Project number:	317871
Project acronym:	BIOBANKCLOUD

Project title: Scalable, Secure Storage Biobank

Project website URL: http://www.biobankcloud.com/

Project Coordinator Name and Organisation:
Jim Dowling, KTH

E-mail: jdowling@kth.se

WORK PACKAGE 1 :

Regulatory and Ethical Requirements for Biobanking Data Storage and Analysis

Work Package Leader Name and Organisation:
JAN-ERIC LITTON, Karolinska Institute (KI)

E-mail: Jan-Eric.Litton@ki.se

PROJECT DELIVERABLE

D1.1 Informatics model specification and ethical guidelines for data protection and data sharing

Deliverable Due date (and month since project start): 2013-05-31, m6
Deliverable Version: v1.0

Document history

Version	Date	Changes	By	Reviewed
0.1	2013-04-05	First draft	Roxana Merino Martinez Jane Reichel	
0.1	2013-04-12	First draft		Jan-Eric Litton

Executive Summary

This deliverable is a continuation of D1.5. In D1.5 we defined a Model Data Management Policy (MDMP) for the BiobankCloud based on the European legal framework for personal data protection. The policy specifies the procedures for service access request and how to evaluate those service requests. This deliverable extends the policy to audit processes and provides the information model that should be handled by the platform in order to guarantee sharing of data among biobanks and researchers as well as the ethics regulations involved in this issue.

Table of Contents

1 Introduction

The BiobankCloud platform will bring data storage and analysis services to biobanks and bio-medical researchers. The data to be processed by the platform come from omics experiments made on biobanked samples. A sample is a part of a specimen taken for diagnosis or experimentation. It can be blood, tissue, RNA, DNA, cell line, etc. In a biobank, a sample has data and information associated to it as sample donor, sample owner, storage conditions, storage location, expiration date, type of sample, quality indicators, etc. The researchers made studies on samples and generate data, results, knowledge and papers that can be reused by other studies. A formal definition of biobank data is a crucial aspect for the sharing of data and knowledge.

BBMRI [1] designed a minimum dataset for biobanks in its preparatory phase. BBMRI.se continued this work that resulted in MIABIS 2.0: Minimum Information About BIobank data Sharing [2]. It specifies the required information to share biobank data at the metadata and aggregated levels excluding personal information about sample donors.

In this deliverable, we establish the specifications for the data model based on MIABIS 2.0 and extend it to cover the specific requirements of the platform. Making the BiobankCloud data model MIABIS 2.0 compliant will guarantee a link between the BiobankCloud and the biobank management systems.

Nevertheless, a data model for biobank data sharing is not complete if it is not supported by regulations about how biobank data can be shared. Section 2 of this document covers this issue.

Section 3 is a continuation of the D1.5. It provides some guidelines for audit to be carried out to evaluate the good practices related to personal data protection in the platform.

Sections 4 to 6 describe the procedure to define the data model in the BiobankCloud platform.

2 Legal and ethics framework for data sharing in the BiobankCloud

Even though the aim of the BiobankCloud is that all personal data within the platform should be anonymous, and therefore falling outside the scope of the EU Data Protection Directive, the legal and ethical standards of said directive will be followed. At this point, only data from Charité, Germany will be used.

As stated in deliverable D1.5, the owner of the data, in our case Charité, will remain the controller in the meaning of the Data Protection Directive, and the BiobankCloud will be the processor. The law applicable to the data uploaded will therefore be German law. Within the function of the controller lies the responsibility to secure that all data subjects whose data is uploaded to the BiobankCloud platform have given their consent, or that an exemption is applicable, all according to German law. If required by the law applicable to the controller, the consent given from the data subjects should also cover the possibility for the controller to share data with other potential users of the BiobankCloud platform. The *SIAC* referred to in section 3.1.1.a should also include information on the extent to which it is possible to share the data as well as all specifications or restrictions regarding the aims of the study requesting data.

The *SRPA* or *SIAC* should further include all necessary information regarding any limitations as to the time the data may legally be stored, according to the law applicable to the controller.

3 Audit related to ethics and legal framework

The aim of the audit processes will be the collection and evaluation of evidence of the BiobankCloud practices and operations regarding personal data protection and research data protection. The evaluation of obtained proofs determines if the platform is maintaining data integrity and security, and is operating effectively to achieve the platform's goals and objectives.
In the BiobankCloud the audit should be carried out by the administrators of the platform. However, it is recommended that the audit processes can also be performed by independent persons (e.g. ethics board).

The audit should cover at least the following areas:

1. Identification of general potential risks
2. Verification of procedure for administration and user authentication
3. Scanning architecture potential vulnerabilities
4. Verification of personal data protection
5. Verification of research data sharing
6. Verification of internal omics data encryption and storage

The last three areas (4, 5 and 6) are going to be explained in details due to their relationship with ethical and legal regulations.

3.1.1 Verification of personal data protection

In the BiobankCloud platform only non-identifiable personal data will be uploaded and managed. The audit for verification of personal data protection should be based on the BiobankCloud Model Data Management Policy (MDMP), specifically in the sections "Selection criteria for individual membership" and "Selection criteria for data store and analysis" and in the two main parts of the data model: study descriptive metadata and omics data.
Studies are accepted in the platform only when all the required information related to personal data protection is provided. Nevertheless, it should be a systematic way to check the integrity and availability of this information.

a) Verifications regarding study descriptive metadata

Checking of the existence of the standard forms:

- **SRPA**: Standard Research Project Approval
- **SIAC**: Standard Information About Consent

- **SINC**: Standard Information on Non-Consented data
- Agreement between controller & processor

Processes:

1. When a requester is accepted as a BiobankCloud user, raise awareness about her/his responsibility regarding privacy protection of sample donors
2. Given a study ID, retrieve the required standard forms
3. Check for the existence of all required standard forms for all the register studies
    a. Retention of data for those studies missing required standard forms
    b. Inform study's owner about the data retention

b) Verification regarding omics data

It implies to keep track of omics datasets and analysis results usage: when, for what purpose and by whom. The purpose can be: reading, updating, uploading, downloading, analysis.

Processes:

1. Given a study ID, report the data usage (datasets and analysis results)
2. The system should have a way to identify the omics datasets that can be used for re-identification of the sample donor

3.1.2 Verification of research data sharing

The data owner (controller) is the only authorized to use her/his data. A tracking of data usage should be implemented for verification purpose.
In case the platform implements data sharing, the regulations for data sharing depend on the law applicable to the controller and they have to be specified in the ***SIAC*** standard form, as mentioned in section 2.
A process should be implemented to keep track of the data usage: when, for what purpose and by whom. It includes study descriptive metadata, omics datasets and analysis results.

3.1.3 Verification of internal omics data encryption and storage

As a principle, all data managed by the platform will be anonymous. Encryption and sharding should be sufficient to render the data as

anonymous in the case of omics datasets. A process need to be implemented to verify this statement.

3.1.4 Data retention

The data associated to a study is not available for use including the omics data and analysis results if:

- When verifying study descriptive metadata, some standard forms are missing (-deleted from the system or corrupted)
- When verifying data anonymization some risks are detected
- When verifying data usage a problem with data ownership is detected

The data owner should be informed and asked to correct the problem.

4 Informatics model specification

As mentioned in the introduction of this document, MIABIS 2.0 [2] will be used and extended to represent the data in the platform.

4.1 MIABIS 2.0

MIABIS 2.0 is a formal definition of attributes describing biobank data that can be shared for research purpose. It suggests a very simple structure with two major entities: Biobank and Sample Collection/Study. An example of use of MIABIS 2.0 can be found in the BBMRI.se Sample Collection Register:

http://bbmriregister.se

MIABIS 2.0 is defined at the metadata and aggregated levels to permit data sharing without compromising personal data privacy.

4.1.1 Data describing biobanks

Code	Name	Allowed Values	Formula Type	Description
MIABIS-01	Biobank ID	Free text in any language	Open text	Textual string of letters starting with the country code (according to standard ISO1366 alpha2) followed by the underscore "_" and post-fixed by a biobank ID or name specified by its juristic person (nationally specific)
MIABIS-02	Name of biobank	Free text in English	Open text	Textual string of letters denoting the name of the biobank in the local language.
MIABIS-03	Juristic person	Free text in English	Open text	Textual string of letters denoting the juristic person e.g. a university, concern, county council etc. for the biobank
MIABIS-04	URL	Free text in English	Open text	Textual string of letters with the complete http-address for the biobank URL
MIABIS-05	Country code	ISO-standard (3166 alpha2), two letter code	Drop list with countries spelled out	Textual string of letters of the two letter code for the country of the biobank according to ISO-standard 3166 alpha2
MIABIS-06	Biobank type	Pathology, Cytology, Gynecology, Obstretics, Transfusion, Transplant, Clinical chemistry, IVF and similar, Bacteriology, Virology, Other	Check-box	Textual string of letters indicating the underlying (primary) medical area within which the samples are collected. Can be several values. The allowed values for this attribute is under debate
MIABIS-07	Contact person	Free text in any language	Open text	Textual string of letters denoting the name of the contact person for the biobank
MIABIS-08	Contact phone	Free text in English	Open text	Phone to the "Contact person", including international call prefix
MIABIS-09	Contact email	Free text in English	Open text	Email address of the "Contact person"
MIABIS-10	Contact department	Free text in English	Open text	Department, or corrsponding (e.g., division), of affiliation of the "Contact person"
MIABIS-11	Contact address	Free text in English	Open text	Street name and street number or PO Box of the "Contact person"

MIABIS-12	Contact ZIP	Free text in English	Open text	ZIP of the "Contact person"
MIABIS-13	Contact city	Free text in English	Open text	City of the "Contact person"
MIABIS-14	Contact country	ISO-standard (3166 alpha2), two letter code	Drop list with countries spelled out	Country of the "Contact person"
MIABIS-15	Hosted studies	StudyID, sample collection acronyme	Open text	Textual string of letters identifying the studies/sample collections that the biobank is physically hosting. Can be multiple values
MIABIS-16	Date of Entry	ISO-standard (8601) time format	n/a	Date in ISO-standard (8601) time format when data about the biobank was reported into a database.
MIABIS-17	Last updated	ISO-standard (8601) time format	n/a	Date in ISO-standard (8601) time format when data about the biobank was last updated in a database.

4.1.2 Data describing Sample Collection/*Study

Code	Name	Allowed Values	Formula Type	Description
MIABIS-18	Sample Collection/Study ID*	Free text in English	Open text	Textual string depicting the unique ID or acronym for the sample collection or study
MIABIS-19	Study name*	Free text in any language	Open text	Textual string of letters denoting the name of the study in English
MIABIS-20	Description	Free text in English	Open text	Textual string of letters describing the sample collection or study aim (max 200 charachters)
MIABIS-21	Sample Collection Responsible / Principal Investigator*	Free text in any language	Open text	Textual string of letters denoting the name of the sample collection responsible or principal investigator
MIABIS-22	Contact person	Free text in any language	Open text	Textual string of letters denoting the name of the contact person for the sample collection or study
MIABIS-23	Contact phone	Free text in English	Open text	Phone to the "Contact person", including international call prefix.
MIABIS-24	Contact email	Free text in English	Open text	Email address of the "Contact person"
MIABIS-25	Contact department	Free text in English	Open text	Department, or corrsponding (e.g., division), of affiliation of the "Contact person"
MIABIS-26	Contact address	Free text in English	Open text	Street name and street number or PO Box of the "Contact person"
MIABIS-27	Contact ZIP	Free text in English	Open text	ZIP of the "Contact person"
MIABIS-28	Contact city	Free text in English	Open text	City of the "Contact person"
MIABIS-29	Contact country	ISO-standard (3166 alpha2), two letter code	Drop list with countries spelled out	The two letter code in format following ISO-standard (3166 alpha2) for the country of the contact person
MIABIS-30	Type of Collection	Case-control, Cohort, Cross-sectional, Longitudinal, Twin-study, Quality control, Population-based, Disease specific, Other	Check-box	Textual string of letters denoting the type of sample collection or study design. Can be one or several values
MIABIS-31	Collection start	ISO-standard (8601) time format	Calendar	Date in ISO-standard (8601) time format specifying when the sample collection starts

MIABIS-32	Collection end	ISO-standard (8601) time format	Calendar	Date in ISO-standard (8601) time format specifying when the sample collection ends, if applicable
MIABIS-33	Planned sampled individuals*	Integer	Drop list with number	Number of individuals with biological samples planned for the study (also see CurrentSampledIndividuals)
MIABIS-34	Planned total individuals*	Integer	Drop list with number	Total number of individuals planned for the study (also see CurrentTotalIndividuals)
MIABIS-35	Sex	Female, Male	Check-box	Textual string of letters denoting the sex of the sample donors. Can be several values
MIABIS-36A	Age Group Low	Integer	Drop list with ages	Age of youngest participant at start of study
MIABIS-36B	Age Group High	Integer	Drop list with ages	Age of oldest participant at start of study
MIABIS-37	Average age	Integer	Open text	Average age of all sample donors in the sample collection
MIABIS-38	Main diagnosis	http://apps.who.int/classifications/apps/icd/icd10online/	Tree-structure multi-select list	ICD-10 codes for the studied diagnoses. Can be several values
MIABIS-39	Comorbidity	Yes, No	Check-box	Textual string of letters indicating if information about comorbidity is available
MIABIS-40	Categories of data collected	Biological samples, Register data, Survey data, Physiological measurements, Imaging data, Medical records, Other	Check-box	Can be several values
MIABIS-41	Material type	Whole blood, Plasma, Serum, Urine, Saliva, CSF, DNA, RNA, Tissue, Faeces, Other	Check-box	Most commonly abundant biological samples in biobanks. Can be several values.
MIABIS-42	Omics experiments*	Genomics, Transcriptomics, Proteomics, Metabolomics, Lipidomics, Other	Check-box	Can be several values
MIABIS-43	Survey Data	Individual Disease History, Individual History of Injuries, Medication, Perception of Health, Women's Health, Reproductive History, Familial Disease History, Life Habits/Behaviors, Sociodemographic Characteristics, Socioeconomic Characteristics, Physical Environment, Mental Health, Other	Check-box	Topics covered by a survey/questionnaire answered by the study participants. Can be several values
MIABIS-44	Medical records	Free text in English	Open text	Free text specifying which medical record data is available in the sample collection/study
MIABIS-45	Registers	Free text in English	Open text	Free text specifying which registry data is available in the sample collection/study
MIABIS-46	Sample handling	Free text in English	Open text	Textual string of letters describing how the samples in the sample collection have been handled as an indication of sample quality. Can be one or several of the following values: Freeze chain, indicating if the samples in the collection have been kept cool from needle

				to freezer. Freeze time, time in hours from needle to freezer. SPREC compliant, if the samples are labeled according to SPREC, Other
MIABIS-47	Storage temperature	Room temperature, +4C, -18C to -35C, -60C to -85C, Liquid nitrogen, Other	Check-box	Can be several values
MIABIS-48	Current sampled individuals	Integer	Drop list with number	Number of individuals with biological samples in the sample collection/study at the date of Last updated (also see Planned sampled individuals)
MIABIS-49	Current total individuals	Integer	Drop list with number	Total number of individuals in the sample collection/study at the date of Last updated (also see Planned total individuals)
MIABIS-50	Date of entry	ISO-standard (8601) time format	n/a	Date in ISO-standard (8601) time format when data about the sample collection was reported into a database
MIABIS-51	Last updated	ISO-standard (8601) time format	n/a	Date in ISO-standard (8601) time format when data about the sample collection was last updated in a database
MIABIS-52	Hosting biobank	BiobankID	Open text	Textual string of letters of the biobank/s storing the biological samples that are part of the sample collection. Can be several

4.1.3 Optional information

Data describing publication

Code	Name	Allowed Values	Formula Type	Description
MIABIS-53	Publication name*	Free text in English	Open text	Textual string of letters denotating the name of the publication. Can be several
MIABIS-54	Publication DOI*	Free text in English	Open text	Textual string of letters denotating the DOI name (Digital Object Identifier) of the publication. Can be several
MIABIS-55	Publication study design*	Cohort, Case-control, Cross-sectional, Other	Dropdown list	Name of the study design used in the publication. Can be several
MIABIS-56	Total sample donors*	Integer	Open text	Total number of donors in the study design of the publication
MIABIS-56A	Cases*	Integer	Open text	Total number of cases (patient, treatment) in the study design of the publication. Used for the Case-control study design

Data describing omics experiment related to publication

Code	Name	Allowed Values	Formula Type	Description
MIABIS-57	Array-based vendor*	Affymetrix, Applied Biosystems, Roche Applied Science, Helicos BioSciences, Illumina, Other	Dropdown list	Name of the array-based vendor
MIABIS-58	Platform*	ASCII	Free text	Platform name or identifier given by the array-based vendor
MIABIS-59	Imputation method*	IMPUTE, MACH, HapMap, Other	Check box	Name of the imputation method. Can be several

4.2 Considerations about implementing MIABIS

MIABIS 2.0 covers the sample specification from the biobank to the research study at the metadata and aggregated levels. With the optional attributes, it also includes the publications and omics experiments.

A hierarchical representation of the data based on MIABIS can be seen as follows:

Fig 1. The white blocks represent the optional entities in MIABIS 2.0

MIABIS 2.0 refers only to human samples. The main users of the BiobankCloud will be biomedical researchers. Not all the experiments are done on human samples. It would be necessary to add the attribute "species" to define the species involved in the study. Another important attribute for the study is the anatomical part where the sample is coming from.
Moreover, the biobank information can be not as relevant for the platform as the study information is. The same can happen with the sample collection information. The data uploading should have the flexibility of handle the data from different starting points:

- Biobank
- Sample collection
- Study

On other hand, MIABIS 2.0 does not include specified omics information. It needs to be added to the data model.

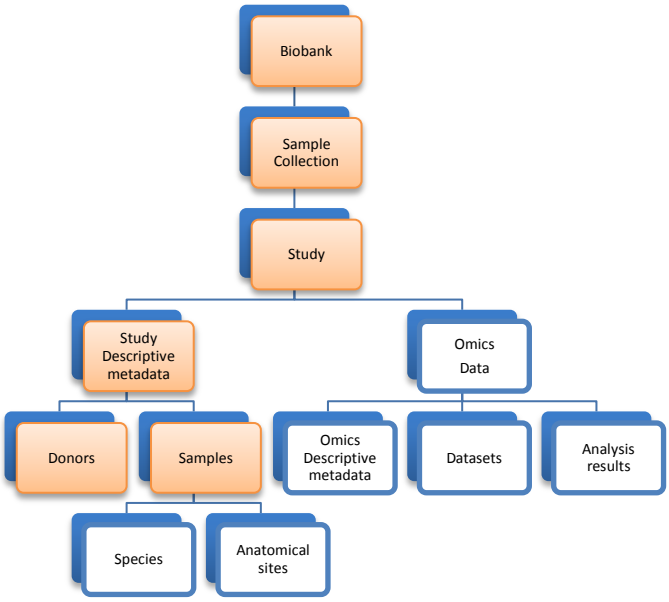

Fig 2. The white blocks represent extensions to be made to MIABIS 2.0

In summary, MIABIS 2.0 model should be modified to fulfil the platform requirements keeping in mind the following considerations:

- Not all the information about biobank needs to be uploaded
- Not all the information about sample collection needs to be uploaded
- The study descriptive metadata should include information about the origin of the sample according to:
  - Species
  - Anatomical parts
- Omics experiment, data, analyses and results need to be added

5 Considerations about the user interface

There are three major processes in the platform:

- Data uploading
- Data processing
  - Search for data
  - Analysis
- Data downloading

The user interface can be:

- Command line based
- Graphical User Interface (GUI)

The main element in the data model is the study and it is composed of two major parts:

- Study descriptive metadata
- Omics data

The data model is designed in the way that most of the relevant attributes describing a study are pre-defined. The free text is avoided as much as possible in order to standardize the information associated to studies and also to standardize the query engine.
One study has one descriptive metadata and can have several omics data associated to it. One omics data can have several datasets each of them have a descriptive metadata and several analysis results associated to it.

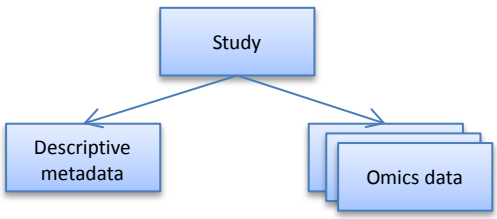

Fig 3 Main component of study information

5.1 Command line based user interface

5.1.1 Data uploading

It is divided into two processes:

- Uploading study descriptive metadata (can include or not biobank and sample collection information)
- Uploading omics data

A command line user interface implies that the user has to pre-process the data in order to structure it according to a standard format defined by the platform (XML, RDF, delimited files, etc.). At the same time, the system has to parse the input files in order to interpret the content.
Some important attributes must be easily identified by the system, example:

- Name and location of files corresponding to standard forms (as described in MDMP)
- Required standard forms for the study
  - For instance, if not all the samples are consented (old samples) then the SINC is required
- Scientific aims of the investigation
- Anonymization method for omics datasets that can be used to identify sample donors

5.1.2 Data processing

The data processing consists of:

- Selection of data to be processed
- Selection of scientific workflow
- Storage of the analysis results

In a command line based user interface, it implies that the user has to be very careful choosing the datasets. It can be difficult to implement a user-friendly way to show the study information and the omics information to facilitate the data selection. The same can happen with the selection of the analysis workflows.

5.1.3 Data downloading

In a command line environment, it could be difficult to select the data to be downloaded if it depends on elaborated search criteria.

A command line-based environment can require more maintenance work that a GUI. It can be more complicated if the platform will implement data sharing. On other hand, from the user point of view, it would require expertise in command line-based environments limiting the use of the platform to bioinformatics experts or advanced computer users.

5.2 Graphical User Interface

A graphical use interface would notably improve the use and maintenance of the platform. From the user point of view, it would be easier to manage

data. Data uploading, downloading, updating and deleting can be made interactively. The information provided can be easily modified and traceable.

Once the main platform goals are fulfilled, it is highly desirable to implement a graphical user interface to make the platform more attractive to the research community.

6 Data Model: First draft

This data model represents research studies on biobanked samples. It covers the data management and processing including the application of regulations defined in the MDMP and is designed keeping into account the future implementation of a disclosure filter for biobank data sharing.

6.1 Actors and roles definition

In the BiobankCloud platform the main user is the researcher. Being an accepted member of the platform, a researcher can upload, download, update and process data. These actions can be easily identified with roles that should be specifically managed by the system.
Other users can be the platform administrators, a person representing an organization that wants to be member of the platform, a person from the ethics board that wants to execute some audit processes.
Researchers, platform, organizations and ethic board; are actors that can play different roles in the system [3].
The deliverable D1.5, section 2.3 defines the controller and processor according to the Data Protection Directive [4]. The processor is the BiobankCloud platform who provides the means of the processing. The controller is the person or organization who "determines the purposes and means of the processing of personal data". Both, processor and controller can be defined as roles in the platform. For the sake of establishing a transparent and reliable chain of command between controller and processors, the BiobankCloud should take care to allocate all responsibilities stemming from the Data Protection Directive in a clear and concise manner.

This data model identifies actors and roles related to data storage and analysis as follows:

Actor	Roles
Researcher	Controller Trusted researcher Guest
Platform	Administrator Processor Auditor Access committee
Ethic board	Auditor Guest
Organization	Guest

6.1.1 Processor

In D1.5 a processor is defined in terms of personal data protection as "the natural or legal person, public authority, agency or any other body which processes personal data on behalf of the controller".

In principle, the processor is the BiobankCloud platform.

6.1.2 Controller

A *controller* in the BiobankCloud platform is a person appointed by the organization to upload data to the platform. When a controller uploads data to the platform creates a "data space". The data space can hold data from one or several research studies that can be access by the controller itself or by a *trusted researcher* (6.1.3).
This role involves the knowledge of the MDMP, specifically "Storage Access request".
The role *controller* can inherit the privileges of the *trusted researcher* in order to use a created data space. Only the *controller* can manage the information related to personal data protection.
A controller can also inherit privileges of the *access committee* role (6.1.6) in order to grant access to trusted researchers to a specific data space.
In case data sharing is implemented, only the *controller* can grant or revoke access to data for sharing. The information regarding data protection is not shareable.

6.1.3 Trusted researcher

This role is meant to allow several researchers manage the same data space. A *trusted researcher* can't upload data to the platform. Only the *controller* has this privilege.
A researcher as a *trusted researcher* analyses and downloads data from one or more data spaces. The *trusted researcher* can update or delete analysis result data but not the information uploaded by the *controller*.
A *trusted researcher* needs to apply for access permission to existing data spaces. The *controller* grants this access.

6.1.4 Guest

This role permits to explore services and information about studies stored in the platform. In case of data sharing implementation, a *guest* could search for data availability using search criteria.
The *guest* role has very limited privileges. This role can only search at the metadata and aggregated level for the studies that have granted access to be visible.
This role has more sense for a web interface where visitors can get information about the BiobankCloud services and stored data.

6.1.5 Auditor

This role can be established to carry out audit processes. For instance, the Ethics Board could access the platform using this role.

6.1.6 Access Committee

As defined in D1.5, an access committee will evaluate the access requests. The *Access committee* role permits to automate the acceptance of organizations and researchers as members of the platform. The *controller* role can inherit privileges from this role in order to grant access to a *trusted researcher* to specific data spaces.

6.1.7 System Administrator

This is the role to maintain and operate the platform. It inherits the privileges of the rest of the roles.

6.1.8 Roles in action

When an organization is accepted as member of the BiobankCloud, based on D1.5, section 2.3, the organization is the *controller*. The organization can appoint one or more persons to carry out the *controller* role in the BiobankCloud platform. For instance, it could be convenient to have a *controller* for each department, for each research group or for each study. When a researcher request membership on behalf of one organization it implies also a request to at least one *controller* to be able to use at least one data space.
A controller can create data spaces, use data spaces and grant access to trusted researchers to the created data spaces.
A researcher applies for membership and gets the *trusted researcher* role. Then, she/he can request access to created data spaces. The *controller* that created the requested data space, accept or deny the access. The *controller* could also grant access to data spaces to *trusted researchers* without request.
Flexibilities can be added to this model in order to easier the access to data but it has to be done observing all the regulations specified in the BiobankCloud ethical and legal framework.

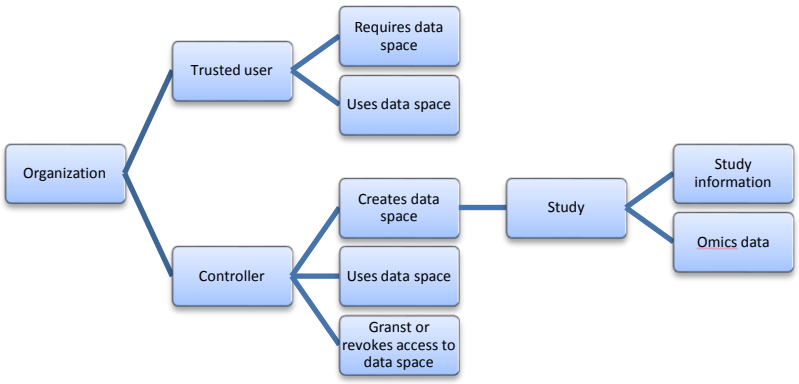

Fig 4. Trusted user and controller roles

6.2 Main use cases

The following use cases are derived from the system requirement related to data storage and analysis.

No	Use Case Name	Primary Actor	Role
1	Apply for membership	Researcher Organization	Guest
2	Store study descriptive metadata	Researcher	Controller
3	Store Omics data	Researcher	Controller
4	Search based on specific criteria	Researcher	Guest Trusted researcher
5	Analyze data	Researcher	Trusted researcher
6	Grant access to study data	Researcher	Controller
7	Grant access to analysis results	Researcher	Controller
8	Revoke access to study data	Researcher	Controller
9	Revoke access to analysis data	Researcher	Controller
10	Download sample data	Researcher	Trusted researcher
11	Download analysis results	Researcher	Trusted researcher
12	Delete the study data	Researcher	Controller
13	Update the descriptive metadata	Researcher	Controller
14	Update omics data	Researcher	Controller
15	Store the data usage	Platform	Administrator
16	Audit the data usage	Platform	Administrator Auditor
17	Add organization	Platform	Administrator Access committee
18	Remove organization	Platform	Administrator Access committee
19	Register users in the BiobankCloud	Platform	Administrator Access committee
20	Terminate users in the BiobankCloud	Platform	Administrator Access committee
21	Suspend users in the BiobankCloud	Platform	Administrator Access committee
22	Update user information	Platform Researcher	Administrator Trusted researcher
23	Login to the BiobankCloud	Reseacher Platform Ethics board	Controller Trusted researcher Administrator Auditor Access committee
24	Logout from the BiobankCloud	Researcher Platform Ethic board	Controller Trusted researcher Administrator Auditor Access committee
25	Get the resources usage	Platform	Administrator

26	Define the user budget	Platform	Administrator

The use cases 2 and 3 represent the data uploading to be carried out by the *controller*:

Use case 2: Store study descriptive metadata

Use Case Element	Description
Use Case Number	2
Use Case Name	Store study descriptive metadata
Use Case Description	A user selects the option “Store descriptive metadata” and upload a study descriptive metadata including the standard forms required by the MDMP
Primary Actor	Researcher
Role	Controller
Precondition	User has to be logged in
Basic Flow	1. System display controller options 2. Researcher select “Store Study Descriptive Metadata” 3. System ask for **SRPA** Document (Approvals for the research project) 4. Researcher uploads the document 5. System asks for **SIAC**: Standard Information About Consent 6. Researcher uploads the document 7. System asks for total number of donors 8. Researcher enter a number 9. System asks for number of donors without consent 10. Research enter a number 11. System calculates the non-consented number. If the number >0, asks for **SINC**: Standard Information on Non-Consented data 12. Researcher uploads the document 13. System checks the standard forms (research authorization and consent) 14. Researcher is granted to store data 15. Researcher upload study descriptive metadata 16. System generates a **StudyID** 17. System asks for the aim of the study 18. Researcher select a keyword or provide a new keyword as aim (-this is for future data sharing)

Alternate Flows	• Privacy protection problem
	  ○ In 13, if there is a problem with the provided information (SRPA, SIAC, SINC)
	  ▪ System informs about the problem and suggest possible solutions
	  ▪ Researcher is not granted to store study data

	▪ No StudyID is not generated

Use case 3: Store omics data

Use Case Element	Description
Use Case Number	3
Use Case Name	Store omics data
Use Case Description	A user selects the option “Store omics data”. Omics data can only be uploaded once the study descriptive metadata has been uploaded and assigned a StudyID. The omics data is associated to a selected StudyID. The user provides omics descriptive metadata and the omics datasets. Several omics data can be associated to the same StudyID.
Primary Actor	Researcher
Role	Controller
Precondition	User has to be logged in The researcher has uploaded at least one study descriptive metadata
Basic Flow	1. System display controller options 2. Researcher selects “Store omics data” 3. System presents a list of studies already uploaded by the researcher 4. Researcher selects a StudyID to associate omics data to it. 5. System ask for data size 6. Researcher select one 7. System presents a list with type of omics 8. Researcher selects one 9. If the omics data can lead to donor identification, system present a list of anonymation methods 10. Researcher selects method 11. Researcher is granted to store data 12. Researcher upload omics descriptive metadata 13. System check for privacy protection (anonymization if required) 14. System authorize data storing 15. Researcher uploads omics data files

Alternate Flows	• Researcher has no StudyID to associate omics data to it
	  ○ System suggests to provide first the study descriptive metadata and redirect to use case 2
	• Data size exceed the maximum allowed
	  ○ System suggests to reduce the size of the data and try again
	• Privacy protection problem. The omics data is not anonymized
	  ○ System suggest anonymization method

6.3 Classes and associations identification

The identification of classes is based on the system requirements for data storing, analysis and sharing. It has been defined in section 6.2 (Main use cases).

	Biobank Cloud	Organization	User	Biobank	Sample Collection	Study	Omics Data	Researcher	Administrator	Contact Information	Country	Regulation
BiobankCloud		maintains	maintains			stores			has	has	has	has
Organization	Member-of (aggregated)		Has (aggregated)	has				has		has	Belong-to	has
User		Belong-to (aggregated)								has	Belong-to	
Biobank		Belong-to			has			supports		has	Belong-to	has
Sample Collection				Is-in		Used-by				has		has
Study					uses		generates	Conducted-by				has
Omics Data						Generated-by		Owned-by				
Researcher			Is-a	Supported-by		conducts	owns					
Administrator	Belong-to		Is-a							has	has	
Contact Information	Belong-to	Belong-to	Belong-to	Belong-to	Belong-to						has	
Country										Part-of		has
Regulation	Established by	Established by		Established by	Established by						Established by	

6.4 First class model

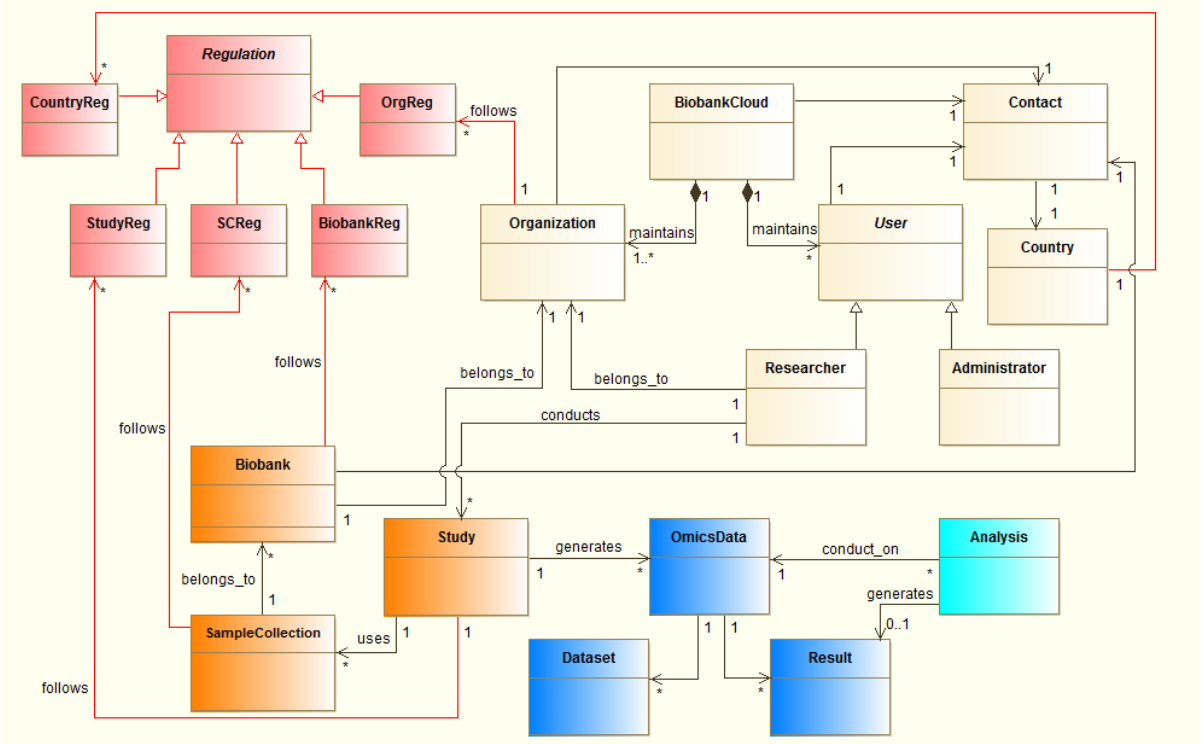

Fig 5 Draft of the class diagram

Generalizations are made only for *User* and *Regulation*. The omics data needs to be formalized according to the type of omics to be processed by the platform. The same applies for the *Analysis* class.

7 Conclusions

This deliverable is a preparation for the design of the informatics model for data storage and analysis to be provided by the deliverable D1.2. It provides WP2, WP3 with guidelines to implement security and storage data structure.

Sections 2 and 3 settle bases for the design of the standard forms to be delivered with D1.3. New regulations regarding ethical issues should be added to the BiobankCloud ethical framework with the help of the BiobankCloud Ethical Board.

Regarding the data model, not all the system requirements are capture in the use cases but only those related to user interaction with the platform, and processes involving data protection and biobank data management. The class diagram only contains the classes and associations. Redefining of the class diagram and definition of attributes and operations will be cover in D1.2.

8 References

1. *BBMRI, Biobanking and Biomolecular Resources Research Infrastructure, EC GA No. 212111, Work Package 5, D5.6 Final report, Annex 16, 2010.*
2. *A Minimum Data Set for Sharing Biobank Samples, Information, and Data: MIABIS, Biopreservation and Biobanking. August 2012, 10(4): 343-348. doi:10.1089/bio.2012.0003*
3. *The Role of "Roles" in Use Case Diagrams. http://infoscience.epfl.ch/record/268/files/WegmannG00.pdf*
4. *Kuan Hon, W, Millard, C. Walden, I, Who is responsible for 'personal data' in cloud computing?—The cloud of unknowing, Part 2, International Data Privacy Law, 2012, Vol. 2, No. 1*