BioMedBridges

Deliverable D10.3

Project Title:	Building data bridges between biological and medical infrastructures in Europe	
Project Acronym:	BioMedBridges	
Grant agreement no.:	284209	
	Research Infrastructures, FP7 Capacities Specific Programme; [INFRA-2011-2.3.2.] “Implementation of common solutions for a cluster of ESFRI infrastructures in the field of "Life sciences"	
Deliverable title:	A prototype federated query interface for information on biosamples, and linking of biosamples and disease terminology to genome	
WP No.	10	
Lead Beneficiary:	3: KI	
WP Title	Integrating disease related data and terminology from samples of different types	
Contractual delivery date:	31 December 2015	
Actual delivery date:	21 December 2015	
WP leader:	Alvis Brazma	1: EMBL
Partner(s) contributing to this deliverable:	1: EMBL, 3: KI, 15: UCPH	

Authors: Roxana Merino, Suyesh Amatya: Karolinska Institutet, Rafael Jimenez, José Villaveces: Elixir, Morris Swertz: Groningen University (Molgenis), Marco Roos: Leiden University Medical Center, Jan-Eric Litton, Petr Holub: BBMRI-ERIC, Søren Brunak: UCPH, Ugis Sarkans: EMBL

Contents

1 Executive summary

This deliverable finalizes the work on linking biological sample-related information across data infrastructures, data availability levels and information types. A federated system for sample discovery across biobanks is described, and we also report on the latest work on linking disease classification to the genome.

We built a prototype for biobank federation through MIABIS 2.0, which is the de facto standard for biobank data sharing in BBMRI-ERIC infrastructure and has also been adopted by other biobank networks and projects. We have developed a sample-centred biobank federation framework that aims to facilitate sample discovery among biobanks members of a federation. At the moment the software is called “MIABIS Connect” and its main features are:

1. Open-source software framework that can be easily adopted by the biobank and research communities.

2. Allows biobanks keeping their idiosyncratic semantics and at the same time be able to share data using MIABIS semantic. The harmonization process is data model agnostic.

3. Biobank data stays in the biobank. Only the results of queries are fetched from the biobanks to be viewed through the common query interface.

4. Requires a minimum involvement of biobank IT support in order to deploy and maintain the federation framework.

Following on the association of ICD10 codes to genes, we have developed a new approach creating associations to microRNAs. Since their discovery in 1994, miRNAs as a target are gaining attraction in the research community. To our knowledge, no-one has worked on associating miRNAs to ICD10 diagnostic codes before.

2 Project objectives

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

No.	Objective	Yes	No
1	Linking disease-related data to molecular information: terminology	X	
2	Linking disease-related data to molecular information: data	X	

3 Detailed report on the deliverable

3.1 Biobank federation: background

At present, it is difficult to access and integrate on demand information from biobanks as there is no common query interface. This makes catalogues like for example the Catalog of European Biobanks (http://bbmri.eu/catalog-of-european-biobanks) [1] and the BBMRI-ERIC Directory (http://bbmri-eric.eu/bbmri-eric-directory-1.0) difficult to update and maintain.

MIABIS 2.0 (Minimum Information About Biobank data Sharing) [2] [3] (https://github.com/MIABIS/miabis/wiki) is a standard that has been developed in response to the need of harmonization and standardization of Biobank information in Europe. Though MIABIS is widely accepted as the standard for biobank data exchange, most of the information from biobanks is not programmatically accessible in the MIABIS format because biobanks in Europe implemented their pipelines, database schemas and LIMS (Lab Information Management System) before the MIABIS standard was published (http://www.ncbi.nlm.nih.gov/pubmed/24849882). Consequently, although the wide adoption of MIABIS would be highly beneficial for the biobank and research communities, the necessary changes in the existing data model of individual biobanks would require a lot of effort.

In this deliverable we report on a pilot that provides a proof of concept for biobank interoperability. MIABIS Connect (https://github.com/MIABIS) is a software framework that facilitates sample discovery. Researchers can use a central service to search for samples in multiples biobanks while biobanks can expose data that can be shared.

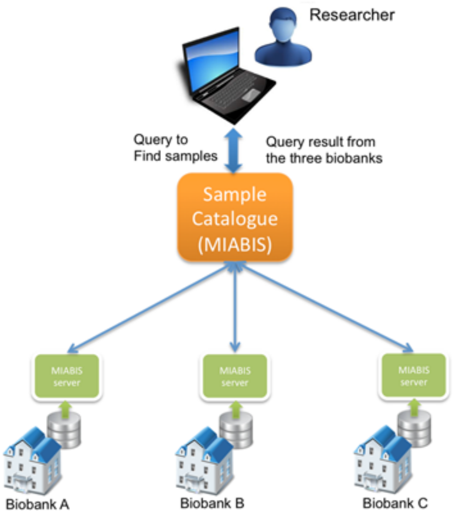


Figure 1 General architecture of MIABIS Connect

MIABIS Connect is a light open-source framework easy to install and use. From the biobank perspective, it requires a minimum of involvement from the biobanks IT support and doesn’t require modifications to the internal data model in biobank informatics management systems to compliant with MIABIS 2.0.

An important feature of this federation framework is that the biobank data never leaves the biobank boundaries. The biobank defines a data repository that MIABIS Connect uses for the federation. Furthermore, the biobank decides which data can be publicly exposed in the federation.

MIABIS Connect software framework is a solution to the difficulty of searching for the most valuable bio-resource in biobanks: the sample. The biobanks normally have their internal management systems for managing samples but not often expose their samples to the research community not only due to ethics and regulatory constraints, but also because there is not a standard software tool that allows exposing their resources without requiring a time-consuming process or economical investment.

MIABIS Connect is a solution to this problem.

3.2 MIABIS Connect

The functional modules of MIABIS Connect are *MIABIS Server* and *MIABIS Client*.

MIABIS Server

The MIABIS Server is a software package to be installed in a server located in the biobank. The biobank exports its data in separated files for samples, sample collections, studies and contact information. It can be done through querying the database and uploading the files to a file repository defined by the biobank.

Then, the file data elements (columns and field values) are mapped to MIABIS only once (or when the data model in the biobank has changed). The resulting map is used to create a de-normalized data file called MIABIS-TAB, which is then converted to an index ready to be queried from a web service.

The software components of MIABIS Server are:

- ***Mapper***: Java program (https://github.com/MIABIS/miabis-mapper) to map biobank data model to MIABIS. The Mapper “discovers” the structure of the data files to be exported and allows the user to map the relevant attributes against the MIABIS attributes. After the mapping process, a ***map file*** is saved for internal use in MIABIS server. The participating biobank uses the Mapper the first time when the data needs to be exported to the MIABIS Server. On the subsequent data exports, the mapping process need not be done and the MIABIS Server uses the existing ***map file***. However if there are changes to the internal data model in the biobank, which affect the format of the files to be exported, then the Mapper has to be used again to do the mapping and create a new ***map file*** reflecting these changes.

- ***Converter***: Java program that reads the biobank files together with the produced map and indexes the results in an instance of ElasticSearch. Alternatively, the tool is able to produce MIABIS-TAB, the MIABIS sample exchange format (https://github.com/MIABIS/miabis-sample-exchange-format).

- The Converter uses a Derby database to facilitate data integration.

- The Converter produces a MIABIS-TAB file (https://github.com/MIABIS/miabis-converter/wiki/MIABIS-SAMPLE-TAB-Format) or populates an ElasticSearch index for real-time search (https://www.elastic.co/products/elasticsearch).

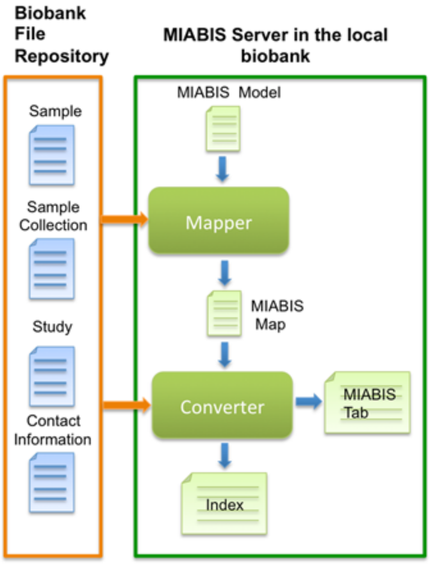
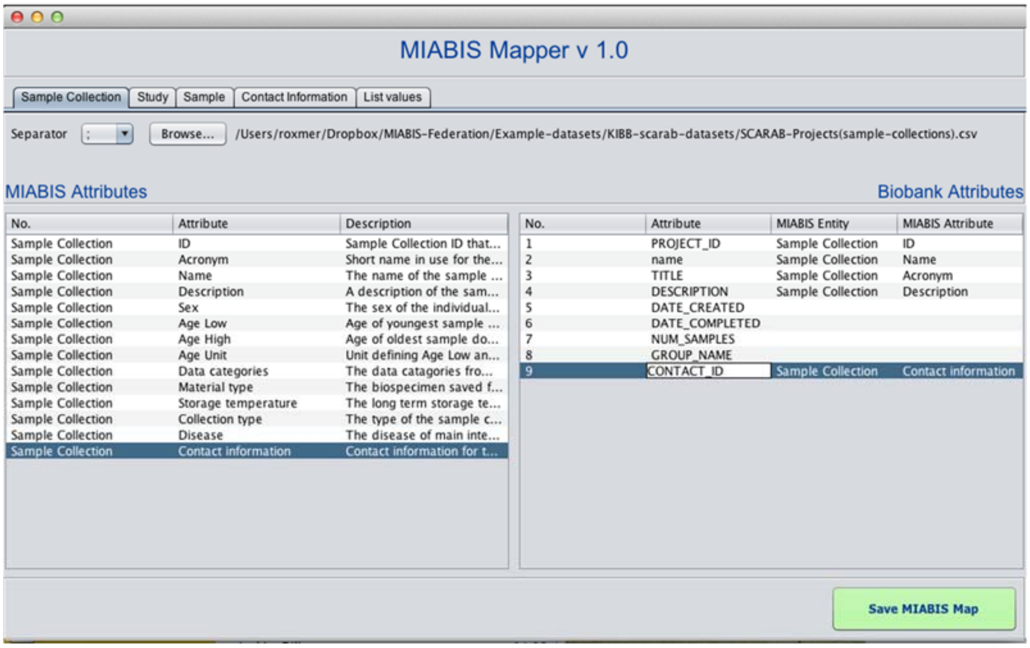


Figure 2 Server components. The biobank systematically uploads data to a file repository. Each time the internal semantic of the biobank changes, the data needs to be mapped to MIABIS using the Mapper. The produced map is used by the Converter to index the data with ElasticSearch for real-time searching from the MIABIS Client.



MIABIS Mapper v 1.0

Separator ; Browse... /Users/roxmer/Dropbox/MIABIS-Federation/Example-datasets/KIBB-scarab-datasets/SCARAB-Projects(sample-collections).csv

MIABIS Attributes

No.	Attribute	Description
Sample Collection	ID	Sample Collection ID that...
Sample Collection	Acronym	Short name in use for the...
Sample Collection	Name	The name of the sample ...
Sample Collection	Description	A description of the sam...
Sample Collection	Sex	The sex of the individual...
Sample Collection	Age Low	Age of youngest sample ...
Sample Collection	Age High	Age of oldest sample do...
Sample Collection	Age Unit	Unit defining Age Low an...
Sample Collection	Data categories	The data catagories fro...
Sample Collection	Material type	The biospecimen saved f...
Sample Collection	Storage temperature	The long term storage te...
Sample Collection	Collection type	The type of the sample c...
Sample Collection	Disease	The disease of main inte...
Sample Collection	Contact information	Contact information for t...

Biobank Attributes

No.	Attribute	MIABIS Entity	MIABIS Attribute
1	PROJECT_ID	Sample Collection	ID
2	name	Sample Collection	Name
3	TITLE	Sample Collection	Acronym
4	DESCRIPTION	Sample Collection	Description
5	DATE_CREATED		
6	DATE_COMPLETED		
7	NUM_SAMPLES		
8	GROUP_NAME		
9	CONTACT_ID	Sample Collection	Contact information

Figure 3 MIABIS Mapper. This tool allows biobanks to map their data elements to MIABIS the first time the biobank will enter into the federation and every time the biobank data model is changed

A batch process will be scheduled to index modified files from the Biobank File Repository.

MIABIS Client

The client is a web application through which the researchers can search for samples using a very flexible query interface based on Kibana [5], an open-source product from Elastic (https://www.elastic.co/products/kibana).

A centralized ElasticSearch instance (tribe node) [6] distributes the queries among MIABIS servers and presents the results in Kibana (as shown in figure 5).

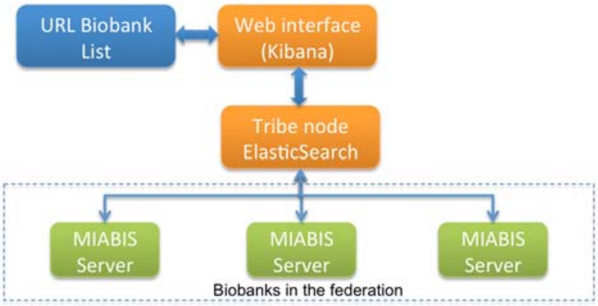
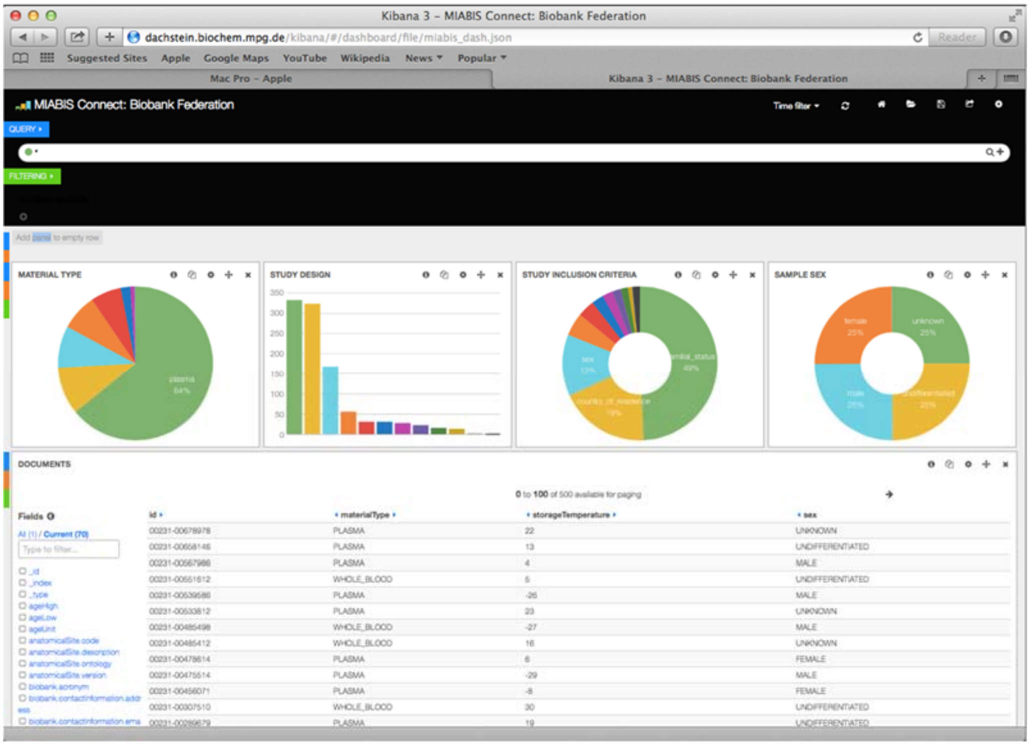


Figure 4 MIABIS Client (orange blocks). Uses a list of URLs from biobanks members of the federation. An instance of ElasticSearch queries the MIABIS servers.

A provisory URL for the MIABIS Client can be found at:

http://dachstein.biochem.mpg.de/kibana/#/dashboard/file/miabis_dash.json



id	materialType	storageTemperature	sex
00231-00678978	PLASMA	22	UNKNOWN
00231-00658146	PLASMA	13	UNDIFFERENTIATED
00231-00567986	PLASMA	4	MALE
00231-00551612	WHOLE_BLOOD	5	UNDIFFERENTIATED
00231-00539586	PLASMA	-26	MALE
00231-00533812	PLASMA	23	UNKNOWN
00231-00485498	WHOLE_BLOOD	-27	MALE
00231-00485412	WHOLE_BLOOD	16	UNKNOWN
00231-00478614	PLASMA	6	FEMALE
00231-00475514	PLASMA	-29	MALE
00231-00456071	PLASMA	-8	FEMALE
00231-00307510	WHOLE_BLOOD	30	UNDIFFERENTIATED
00231-00289679	PLASMA	19	UNDIFFERENTIATED

Figure 5 MIABIS Connect dashboard based on Kibana. The user uses free text or formatted queries to find samples among biobanks in the federation. This service provides a very flexible and rich interface.

The Pilot

- The MIABIS Server was installed in three sites: Karolinska Institutet, Max Planck Institute for Biochemistry and Elixir.

- Karolinska Institutet provided data from KI Biobank (http://ki.se/forskning/ki-biobank). Personal data about contact information was replaced by generated data. KI Biobank exported 1000 samples with associated data.

- Data from MPIB and Elixir was simulated.

- A total of 3500 samples were distributed among the three sites.

MIABIS Connect summary

- MIABIS Connect is a software for biobank federation using MIABIS standard as the central semantic https://github.com/MIABIS/miabis/wiki
- Sample-center data model: Information about Biobank, Sample Collection, Study and contact information related to a sample
- As part of this deliverable we have also designed a so-called **MIABIS-TAB**, a file format definition to describe samples in a file where a row represents a sample and columns are separated by tabulations. At the moment, this format contains sample data and its associated information related to biobank, sample collection and studies. This format can be improved or modified to associate to the sample other MIABIS components: https://github.com/MIABIS/miabis-converter/wiki/MIABIS-SAMPLE-TAB-Format
- MIABIS Connect can be extended to other bio-resources from biobank and research as biobank services, research experiment data, etc.
- This software framework demands minimum involvement of biobank staff or IT support
    - Installation of MIABIS Server
    - Mapping biobank data to MIABIS 2.0
- Biobanks can keep their idiosyncratic semantics and at the same time share sample data
- Once the biobank data is mapped to MIABIS, the biobanks only need to export delimited files from their databases to a data repository defined by the biobank. Biobank data never leaves the biobank
- Standalone solution using minimum software framework based on open-source software
    - ElasticSearch, Kibana from Elastic https://www.elastic.co/

  - Java programs for mapping and indexing https://github.com/MIABIS

- MIABIS Connect is an open-source software framework to be used by biobanks that want to share their resources, and for researchers that need a standardized and easy way to find bio-resources for their research through a common interface. In a more formal way, this software framework can be used by biobank networks to expose and discover bio-resources using MIABIS as central standard to represent biobank data.

- However, with very small changes, this software can be abstracted in a way that can be used to federate any domain that has a formal representation of data or a controlled vocabulary to integrate data among different data providers.

- All the produced software as well as the respective documentation is available on GitHub [7].

Future plans

- Define strategies and requirements to integrate MIABIS Connect with BioSamples (http://www.ebi.ac.uk/biosamples/).

- Implementation of a common REST-API that allows biobanks using Molgenis platform easily join the federation.

- Introduce MIABIS-TAB format to the biobank community and continue its definition.

- Implementation of a security layer in MIABIS Server with control access for files uploading and indexing.

- Improvement of the UI in the server side to facilitate installation and configuration through a control panel.

- In order to compliant with personal data protection issues, a pseudo-anonymisation module should be developed to help biobanks to replace local sample IDs by public sample IDs.

- Creation of a governance system that incorporates ethical and regulatory requirements for biobanks being part of the federation.

- Synergy with BBMRI-ERIC ELSI CS and BBMRI-ERIC Quality CS to implement federation policies.

- Use the BBMRI-ERIC Directory (http://bbmri-eric.eu/bbmri-eric-directory-1.0) as a registry to discover federated programmatic access instances provided by BioBanks. Open not just for MIABIS servers but other programmatic access points provided by frameworks like Molgenis".

- Develop further tools to visualize, analyze and interface with the biobank data that could be reused by the individual biobanks and their users (e.g., BioJS visualization components, galaxy workflows and format converters).

Sustainability plan

- Karolinska Institutet (BBMRI.se) will implement a biobank federation using the MIABIS Connect framework. A pilot at the national level will demonstrate the advantages of this software. It could be extended to the European level through BBMRI-ERIC Common Services IT (CS IT).

- The implementation of the pseudo-anonymization module for sample IDs will be part of the MIABIS Connect pilot in Sweden.

- The software will be maintained (in principle) by Karolinska Institutet (BBMRI.se) Elixir and Max Planck Institute and will be available on github (https://github.com/MIABIS). BBMRI-ERIC could also adopt this software as part of the BBMRI-ERIC CS IT and maintain and integrate new services to the federation.

- Another option is the possibility of sustainability through RD-Connect project. Karolinska and Groningen are part of this project. A federated rare diseases sample catalogue could be implemented using MIABIS Connect solution.

- MIABIS Connect will be a part of the B3Africa project, an EU funded 3 year project that will develop an informatics platform for EU-Africa biobanking and biomedical research collaboration (http://www.b3africa.org/).

3.3 Linking ICD10 to microRNAs

Since their discovery in 1994, miRNAs as a target are gaining attraction in the research community, both as serum biomarkers [8], and for their role in cancer [9] and disease progression [10]. The miRNA molecule acts as a suppressor of messenger RNA translation by physical interaction, and are thus vital to the regulation of biological processes in the cell.

To produce the mapping, we created an intermediate mapping between the ICD10 and Disease Ontology (DO). We followed a slightly different approach compared to the one described in D10.2. Since the submission of D10.2, a publicly available mapping for DO to ICD10 has become available. However, the mapping works in the direction of ICD10 to DO, and not DO to ICD10. We needed the latter. To exemplify, according to DO, diabetes mellitus (DM) (DOID:9351) corresponds to Type 2 diabetes mellitus (T2D) (ICD10:E11). This is indeed correct if going in the direction of ICD10 to DO: T2D is a more specific case of DM. However, what is not true is that DM is a more specific case of T2D. If we had just used the mapping as is, all of the miRNAs associated to DM would also be associated to T2D.

Disease Ontology-miRNA associations were retrieved from [11]. This included both text-mined and predicted associations. Looking at the number of associations across ICD10 chapters (Figure 6), seemingly neoplasms (chapter 2) and the nervous system (chapter 6) has the highest number of associated miRNAs. Compared to the number of gene associations, the number is very high, since there are 10 times more genes than miRNAs in the human genome. The figure also suggests that this mapping can be of great interest to study cancer, Encephalitis and Alzheimers (nervous system), and diabetes and metabolic disorder (endocrine and metabolic) progression in biobank samples.

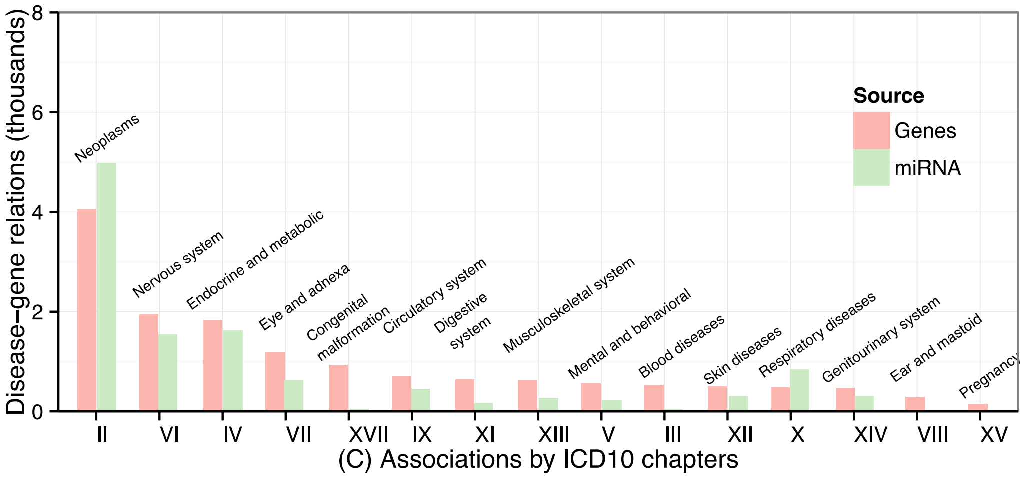


Figure 6 Associations of miRNAs to ICD10 diagnosis codes over chapter, compared to Genes

4 References

[1] Comprehensive catalog of European biobanks. *Nature Biotechnology* **29**, 795–797 (2011) doi:10.1038/nbt.1958

[2] A Minimum Data Set for Sharing Biobank Samples, Information, and Data: MIABIS. BIOPRESERVATION AND BIOBANKING, Volume 10, Number 4, 2012, DOI: 10.1089/bio.2012.0003

[3] Developing a semantically rich ontology for the biobank-administration domain. *Journal of Biomedical Semantics* 2013, **4**:23, doi:10.1186/2041-1480-4-23

[4] Testing Automation for Distributed Applications. https://www.elastic.co/pdf/elastic-white-paper-testing-automation-elasticsearch.pdf

[5] Kibana: https://www.elastic.co/products/kibana?camp=gaw&gclid=Cj0KEQiA-NqyBRC905irsrLr-LUBEiQAWJFYThEDBV-MJNX6g2Hq2zniGD4IeZFVhSfBGOudJyto9IgaAow_8P8HAQ

[6] Tribe Node: https://www.elastic.co/guide/en/elasticsearch/reference/1.4/modules-tribe.html

[7] MIABIS Connect software framework: https://github.com/MIABIS

[8] Gilad S, Meiri E, Yogev Y, Benjamin S, Lebanony D, et al. (2008) Serum MicroRNAs Are Promising Novel Biomarkers. *PLoS ONE* **3(9)**: e3148. doi: 10.1371/journal.pone.0003148

[9] Esquela-Kerscher, Aurora, and Frank J. Slack. Oncomirs—microRNAs with a role in cancer. *Nature Reviews Cancer* **6.4** (2006): 259-269.

[10] Calin, George Adrian, et al. A MicroRNA signature associated with prognosis and progression in chronic lymphocytic leukemia. *New England Journal of Medicine* **353.17** (2005): 1793-1801.

[11] Mørk, Søren, et al. Protein-driven inference of miRNA–disease associations. *Bioinformatics* (2013): btt677.

5 Supplementary information

ElasticSearch: https://github.com/elastic/elasticsearch

Kibana: https://github.com/elastic/kibana

MIABIS Connect:

MIABIS wiki https://github.com/MIABIS/miabis/wiki

MIABIS Java programs for mapping and indexing**:** https://github.com/MIABIS

MIABIS Connect Tools: http://www.biomedbridges.eu/simple-and-versatile-biobank-sample-information-federation-miabis-connect

6 Delivery and schedule

The delivery is delayed: No

7 Adjustments made

The original objective of the task reported in this deliverable was to demonstrate the interoperability of the sample representation at the EMBL-EBI's BioSamples database, with the biobank databases participating in the federated biobank information infrastructure created under BBMRI. The first adjustment to this plan was reported in D10.1: instead of mapping individual biobanks into BioSamples database we produced a mapping between MIABIS (a template defining how biobanks should describe their holdings) and BioSamples database, a result that is applicable to a wider range of BBMRI data sources.

The two problems that BioMedBridges, and WP10 in particular, were addressing in the area of biomedical sample information were: discovery of biosamples, and accessing detailed information about those samples. The plan in WP10 was to depend on the data security framework as developed in WP5. The work of WP5 indeed resulted in a pilot connecting biobank data with the BBMRI catalogue and the BioSamples database, across all information access tiers as considered in WP5: open, restricted, and committee controlled. The security framework incorporates a system for facilitating granting access to committee controlled data (REMS – see D5.4 for details), but does not cater for a federated data access model, or information discoverability. In the biobanking world information federation is crucial, since it gives the data custodians better control over their holdings. Therefore the following adjustments were made to the work that resulted in this deliverable:

- The lead beneficiary for the task was Karolinska Institute (partner #3) instead of EMBL (partner #1), developing a federated query interface for discovering sample information;
- Interface of the BioSamples database to the BBMRI infrastructure is demonstrated in deliverable D5.4.

Another adjustment was the inclusion of new work performed by UCPH in the area of linking disease terminology to the genome, i.e., microRNAs. Thereby, in this deliverable we demonstrate a long information bridge “Samples in BBMRI - Samples in ELIXIR - Genome in ELIXIR”, across infrastructures and data types.

8 Background information

This deliverable relates to WP 10; background information on this WP as originally indicated in the description of work (DoW) is included below.

WP 10 Title: Integrating disease related data and terminology from samples of
different types
Lead: Alvis Brazma (EMBL)
Participants: EMBL, KI, UCPH

This work package will demonstrate the feasibility and provide a prototype for linking disease to molecular information on two levels – terminology and data. It has two tasks respectively – first to link ICD10 terms to gens and protein complexes, and second to link data in selected BBMRI biobanks to samples at the EBI BioSample Database supported by ELIXIR.

Both tasks are related, as to link biobanks to ELIXIR databases, the respective terminologies have to be mapped. For completing the first task we will create a prototype for an interoperable scheme linking ICD10 to genes and protein complexes. This will enable linking biobank and other phenotypic healthcare sector data on individuals to individual genotypes. Interoperability schemes of this kind will be essential for linking BBMRI and ECRIN data to the molecular level and for linking biobank and other phenotypic healthcare sector data on individuals to individual genotypes.

To accomplish the second task, we will select a small number of biobanks participating in the BBMRI infrastructure, with best advanced sample representation in databases. We will develop a model for linking sample objects to the respective objects at the BioSample database developed at EBI. The final deliverable in this work-package will be implementation of these links, allowing user to navigate from the selected biobanks to the EBI BioSample database and vice versa.

The work package will build upon standards developed in WP3, will utilize infrastructure build in WP4 and will benefit from data security framework developed in WP5.

Work package number	WP10	**Start date or starting event:**	month 13
Work package title	Integrating disease related data and terminology from samples of different types		
Activity Type	RTD		

Participant number	1:EMBL	3:KI	15: UCPH
Person-months per participant	26	20	31

Objectives

1. Linking disease-related data to molecular information: terminology
2. Linking disease-related data to molecular information: data.

Description of work and role of participants

Task 1. Mapping between sample information representation in a selected subset of resources. We will map data elements describing sample information in selected biobank databases that participate in the BBMRI federated infrastructure and the BioSample Database at EMBL-EBI. We will work jointly with WP3 to generalize the defined mappings and to develop the minimum standard that would enable to exchange this information.

Task 2. The aim is to create a prototype which maps existing healthcare sector terminologies and their phenotypic descriptions to existing repositories linking diseases, symptoms and genes. More generally the aim is to embed the prototype efficiently into exiting computational linguistics, bioinformatics and clinical environments alike. In particular we will link phenotypic terminology in healthcare sector data to genes we will create a prototype for an interoperable scheme linking ICD9 an ICD10 to gene identifiers in ELIXR database concentrating on genes that have been associated with specific diseases, symptoms and tissues. The prototype will focus specifically on the generic ICD concepts and their mapping to relevant genetic information, and will not include efforts which aim for assigning codes and terminology to free text in healthcare sector data as it usually is done by text mining. The work on the prototype will also include language interoperability aspects on the terminology side, while considering only gene names as they are used in English.

Task 3. The objective of this task is to demonstrate the interoperability of the sample representation at the EMBL-EBI's BioSample database, with the biobank databases participating in the federated biobank information infrastructure created under BBMRI. This will also build upon the work done on terminology and identifier mapping. Working together with WP4 and WP5, we will implement a pilot for federated queries and secure links between a limited number of selected resources in BBMRI and the BioSample Database at EMBL-EBI (ELIXIR). A query system will be developed that will allow sample property based queries across these resources.