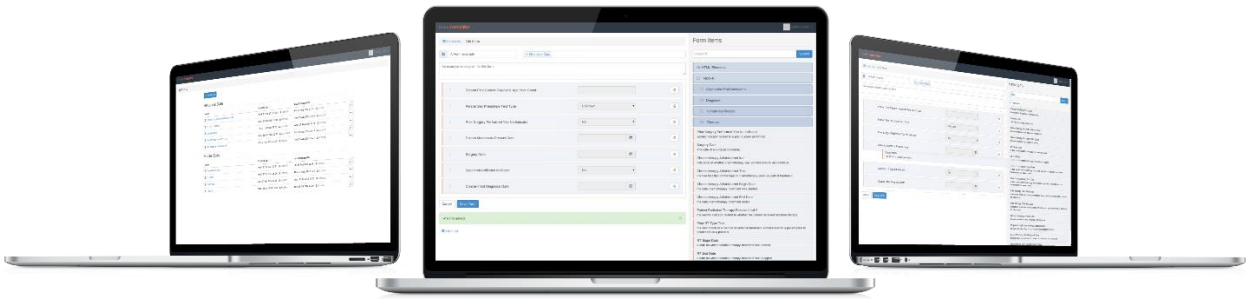

Figure 9 Form Editor - Form list, form detail and form metadata editing

EDC (impression shown in Figure 9) 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 collection of 7.000 cases, a REST interface will be developed. The structure of the POST messages will then include the MDR data item IDs (metadata keys) and the values, which are the content of the colon cancer cases.

[4] Definition of Data Model for Colorectal Cancer Data Gathering in ADOPT BBMRI-ERIC - https://docs.google.com/document/d/1h0iMClZmihow92BP4BNaulz_1EzgWfAT6Ek9YZDrJcE

Authentication and Authorization Infrastructure (AAI)

The Sample Locator, Negotiator and MDR provide services which need user authentication. For this purpose, BBMRI-ERIC CS-IT will provide 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, will be used. 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 10.

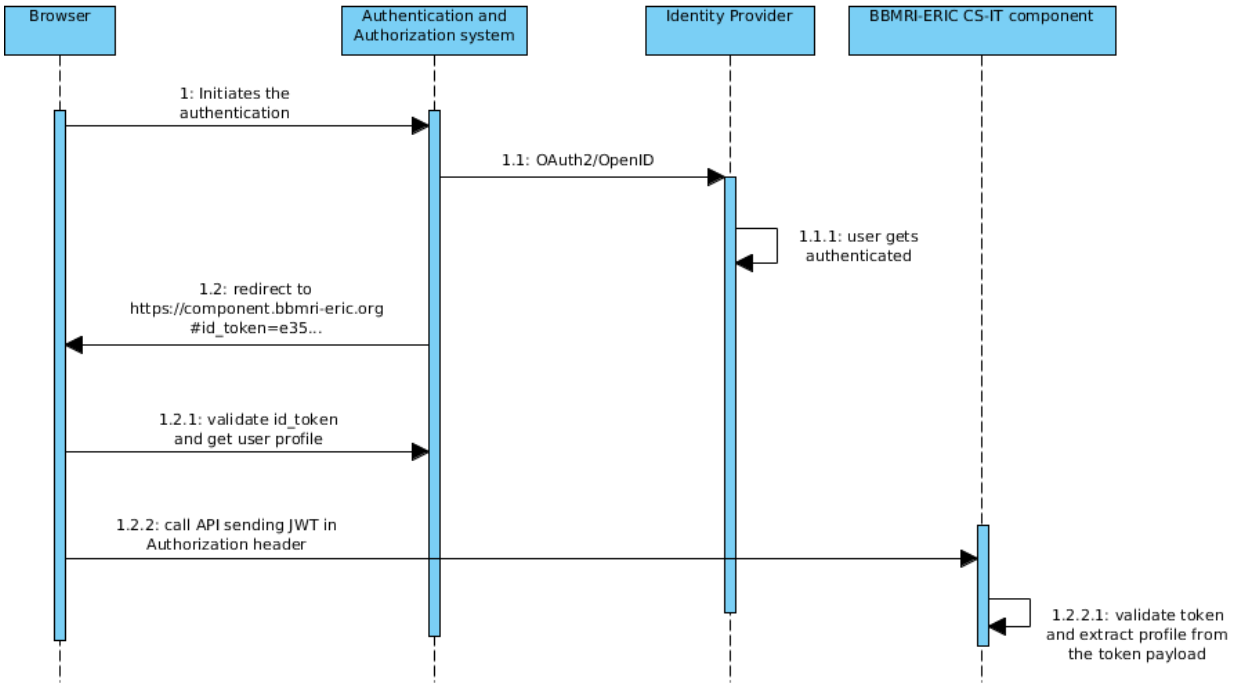

Figure 10 Authentication and Authorization Infrastructure