D8.1 Requirements specification for data harmonisation and terminology mapping tools

BBMRI-ERIC CS-IT Project: D8.1

Requirements specification for data harmonisation and terminology mapping tools

Authors: Niina Eklund, Sebastian Mate, Christina Schüttler, Timo Miettinen, Juha Knuuttila, Hans-Ulrich Prokosch, Kaisa Silander

1 Introduction

BBMRI-ERIC aims to facilitate the joint establishment and operation of research infrastructures of European interest that will provide access to the collections of partner biobanks and biomolecular resources, their expertise and services on a non-economic basis. The mission of BBMRI-ERIC CS-IT project, which started in January 2016, is to deliver expertise, services, and tools relevant to the pursuance of the tasks and activities of BBMRI-ERIC. The project work is divided into 8 work packages (WPs), with WP8 responsible for providing specifications and tools for facilitating collaboration on Data Harmonization, mainly required by WP1 (Directory service) and WP2 (Sample locator).

The aim of this document is to define initial requirement specifications for common terminologies and reporting data sets to support CS-IT architecture plan (Alexandre et al. 2016), based on input from European biobanks and national nodes. The input was provided mainly through two separate questionnaires, one directed to European Biobanks listed in the BBMRI-ERIC Directory, the other directed to National Node Coordinators. In addition, follow-up questionnaires were sent to selected biobanks from each BBMRI-ERIC member country. The results of these questionnaires were considered together with information from additional sources, such as Sample Finder webinars[1], BBMRI-ERIC Use Cases document (Holub et al. 2016) and State of the Art Review of Harmonization Tools (Mate et al. 2016) in order to provide the initial requirement specifications.

1 https://docs.google.com/document/d/1qmZn36M4MEGZuVwPLEs7eRMGuTjANZLVWunW7rlJVXo

2 BBMRI-ERIC CS-IT Planned Architecture

The CS-IT Services planned for BBMRI-ERIC are described in detail in the CS-IT Architecture Notebook (Alexandre et al. 2016). The different components of this architecture have been defined in the BBMRI-ERIC CS-IT plan and the architecture notebook. Below is a short description of

the major components, to aid in reading this document without need to refer to other documents:

BBMRI-ERIC Directory: Catalog-like solution, includes summary level/anonymous information about biobanks and biobank collections/samples, and provides a single access point to the European Biobank Network. Currently, the Directory contains basic information on National Nodes, biobanks and sample collections hosted by biobanks. The information is based on MIABIS Core version 2.0 attributes (Merino-Martinez et al. 2016). An exercise[2] to map MIABIS attributes used in the directory to UMLS attributes has also been done.

Sample Locator: Enables researchers to locate specific individual samples that are fit-for-purpose, i.e. provides availability information for samples and sample/subject-related data from the biobanks to the researchers, while complying with data protection regulations. The Sample locator's federated search tool accesses the different connectors of individual biobanks to execute the queries and provides a summary reply to the researcher that includes all queried biobanks. The federated search concept allows the biobankers, or contributors in general, to retain full control of what data they provide as a response to each individual query, albeit this can be also automated, e.g. using rules what data are returned back without manual approval.

Biobank Connector (data warehouse): A software interface that allows biobanks to connect to the CS-IT Services (=sample locator) while still maintaining full control over their samples and sample/subject-related data as well as who can access the biobank's summary-level data (=query result). For security purposes, it will also log any activity related to the above.

Metadata Repository (MDR): A central database that contains metadata definitions on all the elements (=namespaces) that are used in the Sample Locator and Biobank Connector. There is only one central MDR which has uniform metadata format. Each local connector is linked to the central MDR. Each biobank can have its own private MDR space, if needed, for its own metadata, while the central space will hold all harmonized elements from all connectors.

Negotiator: A communication tool meant to simplify the 1:N communication between a requester and multiple biobanks, to refine sample and/or data availability queries from biobanks, based either on data available in the Directory or in the Sample Locator.

"Harmonization" is a vast concept and it can have many different meanings. The scope of harmonization defined in this requirement document deals with harmonizing a selected set of pre-defined variables that describe biobanks, biobank sample collections, biobank samples and sample-related and sample donor-related data between European biobanks to support joint queries based on this selected data set across all biobanks.

Ontology mapping and harmonization tool/s: Software tool/s that allow/s mapping of local elements agreed upon as BBMRI-ERIC reporting data sets (compulsory minimal datasets and optional case-specific datasets), known ontologies and standards. The tools could also support comparison of attributes between two different sample collections or data sets and their harmonization into unified harmonized elements to promote the use of different biobank collections in the same study. The mapping/harmonization process includes the following steps: definition of similarity level between elements, selection of best match, transformation rules, renaming of elements, supporting language versions, etc. The harmonization tools should be provided as part of the MDR-Connector schema.

For more information about the planned architecture, please refer to the CS-IT Architecture Notebook (Alexandre et al. 2016).

An important aspect that has to be taken into account when planning BBMRI-ERIC's common infrastructure services is that all the processes will aim at using open source tools easy to customize and to fit into the biobanks' existing databases. In addition, the local tools provided by BBMRI-ERIC to biobanks should be easy to install, maintain and use, and proper user support should be available.

[2] MIABIS-Directory-UMLS mapping

3 BBMRI-ERIC Network Biobanks' Use Cases

The intended outcome of this work was to form reasonable definitions for harmonization and ontology mapping issues relevant to European biobanks, which can be used as a basis of the architecture development. While forming the definitions, there is also a need to focus on the different groups of users: biobanks, researchers, the academy, governments and industry. These groups have very different needs regarding the use of biobank samples and sample related data. The basic description of the main user groups of BBMRI-ERIC ICT-services has been detailed in the BBMRI-ERIC Use Cases -document (Holub et al. 2016) and includes the following: Bio/Medical researchers, Data driven researchers, Industrial R&D and Biobankers.

It is evident that the capacity of biobanks has to be kept in mind: adopting and sustaining of the ICT-systems developed needs to be done within their own available resources: time, budget and capable personnel. Binding of biobanks to the development process and to the use of harmonization mechanisms to be developed should be done by incorporating them as stakeholders into the requirement definition work and by ensuring that they will benefit from the developed CS-IT services.

The European Biobanks' use cases for requirement definitions were formed by conducting surveys of BBMRI-ERIC network biobanks and BBMRI National Nodes.

3.1 Survey of BBMRI-ERIC Network Biobanks

This section describes in details the survey we performed of BBMRI-ERIC Network Biobanks and the survey results. A questionnaire regarding availability of samples and related data and harmonization processes done at the biobank was sent to all biobanks listed in BBMRI-ERIC's Directory.

3.1.1 Motivation & Methods

The motivation for conducting the survey described here was to chart the current level and status of biobank sample and data harmonization: what sample-related and subject-related data do biobanks have, do biobanks harmonize their data in some way, what ontologies they use, and what depth is used in data harmonization. The main aim was to chart the user experiences and collect user stories to aid in defining the initial specifications of BBMRI-ERIC CS-IT common ontology/ies and reporting data set/s to support data availability services of European biobanks.

This survey research was executed with quantitative methods. Convenience sampling was used by sending all biobanks listed in BBMRI-ERIC Directory an e-mail with a link to a web-based questionnaire. The questionnaire was built using Webropol[3] e-survey and reporting tool and it was structured using 2 questions covering contact information of biobank and harmonization expert, 6 open questions and 2 multiple-choice questions related to biobank's sample collections and data harmonization. All questions were optional and could be answered if the question was applicable to the respondent's biobank.

The biobank sample and data harmonization questionnaire can be found in appendix 1.

[3] www.webropol.com

3.1.2 Analyzing Survey Results

The survey was sent to 389 biobanks listed in the BBMRI-ERIC Directory 2.0[4], as well as four biobanks outside of Directory in order to capture use cases from all BBMRI-ERIC member countries. Reply time was limited to one month, but then extended by 3 weeks to obtain better representation. However there was an initial loss of 28 (7.2 % of sample) biobanks in the sampling, because of invalid contact email addresses. From this convenience sampling we received 84 replies regarding 85 biobanks/research collections, thus the unadjusted answer rate was 21.6 %, and answer rate adjusted with loss was 23.3 %.

The survey results were analyzed with Webropol reporting tools and MS Excel-based summaries as well as proper study and discussion of each reply.

[4] http://old.bbmri-eric.eu/bbmri-eric-directory-2.0

3.2 Interview of Selected Biobanks

Based on the initial questionnaire answers, 1-2 biobanks or research groups were selected per country for more detailed interviews. Also some additional biobanks or research groups that did not reply to the web-based questionnaire were sent the interview request. The additional interview selections were made based on the author team's previous knowledge of ongoing harmonization efforts in the field of biobanking.

3.2.1 Motivation & Methods

The motivation for interviewing selected biobanks was to learn in greater detail what harmonization tools are used, how they are implemented and if they would be potentially useful to BBMRI-ERIC CS-IT. The interviewee selections were made based on the use of some harmonization tools or initial efforts to manually harmonize biobank samples and data in some way. 1-2 biobanks or research groups per country were selected as a use case for more explicit interview.

Some biobanks had otherwise interesting replies to the initial harmonization questionnaire, but they reported that they didn't have any tools in harmonization or they are not doing any sample or data harmonization. For these biobanks, a customized interview was made to chart out the reasons why no harmonization efforts were made. Both interviews and their questions are presented in Appendix 2a and 2b.

The interviews were mainly performed via e-mail. The interview requests were sent to selected biobank's or research groups harmonization or IT-expert named in the initial harmonization questionnaire. The interview requests included the biobank's questionnaire replies and the interview form attached. Only one interview was conducted face-to-face when piloting the interview questions. Small alterations to the interview questions and to question ordering were made after this initial pilot.

3.2.2 Analyzing Interview Results

Interview requests were sent to total of 21 biobanks or research groups. In total 14 responses were received, making the answer rate 66.7 %.

Since most of the interviews had to be customized according to the questionnaire replies in order to capture the full functionality of the harmonization processes and tools used in the biobanks in question, the analysis of interview replies wasn't a fully straightforward procedure. The interviews were transcribed and each question was thoroughly analyzed, summarizing the data in MS Excel. Also a study and discussion for each reply was conducted.

3.3 Survey of BBMRI-ERIC National Nodes

A separate survey to chart national sample and data availability services was sent to all BBMRI-ERIC National Node Coordinators. This section

covers the questionnaire sent to National Nodes and the results of the survey.

In addition to conducting a survey of availability services in BBMRI-ERIC National Nodes, several biobanks and biobank networks have been invited to present their own sample locator/catalog tools and/or availability services in series of webinars[1] held during January and February 2016.

3.3.1 Motivation & Methods

Motivation to execute a survey of National Node availability services was to find out if some national efforts around the availability biobank/research collection samples and associated data are already ongoing. These already existing or going to be developed services could also be seen as examples to learn from, especially when considering what level of information about available samples and data can be shared publicly and what should be given only to registered users and with biobanks' approval.

This survey was executed in similar way as the harmonization questionnaire for biobanks and by using quantitative methods. Convenience sample was formed sending all BBMRI-ERIC National Node coordinators an email with a link to a web-based questionnaire. The questionnaire was built using Webropol[3] e-survey and reporting tool. The questionnaire was structured into 5 open questions, one of them asking them from which National Node the respondent is. All questions in this questionnaire were optional and could be answered if the question was applicable in the respondent's Node.

The questionnaire about National Node availability services can be found in appendix 3.

3.3.2 Analyzing Results

The electronic questionnaire link was sent to all 14 BBMRI-ERIC National Node Coordinators as well as to all 5 observer members. The reply period was limited to one month, but later extended by two weeks to obtain better representation. In total we received 9 replies to the questionnaire regarding the National Node availability services, making the answer rate 47.4 %.

The survey results were analyzed with Webropol reporting tools and MS Excel-based summaries as well as proper study and discussion of each reply.

3.4 Survey Results

The summary of surveys and interview results are presented in this section. The results are divided by the analysis question in mind.

3.4.1 Survey of BBMRI-ERIC Network Biobank sample and data harmonization

We received in total 84 replies to the biobank/research collection harmonization questionnaire regarding 85 biobanks/research collections.

We managed to capture replies from all BBMRI-ERIC member countries, so the survey result summaries can be considered representative. Figure 1 shows the replies received by countries.

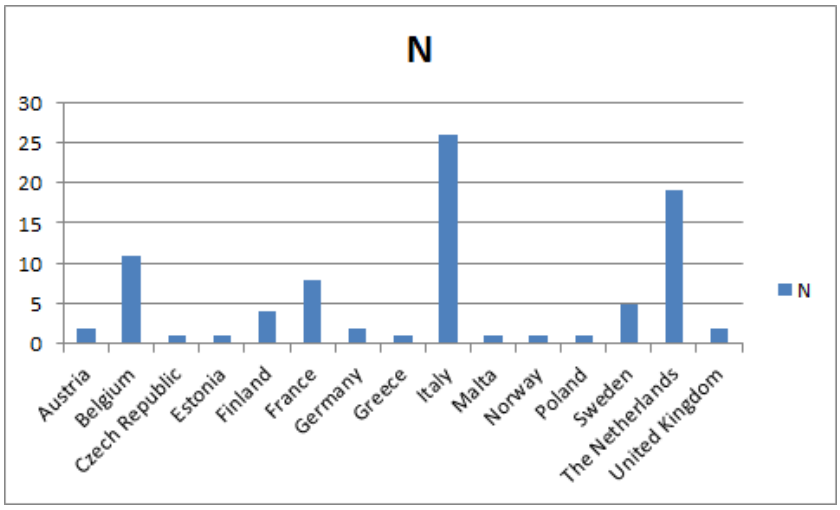

Figure 1. Harmonization questionnaire replies are divided by biobank/research collection country.

Using the BBMRI-ERIC Directory 2.0 as a reference, we divided the replies in four main categories based on whether the respondent biobank/research collection was classified as (having) clinical collection, population collection, research-study collection or non-human collection. The classification is presented in table 1.

Table 1. Biobanks/research collections divided in four major classes according to their main collections. **NB.** There are few biobanks that have collections fitting multiple categories.

Biobank type	N
Hospital biobanks/collections and cancer-related biobanks/collections	58
Population-based/survey/epidemiological biobanks/collections	17
Research biobanks/collections, a general term for all kind of research-based collections such as disease-specific family collections (rare and more common diseases), case-control studies, clinical trial studies, etc.	10
Non-human biobanks/collections, such as veterinary biobanks, microbial collections, plant biobanks, etc.	4

The biobanks that reported having clinical collection(s) or those that were associated with a hospital were classified as clinical biobanks (N=58). In a similar way, biobanks that reported to have samples from population were classified as population-based biobanks (N=17). Research study collection -category however is more complex: all case-control based, disease based or family collections were classified in this category (N=10), since they don't fit neatly either the clinical or population based biobank category. Non-human collection category holds all collections in which the sample origin organism isn't human (N=4). Biobanks could fit into several different categories, thus making the combined N=89.

We performed this classification to see what are the main differences between biobank types. By doing this differentiation we can strengthen the common ground between biobank types and see if biobank type-specific services or tools are needed.

3.4.1 Available Samples and sample related data in BBMRI-ERIC Network Biobanks

In the harmonization survey we had asked the biobanks what type of sample related data do they store.

Table 2. Available samples in BBMRI-ERIC Network Biobanks

	Sample type mentioned (% of certain type of biobank)									
Biobank type	**DNA**	**RNA**	**Serum**	**Plasma**	**Blood**	**Cell**	**Tissue**	**Tumour**	**Saliva**	**Urine**
Clinical biobank	31.0	19.0	23.8	17.2	27.6	17.2	24.1	19.0	1	2
Population biobank	23.5	11.8	23.5	17.7	11.8	5.9	5.9	0	1	2
Research study collection	50.0	20.0	30.0	30.0	30.0	30.0	0	0	0	2
Non-human biobank	0	0	0	25.0	0	50.0	75.0	50.0	0	0

Table 2 shows what sample types the biobanks have information about. It is important to note that all biobanks didn't report what sample types they collect and store since it wasn't directly asked in the questionnaire. We intentionally left this question out because the available sample types should be listed under each biobank in BBMRI-ERIC Directory 2.0. This exercise was done to see if there are differences between sample types collected between different biobank types. As can be seen in table 2, there are indeed few such examples: tumor samples can be found in clinical biobanks (19 % of clinical biobanks reported to have sample of this type) and non-human (veterinary) biobanks (2/4 of non-human biobanks reported to have sample of this type), but they were absent in population biobanks and research collections, also tissue samples seem to be more common in clinical and non-human biobanks (24,1 % and 75 % consecutively) compared to population-based biobanks and research collections (5,9 % and 0% respectively).

3.4.2 Available Sample Donor-related Data in BBMRI-ERIC Network Biobanks

We wanted to find out what kind of sample-donor related data the biobanks have collected (for the questionnaire questions see question 3 in appendix 1). We asked if some common categories of data collected from sample donors would be available in biobanks. The data categories of interest were presented under three larger categories, namely Medical data: a type of data collected in hospitals and from hospital databases, General information: data that is usually collected via questionnaires, interviews or laboratory measurements made in research laboratory and Other data: data that doesn't belong to the above-mentioned categories.

Table 3. Comparison of biobank types and availability of different types of sample donor-related data categories. **NB.** Some biobanks/research collections may have replied to overlapping categories. The percentages represent % of biobanks of certain type, not % of all biobanks.

	Medical data					General information					Other		
Biobank type	**Medical history (%)**	**Diagnoses (%)**	**Medication (%)**	**Medical procedures (%)**	**Laboratory values (%)**	**Demographics and other background information (%)**	**Lifestyle and environmental information (%)**	**Findings (%)**	**Laboratory values (%)**	**Medication (%)**	**Data from official registries (%)**	**Other (%)**

Biobank clinical	79,3	89,7	63,8	65,5	72,4	55,2	39,7	17,2	56,9	43,1	29,3	12,1
Biobank population	82,4	76,5	70,6	41,2	58,8	82,4	88,2	47,1	70,6	52,9	70,6	29,4
Research collection	80	80	80	20	60	90	80	40	70	50	40	20
Biobank non-human	50	50	50	50	50	25	25	0	50	25	0	50

Categories of available biobank data presented in table 3 are also presented in figure 2 that demonstrates the percentages of biobank types that have a certain data category stored in their biobank. From this figure we can see that different types of biobanks tend to collect different types of data: this division seen in active collection of different data categories translates to that hospital biobanks/collections and cancer-related biobanks/collections usually have information on tissue type, diagnosis, treatment and treatment response, disease-related medication, disease-related biomarkers, etc. Population-based biobanks/collections usually contain background information on the sample donor, such as geographic area, lifestyle-related attributes (such as smoking, alcohol use, exercise, diet), other questionnaire data (such as health status, well-being, mental health, sleep, etc.), anthropometrics, serum/plasma measurements (such as lipids, infection biomarkers), genomic data, etc. Research biobanks and non-human biobanks may contain any attributes that are relevant for the specific research topic or type of collections, and are therefore very varied. The table 3 in chapter 3.4.2 shows comparison of these types of biobanks/sample collections and availability of different types of data items in the surveyed biobanks.

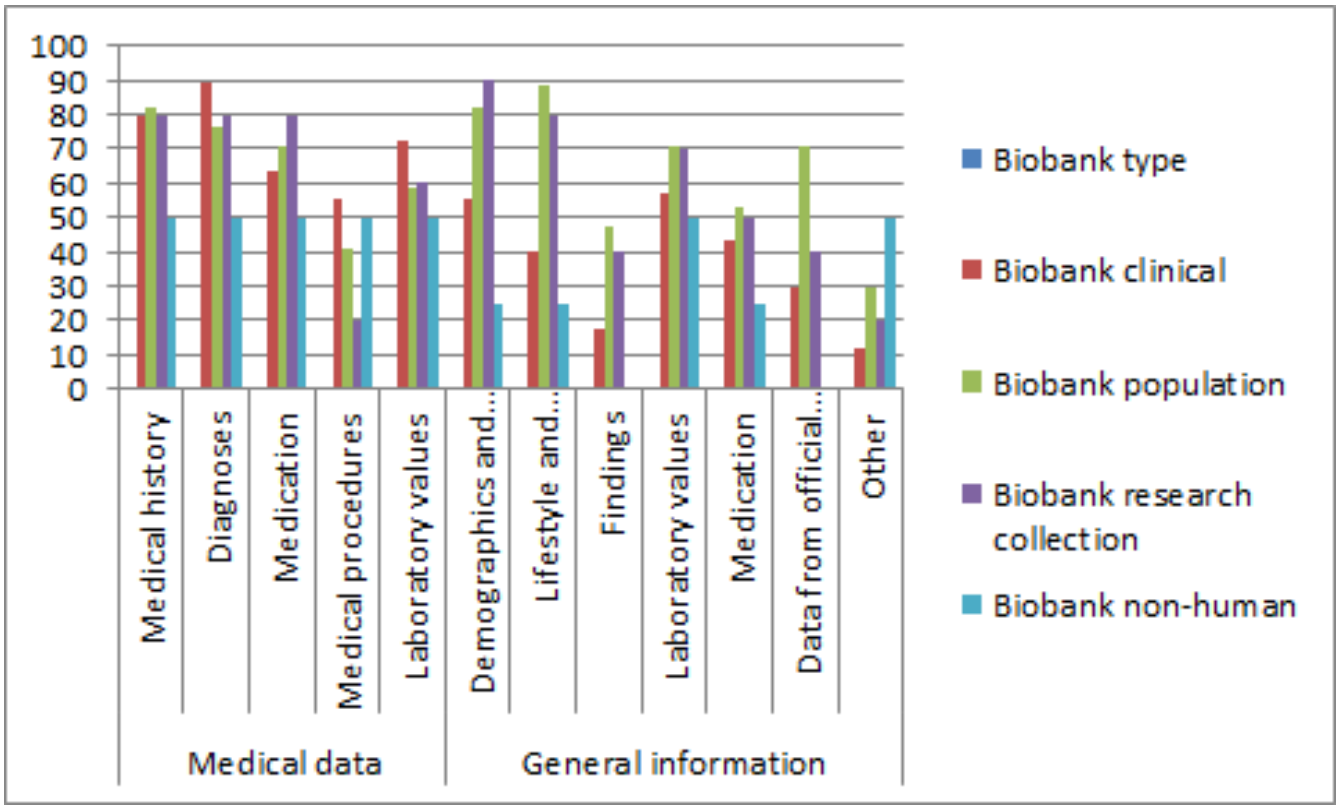

Figure 2. shows the percentage of certain types of biobanks that collect certain data categories.

Even though a certain number of all biobanks seem to collect data from almost all categories, there are according to the biobanks' replies to the harmonization questionnaire only very few attributes that are in common across all biobanks (sample type, age at sampling, gender). This relates to both sample-related and sample donor-related data.

3.4.3 Harmonization processes in BBMRI-ERIC Network Biobanks

With the harmonization questionnaire we tried to capture the current situation of sample and sample-donor related data in BBMRI-ERIC Network biobanks. Summarizing the results showed that there is already a number of harmonization activities ongoing in the field, even though the processes and their aims are very heterogeneous.

Any experience with harmonization activities or plans starting harmonization processes

Table 4. Biobank replies regarding their experience with harmonization activities.

Response option	N	%
Experience in harmonization activities	54	64.3
No experience in harmonization activities	27	32.1
Harmonization process in progress	3	3.6

When asked about their harmonization activities, more than a half of the questioned biobanks answered that they have some kind of harmonization process (table 4.). However, the approach varies. The most common procedures are to collect the data in one single database (n = 7) or to manage their data by using specific IT tools and software (n = 6). Besides IT related solutions, the biobanks work with standard operating procedures (SOPs) (n = 5), specific scripts (n = 4), guidelines and standards (n = 4) as well as templates in order to facilitate the harmonization process. Other types of data harmonization processes include the implementation of institutional platforms, the extraction of data from the HIS, and manual work. At the time we carried out this survey, three biobanks were in their planning phase of implementing harmonization processes or at least were aware of the necessity to take up harmonization activities.

Harmonization processes can be either manual or automatized using IT-tools

Table 5. Biobank replies regarding the handling of their harmonization process.

Response option	N	%
Biobanks that handle the harmonization process manually	26	31
Biobanks that handle the harmonization process manually as well as automatized (semi-automated processes)	12	14.3
Biobanks that handle the harmonization process automatized	8	9.5
Other methods	7	8.3
NA	31	36.9

The way biobanks handle their harmonization process is quite heterogeneous, as can be seen in table 5. In general, there are three different approaches to harmonization: Manually, automatized, and a method combining both. Most of the interviewed biobanks handle the harmonization process manually, at least in local level. Some biobanks mentioned that an automatized harmonization process takes place on national level (e.g. in Belgium). The majority of biobanks, whose process is partly manual and partly automatized, have automated processes to retrieve data (e.g. demographic data, patient ID) from a HIS or a databank, so they only have to add specific data manually. Moreover, the data verification is mostly done manually. Only a few biobanks stated that they have a fully automated harmonization process. One of them described its process in more detail and named Perl, PHP, PostgreSQL and web interfaces as development tools.

IT-tools used in harmonization processes are in-house developed, open source or commercial softwares

The usage of harmonization tools in biobanks is moderate. About half of the biobanks/research collections that replied to the questionnaire mentioned one or more tools that they use within the harmonization process. It is noteworthy that there is not any specific more widely used tool that would be preferred by biobanks. Regarding commercial tools, especially software like MS Excel, MS Access and SPSS was mentioned. Moreover, some responders indicated biobank or database software and information systems as well as solutions like Nautilus LIMS™ (Thermo Scientific), CRESALYS® (Excilone), DataBiotec (Oriam), KaliLab (Netika), DIAMIC (INFOLOGIC-SANTE) as well as WinlabWeb (Tesi) to be in use. The second most common type of tools are open source software, namely BiobankConnect (Molgenis), PostgreSQL, Perl, PHP scripts, MySQL, R, an ICDO-3 to ICD10 converter tool from IARC (International Agency for Research on Cancer), tools provided by the Belgian Virtual Tumourbank or the Cineca consortium (Italy), HyperCLDB (Cell Line Data Base, Interlab Project)) as well as other web-based tools. A few biobanks have developed in-house solutions to meet their need regarding the handling of their biobank processes, while some are still in the process of planning to implement a specific harmonization tool.

3.4.4 Common terminologies, ontologies and guidelines used in BBMRI-ERIC Network Biobanks

The harmonization processes require the biobanks to use some common standards and guidelines. In the survey we asked the biobanks (N= 57, see table 4) who reported having doing harmonization what type of standards, guidelines, reference terminologies and ontologies the biobanks have applied to their sample and data collection procedures. As expected, most biobanks who are doing data harmonization are applying some specific standards, SOP's and reference terminologies/ontologies to sample and data processing. We also expected correctly the most popular standards/guidelines to be ISO Standards, OECD and ISBER Best Practices, reference terminologies and ontologies used to be ICD-10 and it's previous versions, MeSH, and SNOMED. In addition to those options already named in the questionnaire, biobanks reported using some other specific terminologies/ontologies/guidelines/standards. Clinical biobanks named using (or planning to use) SPREC 2.0, TNM, HGVS and ISCN-2013, non-human biobanks mentioned OIE standard requirements and one population biobank stated that they may have their own classificators as well as national classificators from EMR. Also one research-collection reported that they handle their data so that it is HL7 compliant.

Table 6. Are there common standards, guidelines, ontologies and terminologies used in BBMRI-ERIC Network Biobanks?

Response option	N	%
Biobanks reported using standards and/or guidelines (i.e. ISO Standards, OECD and ISBER Best Practices)	39	68.4

Biobanks reported using reference terminologies and ontologies (i.e. SNOMED CT, ICD-10, MeSH)	49	86.0
Biobanks reported using other specific reference terminologies/ontologies/guidelines/standards in the harmonization process?	16	28.1

A more detailed table describing some of the key ontologies that are used in the European biobanks can be found in section 4.4 Ontologies in common use at European Biobanks.

3.4.5 Availability Services in BBMRI-ERIC Network Biobanks

The survey of BBMRI-ERIC National Nodes was conducted to find out, if some national efforts around the availability biobank/research collection samples and associated data are already ongoing, and to learn especially what level of information about available samples and data can be shared publicly and what should be given only to registered users an with biobanks' approval.

National Node engaged in sample locator/catalogue development

According to the 9 replies to the BBMRI National Node Coordinators' questionnaire about National Availability Services, there are already ongoing efforts regarding the development of national sample locators or catalogues (table 7). Seven National Nodes have already set up national catalogues and one Node is planning to set up one soon. Only one National Node is not engaged at the moment with setting up availability services.

Table 7. BBMRI-ERIC National Nodes sample locators/catalogs available.

National sample locator/catalog service(s) in your country	N	%
Yes	7	77.8
No	1	11.1
Planning in process	1	11.1

Many of these catalogues are presentations of biobanks or sample collections within their country's biobanks. It is important to note that one Node could have multiple locators/catalogs that would serve the whole Network or a certain sub-section of it, i.e certain type of biobanks (e.g. clinical biobanks). Two Nodes reported that in their catalogues only a minimal information set of available biobank data is presented.

Stakeholders (also companies) used in national availability service/catalogue requirement definition work

When asked about using different stakeholders including companies and industry, in the requirement definition of national services for biobanked samples, 4 Nodes replied that they had used different stakehoders in the definition work and two nodes had used only biobanks in defining requirements. Only one node did not include stakeholders to the definition work (see table 8).

Table 8. Stakeholders included in national service requirements specification.

Including stakeholders to sample locator/catalog requirement definition	N	%
Yes	4	44.4
No	1	11.1
Only biobanks	2	22.2
NA	2	22.2

Typical stakeholders that were included in the biobanked samples' national service requirement definition were biobanks themselves, industry, vendors and academia.

Level of publicly available data and data given after biobank approval

The level of availability data, here meaning the information about samples and related data, shared publicly is very heterogeneous, as can be seen in table 9. Some biobanks have selected that all data in sample locators/catalogues is publicly available. This however does not necessarily mean direct access to individual-level data; i.e. one sample catalogue offers complete lists of available collections and the data categories and variables related to them, but no individual-level data is present. One National Node replied that they give aggregate data to non-registered users. Likewise one Node replied that they give the minimum dataset, but for authenticated registered users only. Two biobanks share publicly only (k-)anonymized data.

Table 9. Level of publicly shared data availability in national sample locators/catalogues.

Level of availability data shared to public	N	%
All public	3	33.3
Aggregate data for non-registered users	1	11.1

Minimum dataset for registered users	1	11.1
Anonymized data	2	22.2
NA	2	22.2

After the availability query has been approved by biobank, biobank network or National Node, the level of queried data increases. Three National Nodes reported that after biobank/Node has approved the query, individual level data on samples and sample-donors can be given to the user. One Node stated that after the approval Directory 2.0 defined data can be given. Also one Node replied that after biobank/Node approval they will give data depending on the sample donor consent and one Node will give data if the study committee approves the data request (see table 10).

Table 10. The level of data that is given after biobank/National Node approval. **NB.** Some Nodes have replied to several categories.

Data given after biobank/Node approval	N	%
Depends on the consent	1	11.1
Depends on the committee decision	1	11.1
Minimum dataset	1	11.1
Individual level data	3	33.3
Data defined in Directory 2.0	1	11.1
NA	2	22.2

The level of publicly shared data and data that is given only after approval varies largely between the National Nodes, and the factors affecting this variation (i.e. legislation) must be carefully taken into account.

3.5 Interview results

Based on the initial results of harmonization survey, selected biobanks from each BBMRI-ERIC National Nodes were invited to interview about the harmonization processes applied in their biobank or the lack of harmonization procedures.

3.5.1 Biobanks that have initialized harmonization processes

We asked some more detailed questions from selected biobanks that have started harmonizing their data either manually or with the help of some IT tool(s) to find out what were the main advantages and bottlenecks in their harmonization processes. Main findings were, that almost all biobanks have had harmonization processes set up and IT tools applied to them for several years, so they are already quite experienced in this activity. Most biobanks had tools or scripts for aiding in the variable harmonization, though manual steps were required throughout the process.

Advantages and disadvantages of biobank-specific IT tools in harmonization processes

Usually the tools and scripts used in aiding manual harmonization processes were in-house developed and specifically customized to fit biobank's own need. Only a few open source (i.e. Opal/Mica, OSSE framework) and commercial softwares (i.e. MS Access, SLims, Cresalys) were mentioned. What is noteworthy, these in-house developed tools and scripts are specifically developed to be used only inside each biobank and they're usually not designed to be open source or easily adapted to other biobanks' needs, and the development/optimization process of these tools is continuous based on constant user requirements and feedback.

Main advantages of in-house developed tools and scripts:

- Fast development of the tool based on "real everyday lab needs"
- Modularity of the tools and scripts
- Biobank staff can easily read the scripts and modify them if harmonization process needs it
- The tools and scripts follow biobanks' own workflow and they fit biobanks' quality system
- The biobanks have full control of their data. For IT specialists the tools are easy to implement.

These in-house developed tools and scripts were also mentioned to be easily interconnected to other softwares, however no commercial databases support them fully.

Advantages of commercial and open source tools:

- Easy to use and personalize
- Web-based management of data
- Audit trail
- Security
- Community

Main disadvantages of in-house developed tools ad scripts were namely the following:

- The tools are pretty much work in progress: there's a constant need for customization, bug-fixes, .
- Harmonization processes may fail due to server/storage capacity issues
- It is a tedious and time-consuming process to build a tool from scratch
- Programming skills of biobank staff are essential
- Error reporting needs development
- Many manual steps are still needed in the harmonization process

Almost all the in-house developed IT tools were described to be "quick fixes" to solve the problems in harmonization pipeline since no suitable open source or commercial tools was available.

Disadvantages of commercial and open source tools:

- Licence costs/costs of use
- Amount and cost of server space needed
- Limited documentation of how to use the tools and API's
- "Fragile" features
- Some open source tools were mention to benefit from further development by an external service provider.

The biggest improvement to biobank-specific IT-tools and scripts would be diminishing the number of steps needing manual work. Main improvement for both commercial and open source developed tools would be a lower cost of licence and use as well as better documentation of how to use tools and APIs.

One biobank, which has worked almost 10 years with commercial software Cresalys, mentioned that one of them is still a work in progress, which is assumed to mean that the customization to suit the biobank's needs is still unfinished. The biobank would like this tool to be more customizable and it's not quite user-friendly by their standards. They also say that they provide the company with software development requests and suggestions, but they are not always accepted at the end. The software is also maintained by the company itself, so it may be difficult to involve this database to the processes developed in CS-IT (i.e. performing the ETL process)

Main challenges and bottlenecks in harmonization process

When assessing the harmonization tool interview results, we found out that there are some major bottlenecks that need to be properly addressed when defining the BBMRI-ERIC CS-IT harmonization tools. These bottlenecks consists of the following categories:

Retrospective harmonization of data: The biobanks interviewed mentioned that it is the data sources in retrospective harmonization that usually create one major bottleneck in harmonization processes, not the tools themselves. The cleanup of retrospective datasets maintained in numerous spreadsheets is tedious task: the data may need several manual check-ups and validation. The manual preparation of retrospective datasets to harmonization pipeline and preparation of those datasets that may include uncommon or local terms that are not used in harmonization tools' library is very time consuming work needing expertise and knowledge of the data.

Data schema: Storing of different datasets may also cause another bottleneck regarding what data schema would be used in the final harmonization product. All collections to be harmonized may have a different data schema. There's also a dire need of defining a minimal dataset for the harmonization process from multiple different datasets.

Capacity: Large amount of data creates capacity issues for storage when saving harmonized datasets. The amount of server/file folder space must be carefully adressed. One issue might also be lack of funds to finance the collaboration between the experts of different fields (IT, clinic, laboratory, pathology) to ensure the high quality of harmonized dataset.

Tools: When developing and customizing the biobank-specific IT tools according to the user feedback, it must be ensured that the efficiency and modularity of these tools are kept intact. The lack of standardization of data format by different local software providers can cause yet another bottleneck.

In addition, regulatory requirements that increase complexity of the databank used in harmonization pipeline may cause issues in some countries.

IT support in biobanks

All interviewed biobanks but two claimed to have at least some level of IT-support. The most common form of IT support was to have one dedicated person doing all IT related issues inside the biobank. Few biobanks also mentioned having one or several part-time IT persons helping them with system administration and development. Some of the interviewees actually were these persons responsible for It services in their biobank and they seem to be quite busy already "I am the only one IT guy in our biobank so I am doing everything that needs to be done, lots of coding". Also other discussions the author team had with harmonization experts outside this interview showed that there is a major concern of how much the planned CS-IT architecture and data harmonization processes would burden the biobanks IT person. When planning the processes to be applied into the CS-IT services we must keep the ICT solutions clear, simple and easy to adapt and maintain so the biobanks won't lose their interest in using the developed CS-IT tool package.

3.5.2 Biobanks not planning to do data harmonization

Data harmonization procedures are not in effect at every biobank. Based on the initial harmonization survey results, we wanted to find out why some biobanks are not doing or planning to do data harmonization. The main challenges named to be the reason not implementing any harmonization processes were lack of time and capable personnel, development stage or setting up the local IT infrastructure or that direct access to data to be harmonized is not allowed. Also having no or limited knowledge about existing harmonization tools/platforms may cause a biobank to have a lack of interest to initialize any harmonization procedures. Some of these named reasons came up also as main bottlenecks in the interview results for biobanks that have already initialized harmonization processes, so these issues must be carefully taken into account when designing a harmonization processes in the CS-IT level. The interview questions for biobanks not doing harmonization can be found in appendix 2b.

4 Semantic Considerations for BBMRI-ERIC CS-IT Architecture Plan

4.1 Minimal Data Requirements

In order for the sample locator queries to be successful and informative, it is important that a minimal set of elements related to biobank sample collections and sample- and sample-donor related data would be compulsory. In addition to agreeing on the minimal set of attributes, it is also necessary to agree on the coding of the values. E.g. if for instance Dutch biobanks use PALGA codes, Italian biobanks use ICD-10 codes, and Finnish biobanks use SNOMED CT (old version) to code disease, it is necessary to harmonize and map all these codes before queries can be done across countries (see sections on ontology mapping 4.4 and 4.5). Based on the questionnaire replies, the following minimal data sets are suggested:

MIABIS Biobank minimal information

For the Sample Locator query, the MIABIS Biobank minimal information as it is in the Directory v2.0 could be sufficient, i.e. Biobank name, country, and Biobank type (clinical, population, research study, non-human, stand-alone collection).

MIABIS Sample collection minimal information Note that certain biobanks are built around a single sample collection, i.e. sample collection equals to biobank name for these.

For the Sample Locator query, the following attributes, as they are in Directory v2.0, presented in table 11 could be sufficient:

Table 11. Minimal sample collection information for the Sample Locator query.

Attribute name	Attribute description	Options
Collection name	Name of collection	
Collection type	Type of collection	The type of the sample collection. Can be several values MIABIS-2.0-16
Collection size	Size of collection in 10^x	
Collection age low	Age of youngest participant	Age of youngest sample donor at time of sample donation
Collection age high	Age of oldest participant	Age of oldest sample donor at time of sample donation
Collection sex	Sex of participants	Categorical: male, female, unknown, undifferentiated
Collection available data	Type of data available in collection	Categorical: https://github.com/MIABIS/miabis/wiki/Structured-data-and-lists#data-categories

Minimal sample related data:

A MIABIS-based suggestion for minimal sample related data is presented below in table 12.
Table 12. Minimal data related to samples.

Attribute name	Attribute description	Options
Stored material type	Type of material from MIABIS	The biospecimen type saved from a biological entity for testing, diagnostic, propagation, treatment or research purposes. Can be several values MIABIS-2.0-14
Organ of origin	Anatomical Site	Code for the anatomical source of the material from an established ontology (http://www.ontobee.org/ontology/UBERON)
Human/non-human sample	Organism of origin	categorical, can include the binomial nomenclature of the species (e.g.Homo sapiens)
Sample collection/biobank name	Name of the sample collection or biobank to which the sample belongs to	
Sampling information	Circumstance of sampling	categorical, includes several different options such as diagnostic or treatment-related, cohort/population survey, family collection, research collection, other

For sampling information, we are considering to use the same categories used for MIABIS CollectionType. However, there might be an issue with

noticeable overlap in the MIABIS categories mentioned and that a sample can belong to several categories. The issue regarding the overlap could possibly be accommodated via IT-solution.

Sample donor related data:

Table 13. Biobank data related to sample donor.

Attribute name	Attribute description	Options
Age at sampling	Age of the sample donor when the sample was taken	years or 5y-age groups
Gender	Gender (biological sex) of sample donor	male, female, unknown, undifferentiated (defined in MIABIS and various ontologies)
Disease diagnosis at sampling	Diagnosis using established ontology	e.g. ICD-10, ICD-O, SNOMED-CT, OMIM, Orphanet, no diagnosis
Available data categories	medical records, imaging data, survey/questionnaire data, national registries, genealogical records, physiological and biochemical measurements, genomic data	yes/no

It is important to note that some biobanks rely on simple solutions for data handling (e.g. MS Excel files) and many have limited personnel resources, for example, only 1 person in charge of all data aspects. Therefore, the compulsory data set should be kept small and simple to allow all biobanks to participate in the Sample Locator system.

4.2 Optional Data Suggestions

The two more frequent types of biobanks in our survey are the hospital biobanks and population biobanks. It would be useful if a set of optional, agreed upon attributes will be selected to catch the necessary information in these very different types of collections at a better granularity than the minimal data set (based on the data in the tables 11, 12 and 13). The more detailed the information is about the biobank collections, the more useful the query results will be. Otherwise there is a risk that the minimal common data provided will be too shallow and will not contain the necessary information to really identify the biobanks/collections of interest.

Is it possible to make an initial filtering option that allows the researcher to specify the following issues?

- human samples or non-human samples
- disease (diagnosis) -specific or not disease-specific

Based on these initial selections, the user will get access to a more informative query that will include optional data sets that are more specific to the type of sample collection/biobank in question. An example of optional data set for cancer-related collections could be the colon cancer data set that is described by BBMRI-ERIC CS-IT WP3 (Proynova 2016) or the basic data used in the Belgian Tumor Bank[5].

[5] http://www.fournier-majoie.org/en/about-us/partners/belgian-virtual-tumour-bank-belgian-cancer-registry

4.3. Metadata Models

Based on the planned architecture, the MDR shall follow the ISO 11179 definition of metadata items, particularly oriented by the “Registry metamodel and basic attributes”. The minimal required data elements should be based on accepted metadata models, such as MIABIS attributes [6]. The requested metadata models should be made available to all participating biobanks in several different file formats, and the process for describing new data elements should be made as user friendly as possible.

Examples of metadata models currently in use:

- WP3 Colon cancer collection metadata model for clinical data[7]
- MIABIS metadata model in BBMRI-ERIC Directory for describing biobanks and sample collections
- KITE metadatamodel for population-based collections[8]

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

[7] 2016-06-29 Semantic Architecture Telco

[8] https://kite.fimm.fi

4.4 Ontologies in Common Use at European Biobanks

Based on the questionnaires, ontologies presented in table 15 are in use in European biobanks and should be supported by the BBMRI-ERIC CS-IT Services.

Table 15. Commonly used ontologies in European biobanks based on the harmonization questionnaire replies and discussions the author team has had with ontology experts.

Ontology	Aim	Functionalities
DSM-IV (https://www.psychiatry.org/psychiatrists/practice/dsm)	Diagnostic and Statistical Manual of Mental Disorders is the standard classification of mental disorders used by mental health professionals. DSM can also be used for research in clinical and community populations. It is also a necessary tool for collecting and communicating accurate public-health statistics.	DSM consists of three major components: the diagnostic classification, the diagnostic criteria sets, and the descriptive text.
ICD-10 (+previous versions) (http://www.who.int/classifications/icd/en/)	ICD-10 is the 10th revision of the International Statistical Classification of Diseases and Related Health Problems (ICD), a medical classification list by the World Health Organization (WHO).	It contains codes for diseases, signs and symptoms, abnormal findings, complaints, social circumstances, and external causes of injury or diseases.
ICD-O,(and ICD-3 +previous versions) (http://codes.iarc.fr/)	The classification of neoplasms used in ICD-O links closely to the definitions of neoplasms used in the WHO/IARC Classification of Tumours series which are compiled by consensus groups of international experts and, as such, the classification is underpinned by the highest level of scientific evidence and opinion.	ICD-O consists of two axes, the topographical code which describes the anatomical site of origin of the tumour and the morphological code, which describes the cell type or histology of the tumour, together with the behaviour.
MeSH (https://www.nlm.nih.gov/mesh/)	MeSH is the National Library of Medicine's controlled vocabulary thesaurus. It consists of sets of terms naming descriptors in a hierarchical structure that permits searching at various levels of specificity.	The MeSH thesaurus is used for indexing articles for the MEDLINE®/PubMED® database. It is also used for the NLM-produced database that includes cataloging of books, documents, and audiovisuals acquired by the Library. Each bibliographic reference is associated with a set of MeSH terms that describe the content of the item. Similarly, search queries use MeSH vocabulary to find items on a desired topic.
MIABIS (https://github.com/MIABIS/miabis/wiki)	MIABIS represents the minimum information required to initiate collaborations between biobanks and to enable the exchange of biological samples and data. The aim is to facilitate the reuse of bio-resources and associated data by harmonizing biobanking and biomedical research.	MIABIS attributes describing biobanks and sample collections are already in use at the BBMRI-ERIC Directory 2.0

OIE-World Organisation for Animal Health standard requirements (http://www.oie.int/)	The OIE has been developing and compiling a number of technical background documents and guidelines on a variety of specialised topics related to animal health. This scientific information is mainly dedicated to veterinary services and technical experts and is regularly updated.	An OIE Reference Centre is designated either as: “OIE Reference Laboratory” whose principal mandate is to function as a world reference centre of expertise on designated pathogens or diseases; or “OIE Collaborating Centre” whose principal mandate is to function as a world centre of research, expertise, standardisation of techniques and dissemination of knowledge on a specialty. The network of Collaborating Centres and Reference Laboratories constitutes the core of OIE scientific expertise and excellence. The ongoing contribution of these institutes to the work of the OIE ensures that the standards, guidelines and recommendations developed by the Specialist Commissions and published by the OIE are scientifically sound and up-to-date.
OMIM (http://www.omim.org/)	OMIM is a comprehensive, authoritative compendium of human genes and genetic phenotypes that is freely available and updated daily.	OMIM contain information on all known mendelian disorders and over 15,000 genes. OMIM focuses on the relationship between phenotype and genotype.
Orphanet (http://www.orpha.net)	Orphanet is the reference portal for information on rare diseases and orphan drugs, for all audiences. Orphanet’s aim is to help improve the diagnosis, care and treatment of patients with rare diseases.	Orphanet services are: An inventory of rare diseases, An encyclopaedia of rare diseases, An inventory of orphan drugs, A directory of expert resources, An assistance-to-diagnosis tool, An encyclopaedia of recommendations and guidelines for emergency medical care and anaesthesia.
SNOMED-CT (http://www.ihtsdo.org/snomed-ct)	SNOMED CT is the most comprehensive and precise clinical health terminology product in the world, owned and distributed around the world.	IHTSDO Tooling consists of the services and tools that support the SNOMED CT International and Spanish Edition product life cycles, and the systems that are used to support the IHTSDO, Members and the Community of Practice.
SPREC (http://www.sprecware.org/?q=Sprec)	“Standard PRE-analytical Code” (SPREC), is a proposal for a standard coding of the pre-analytical options which have been adopted, in order to track and make explicit pre-analytic variations in collection, preparation and storage of specimens.	Each biospecimen is assigned a seven-element-long code that correspond to seven pre-analytical variables and contains a string of letters(different for fluid or for solid tissues) in a defined order, separated by hyphens.
TNM Cancer Staging System (http://www.cancer.gov/about-cancer/diagnosis-staging/staging)	The TNM system is the most widely used cancer staging system. Most hospitals and medical centers use the TNM system as their main method for cancer reporting.	In the TNM system: The T refers to the size and extent of the main tumor. The main tumor is usually called the primary tumor. The N refers to the the number of nearby lymph nodes that have cancer. The M refers to whether the cancer has metastasized. This means that the cancer has spread from the primary tumor to other parts of the body.

UBERON (http://uberon.github.io/)	Uberon is an anatomical ontology that represents body parts, organs and tissues in a variety of animal species, with a focus on vertebrates.	Uberon has been constructed to integrate with other ontologies, such as OBO Cell Ontology, Gene Ontology, Trait and Phenotype ontologies. One of the main uses of Uberon is translational science, we have extensive coverage of structures shared between humans and other species.
UMLS (https://www.nlm.nih.gov/research/umls/)	The UMLS integrates and distributes key terminology, classification and coding standards, and associated resources to promote creation of more effective and interoperable biomedical information systems and services, including electronic health records.	UTS Metathesaurus, semantic network, Specialist Lexicon and Lexical Tools

In addition, there are country(/language)-specific ontologies that are used in various biobanks and mapping of them to the common ontologies mentioned above should be explored. These include:

- THESAURUS ADICAP (France, an oncology-related ontology)
- LOINC (Logical Observation Identifiers Names and Codes, maps different language versions to one unified term)
- OBIB, OMIABIS

The list of ontologies used by BBMRI-ERIC should be reviewed periodically to verify that it is current, and new ontologies could be added as needed.

4.5 Harmonization Processes and Ontology Mapping Tool requirements to be Incorporated into CS-IT Architecture Plan

As stated in the introduction, the harmonization processes to be included in the CS-IT initial phase are meant to support the participation of European biobanks in the CS-IT Sample Locator System, using pre-defined sets of variables that describe biobanks, biobank sample collections, biobank samples and sample-related and sample donor-related data. This selected data set should rely on existing supported ontologies (mentioned in section 4.4) when possible and on the MIABIS attributes already existing in the Directory. Below we describe the requirements for these processes and tools.

The mapping/harmonization process includes the following steps: definition of similarity level between attributes, selection of best match, conversion rules, renaming of attributes, supporting language versions, etc.

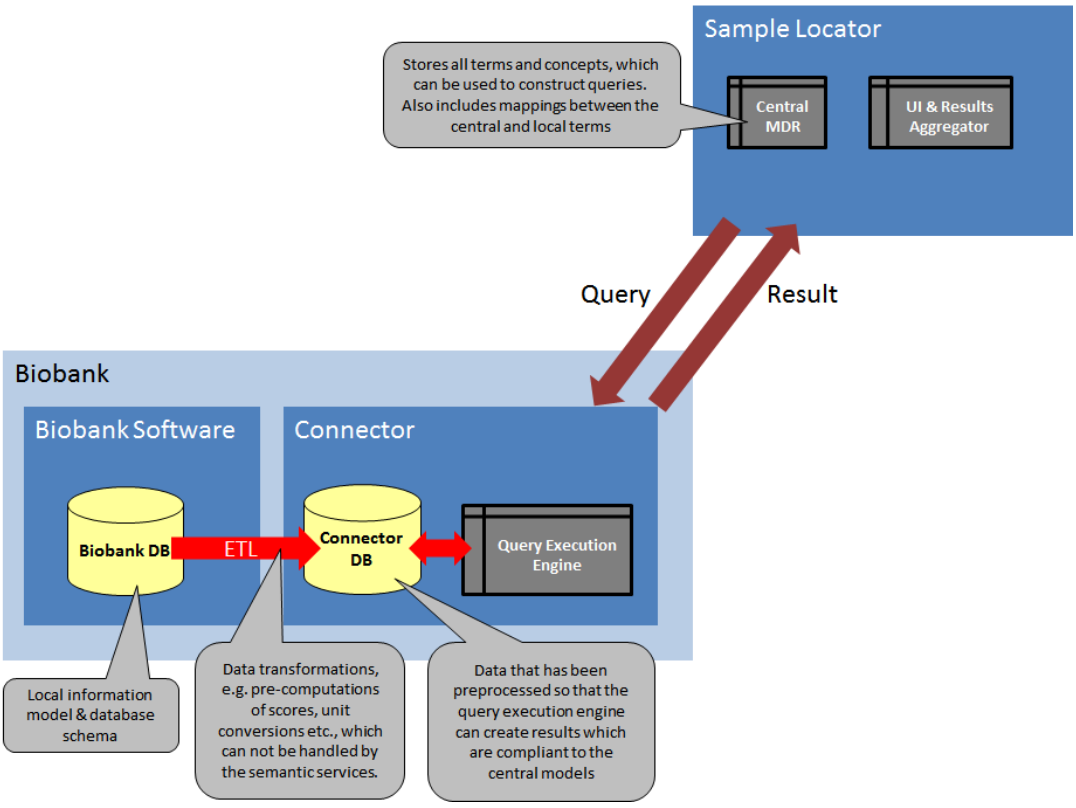

Figure 3. CS-IT Architecture Plan with semantic aspects.

4.5.1 Harmonization process requirements

Harmonization processes required for the CS-IT architecture are needed at several levels:

1. Mapping of local variables to the selected minimal data sets used in the Sample Locator and the Directory
2. Mapping of local ontologies to the selected ontologies supported by BBMRI-ERIC
3. Harmonizing the variables' values with the values required for the directory and Sample Locator (examples: Gender marked as M/F should be converted to the options supported by MIABIS; Converting measurement units; Harmonizing date information; etc.)

The harmonization process from a biobank's point of view could include the following steps:

1. Select the variables that correspond to those variables required for the Sample Locator
2. Collect all the necessary metadata for the required variables as it corresponds to the biobank's own data, tabulate it in the correct format required by MDR.
3. Upload the variable metadata into the MDR, make sure the upload went correctly using data upload report tools and by manually checking the data.
4. Map the local variables with the required variables in the MDR using a mapping tool (such as BiobankConnect). [Steps 3 and 4 can happen also in the opposite order, if the mapping tool is not part of the MDR, but operates as a separate tool].
5. One variable at a time, and using the mapping tool, compare the units/options for the MDR required dataset, and decide whether the variable is compatible as such or are conversion rules needed before the variable can be used. Write the conversion rule. Save the updated local variable metadata into your local space in the MDR.
6. Map the local ontologies used in the biobank's data to the supported ontologies of BBMRI-ERIC, using an ontology mapping tool (see 4.5.2). This is an important and difficult step if there is no ready mapping of a local ontology to the supported ontologies. For local ontologies that are already mapped to the supported ontology, that mapping could be readily used and the links between the ontologies saved in the MDR. During the individual-level data upload process, there should be an automatic conversion process that converts the local ontologies to the supported ontology.
7. Set up a data warehouse software (=connector) provided by BBMRI-ERIC onto the biobank's own server, and verify there is proper connection between the data warehouse and the Sample Locator software (local or central?) and the central MDR. [How can the software and the connections be tested?]
8. Prepare the individual (or sample) –level data files for the biobank's sample collections. If there are various sample collections, each one including slightly different variable names and value options, harmonize the variables for all collections as needed (points 1-6). Depending on the available data the biobank wishes to share, it is possible that different data tables would need to be prepared for a given collection, e.g. individual-level data, and sample level data.
9. Upload the individual-level data to the datawarehouse using the available import tools. The import tool should at best do the transformations of the data based on the transformation rules automatically, during the import process. Use the upload reporting tools to verify the data has been properly imported, the transformations of the data properly done as specified in the MDR, and the ontology mapping correctly done.

10. Test the Sample Locator to see that the data is now available for sample availability queries.

In order for biobanks to participate in the sample locator system, a detailed instruction document that explains all steps should be prepared. The following issues should be included:

1. Information on which variables are compulsory to provide about biobank, collection and sample-/individual –level data
2. Information on required metadata model for each compulsory variable
3. Information on supported ontologies
4. How local ontologies are mapped to supported ontologies
5. How local variables are mapped to compulsory variables
6. How transformation rules for variable values are built
7. How metadata is imported to the MDR and how individual-level data imported to the local datawarehouse.
8. How the local softwares should be installed and checked.
9. What are the server requirements

However, we must keep in mind that the data harmonization procedure to be developed is going to be simple enough in order to facilitate retrospective harmonization of variables that usually is a time-consuming manual task and therefore can cause a massive loss of data uploaded to MDR. Also the procedure must be compliant with already existing standards, SOP's, guidelines and terminologies that biobanks already use so that the new format of storing sample and data can easily be implemented with the existing tools when collecting new samples and data.

4.5.2 Ontology mapping tool requirements

A software tool that allows mapping of local data to BBMRI-ERIC endorsed ontologies. This tool will help automatize the process of converting local data versions to the standard versions. Examples of use cases for ontology mapping:

- Converting ICD-8 and ICD-9 versions to ICD-10
- Converting older SNOMED versions to the newest SNOMED-CT
- Converting older UMLS version to the newer UMLS
- Converting the French THESAURUS ADICAP to ICD and SNOMED, when applicable
- Converting different versions of MIABIS attributes to the current MIABIS format
- Mapping local attributes to any of the ontologies that are currently in use in BBMRI-ERIC
- Supporting different language versions of all mentioned ontologies, when available

Ontology conversions have already been made for some of the ontologies mentioned above and these mappings should be utilized in the CS-IT mapping tool, e.g. run in the background and provide the correct mapping easily. For example, ICD-9 has been mapped to ICD-10, and currently there is an international effort to develop ICD-11[9] (due in 2018), including translations into other languages.

The ontology mapping tool can also support comparison of attributes between two different sample collections or data sets and their harmonization into unified harmonized attributes to promote the use of different biobank collections in the same study.

The mapping/harmonization process includes the following steps: definition of similarity level between attributes, selection of best match, transformation rules, renaming of attributes, supporting language versions, etc.

[9] http://www.who.int/classifications/icd/revision/en/

4.6 Additional considerations for development of CS-IT tools

In this section are presented some additional thoughts and considerations raised in the development of CS-IT tools.

4.6.1 Aggregate data vs. individual level data

The use of individual level data puts many limitations on how the data can be shared, with whom the data can be shared, and how to provide the common query tool that accesses the biobank's individual level data. Even if the individual level data is not directly shared with the researchers, providing direct query access to it even through firewalls, access control and anonymization procedures is still a problematic issues. Different countries have different legislations that should be taken into consideration when developing tools that rely on access to individual level sensitive data.

One option to consider is the use of aggregate level data as the basis for sample availability queries. This has been successfully implemented, for example, in Auria Biobank in Finland[10]. Auria Biobank's catalog of available tissue samples is public and is fully based on aggregate level data that includes the following fields: organ class, organ subclass, diagnosis class, neoplasm type, age and gender.

It would be good to consider extending BBMRI-ERIC's Directory with aggregate level data on samples and sample-related data. Several different aggregate data models could be developed to accommodate the different types of sample collections available in European Biobanks.

[10] https://www.auriabiopankki.fi/services/biobank-samples/?lang=en

4.6.2 Practical consideration for queries

Practical issues to resolve in implementing the sample locator queries:

Linking sample donors to samples: How the links between sample donor information and sample-related information will be established in the Connector? For example, in a typical sample collection scenario, several different samples are taken from the sample donor (X number of blood tubes for serum/plasma/DNA/RNA and/or other liquids, and/or tissue sample that is divided into several different pieces that are processed differently). Thus, for every sample donor there will be a number of different types of samples available. The query should take this into account and not consider each information row about sample as independent sample donor. Will the connector have different data tables for different type of data (=subject-related data and sample-related data)?

Sample aliquots: It is suggested that sample aliquots will be disregarded for the level of data granularity that is needed for sample queries. Information on aliquots is usually difficult to obtain and it is constantly changing.

Directory queries vs. sample locator queries:

- Will the directory have proper query tools that are based on information provided about sample collections?
- How will these queries be aligned/linked with query tools that correspond to data present in the sample locators (i.e. individual-level and sample-level data)?
- Can queries be done simultaneously for data present in the Directory and data present in the local connectors, or will these be two separate queries?
- It is possible that some biobanks will have their local connectors linked to the central sample locator query system, while others will have information only in the Directory, while other biobanks will have information in both. The query should catch both cases and possibly consolidate duplicate results.

Queries of multiple keywords: It is important that multiple keywords and options could be queried simultaneously. For example, several different ICD-codes, different types of samples, different age groups, etc.

4.6.3 Connector and MDR User Requirements

The Connector is a software that should be installed and maintained at local biobanks. MDR however, is a central component that can have local space for those biobanks that need to operate MDR locally. Since the level of IT-expertise in biobanks varies, and not all biobanks have a dedicated IT-person or even proper database tools for their use, it is very important that the Connector and MDR software provided by BBMRI-ERIC CS-IT would be easy to install and use. Under are listed some specific issues to consider:

- Easy installation package with good documentation and online support from CS-IT
- Good User Manual written in several languages that provides detailed information on how to use the software
- Different import tools that allow uploading metadata on variables into the MDR one variable at a time or in a table format (i.e. csv), as well as tools for uploading individual level data into the Connector in table format (Command-line functionalities may be problematic). For the import of metadata in table format, it is important to make sure that all required metadata elements are included in the import. A tool for mapping the biobank's own variables to the common CS-IT variables should work in sync with the MDR.
- Informative import process error logs that pinpoint where the problem in import is (row number in table, attribute name) and how it could be solved
- Is it possible that the connector and the MDR will be accessed via one user interface, or even that they will be included in the same software tool that supports both metadata and individual-level data? This type of software tool is available in several biobanks and projects, such as MIABIS Connect, i2b2.
- Good and easy to use search and edit tools in both software that allow searching for variables and individual level data, editing them, deleting them, etc. (Command-line functionalities are problematic)
- IT-support from CS-IT team to obtain the necessary access of the central Sample Locator to the local CS-IT databases, through the biobanks' firewall systems. The architecture is designed considering these challenges - e.g. Information exchange architectural mechanism - local systems will be polling central components for data exchange.
- A unified user interface that incorporates in it the MDR, the Connector and the ontology mapping and harmonization tool/s. It is important that the user will not have to switch between different software continuously, and that each tools will be a component of the same system. This issues is already planned in the CS-IT Architecture Notebook (Alexandre et al. 2016).

4.6.4 Interlinking CS-IT Directory to other catalogues

Based on the questionnaire replies, many biobanks are already registered/participating in existing catalogs and networks. It would be important to have a way to link the biobanks in the Directory also with these already existing catalogs. As such, these catalogs usually have been designed to serve specific topics and usually contain much more information than that provided in the Directory. As such, the information available in them would be useful for the researcher.

A few examples of such catalogues:

- Telethon Network of Genetic Biobanks (http://biobanknetwork.telethon.it/)
- Euro Biobank Sample Catalogue (http://www.eurobiobank.org/en/services/CatalogueHome.html)
- BBMRI LPC Catalogue (http://www.bbmri-lpc-biobanks.eu/cataloque.html)
- RD-Connect catalogue (http://catalogue.rd-connect.eu/)
- The German Biobank Registry (https://www.biobanken.de/en-gb/home.aspx)
- Kite - Finnish catalogue of sample collections (https://kite.fimm.fi)
- Belgian Virtual Tumor Bank (BVTc, http://virtualtumourbank.kankerregister.be/tumourbank.aspx?url=BVT_home)
- BBMRI.at catalogue (http://catalog.bbmri.at/)
- BBMRI.nl catalogue (https://catalogue.bbmri.nl/)

5 Points to consider for the CS-IT architecture

Based on the above mentioned information and on other discussions within the BBMRI-ERIC CS-IT team, there are several points to consider regarding the CS-IT architecture requirements, which are discussed below.

5.1 Ontology considerations

BBMRI-ERIC could support ontologies with the following means:

a. The BBMRI-ERIC supported ontologies should be stored somewhere in the CS-IT system e.g. the MDR.
b. Existing mappings between different versions of the same ontology and different ontologies should be stored and used, possibly using a CS-IT supported ontology mapping tool.
c. Existing ontology mapping tools should be used for CS-IT needs where possible without re-inventing the wheel. There are solutions that could support ontology mapping such as Protege and RDF.
d. The use of meta-ontologies (such as the UMLS, which encompasses most of the supported ontologies and many others) could be considered. If these meta-ontologies are used, they require large capacity to handle and may slow down any application involved. An alternative is to use these meta-ontology mapping via the Bioportal API.
e. Tool to support mapping of biobank variables to the supported ontologies should be available. This tool should be able to also link different ontologies to each other when possible and also have an option for distinct values to be linked to each other manually and saved (=ontology curation) for further use. Minimum level of functionality would be an option to link the data elements with supported standardized ontologies.
f. There is existing literature discussing guidelines for updating ontologies that could be used.

Ontologies can be reused to support the mapping of central namespaces to local namespaces.

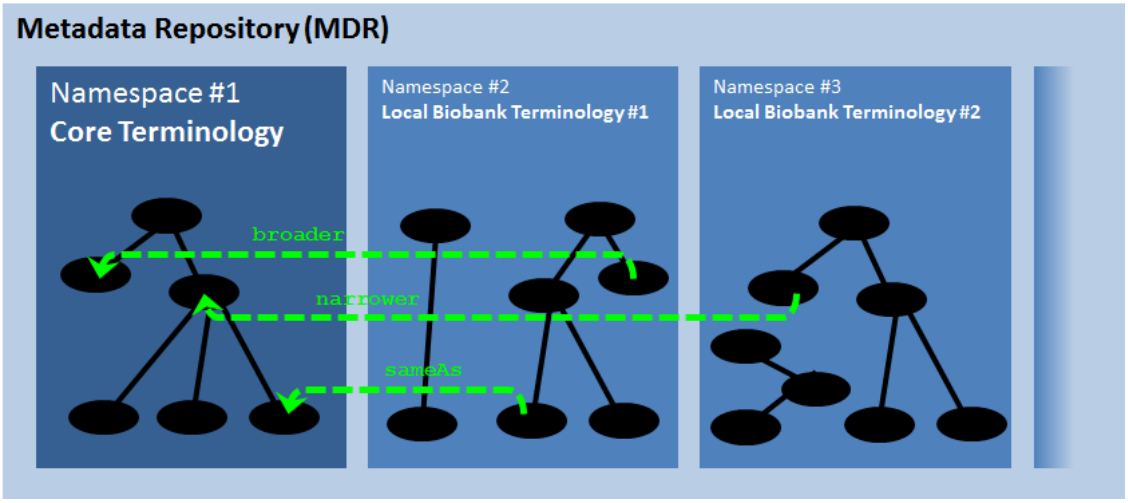

Figure 4. Namespaces and how they relate to each other in MDR.

5.2 MDR considerations

a. Metadata model: The metadata models used in the MDR and the Directory should be compatible to each other. Additional metadata models could be based on existing standards when available.
b. The biobanks' metadata will be stored in one central MDR instance, with a local space provided for those biobanks needing it. MDR should be the main contact database for the biobank user, while the mappings with ontology references could be working in the background of the MDR, so that the user is not necessarily aware of them.
c. There should be only one central set of namespaces in use that the majority of participating biobanks could use as such. Some more specific ontologies may be considered if necessary.
d. The namespaces biobanks use/upload can be unique or they can use the central namespaces. However, the biobanks need to map their own namespace to the central MDR. A process to support the mapping of local namespaces with the core terminology should be provided, which includes also transformation rules by which values of a given variable can be changed into a standard variable (for example, unit conversion, change of enumeration, etc.). An example of such mapping tool is BiobankConnect.
e. Import functionality of any new variable metadata into MDR should support both manual, single variable addition/edit, as well as an upload of tabulated data in an agreed upon format. Easy to use user support tools are needed in the import process.

5.3 Additional levels of sample availability data

The Directory could support additional levels of specific aggregate data on sample availability, that currently are not supported by it. An example of aggregate level of data sharing that currently is not supported by the directory is the aggregate data in the sample catalog of Auria Biobank for pathological samples[11]. The Auria catalog is based on aggregated data about tissue samples and provides information on how many samples are available per diagnosis/gender/organ/hospital/etc. These additional levels of data sharing would allow biobanks that are unable to participate in the sample locator effort e.g. due to personnel resources, lack of needed expertise or other limitations, to still provide basic availability data.

Different aggregate sample availability sets could be developed to support different types of sample collections, such as epidemiological DNA collections, different types of cancer collections. An example for this kind of availability set is the colon cancer database currently built by ADOPT, disease-specific collections, etc.

Joint sample availability queries that are based on sample-level data require a high degree of data harmonization across all biobanks. The added amount of effort (personnel-wise, work time and other resources) requested from all participating biobanks, which is linked with producing updated and harmonized sample-level data could be considerable compared to providing a simple table with aggregate level sample availability data.

[11] https://www.auriabiopankki.fi/katalogi/

5.4 Considerations regarding the biobanks' data warehouse (=Connector)

The data warehouse needs to contain both subject-related and sample related data. The link between subject ID and various sample IDs can be done in several ways. It's important to note that some biobanks may have several different kinds of samples from one sample-donor, and even different biobanks may have samples from the same subject. It will be a big effort for biobanks to prepare sample-level data for the data warehouse. One option is that the subject-level data will only contain information on the type of samples available for each given subject. The data granularity levels could be different from different biobanks, though.

5.5 Directory and sample locator

According to the current CS-IT Architecture Notebook (Alexandre et al. 2016, see figure 1 titled "Overview of WP1, WP2, WP8 and auxiliary components"), the Directory and Sample Locator will not be connected such that the user would have an option to have their queries access information in both Directory and Sample locator and provide a reply that includes information from both sources. This would, however be preferable from the end-user point of view. Otherwise, the user will have to consolidate the results of 2 separate queries in order to obtain the necessary information. The joint query option would be especially important if a middle level of data sharing is made available through the use of specific aggregate datasets of sample availability, as can be seen in 5.1.).

5.6 Piloting the current architecture plan

In order to test the functionalities and processes planned for the CS-IT infrastructure, some of them listed above, pilots are critical. In these pilots, the feasibility of the process, as well as the necessary personnel and their expertise, the time needed for each step of the process, and the estimated cost should be evaluated. A few suggested pilots are listed below, which should use currently available IT-tools.

1. Harmonization process pilot with the participation of several biobanks, which includes selected variables from the minimal and optional datasets (sections 4.1 and 4.2). The pilot would aim to test if the planned processes for data harmonization are within the capacity and capability of the participating biobanks.
2. Use of the central MDR and local connector and necessary upload of metadata and sample-level data by several biobanks, with connection to the current sample locator software to perform actual queries on the sample-level data.
c. Description of the current planned process that allows biobanks to provide sample availability data, including all the steps and architecture components involved. This description should be reviewed by several biobanks of different type (clinical, population, research). The biobanks could provide feedback on how feasible the process is from their point of view.

These pilots should be conducted in collaboration between WP2, WP6 and WP8.

The importance of pilots is highlighted by experience from similar biobank networking efforts that have already been done, for example in Paris, France, where a network of data warehouses was built with i2B2 and Shrine (information from Nicolas Malservet). The strategy was to first install a common tool to store data (here i2B2), and then build the network with commons APIs. The staff and resource required were significant:

- There was data and software heterogeneity between the biobanks
- BBMRI.fr catalogue experiment results with only 30 data items weren't satisfactory
- "The approximate cost to develop each connector was 7 days, if we can have all the informations to access to the source database. If we need to do it with 80 biobanks, and maintain it each 6 months in France, we need 2 full time developers with a huge budget to travel."
- The estimated cost into the harmonization process needs to be integrated. The best way to harmonize the data into the network to make a sustainable system. The harmonization process and the strategy behind it will drive many technical choices.

References

Alexandre et al. 2016 Alexandre, D; Lablans, M; Ückert, F; Holub, P; Hummel, M; Swertz, M; Proynova, R; van Enckevort, D; Jetten, J. BBMRI-ERIC CS-IT Architecture Notebook. Deliverable report for ADOPT D3.3, submitted September 2016.

Holub et al. 2016 Holub, P; Quinlan, P; DiezFraile, A; Hummel, M; Kuhn, K; Litton, JE; Müller, H; Swertz, M; Ückert, F; Valík, D; Vojtíšek, O; Zanetti, G. BBMRI-ERIC Use Cases.

Mate et al. 2016 Mate, S; Leb, I; Schüttler, C; Silander, K; Miettinen, T; Eklund, N; Knuuttila, J; Prokosch, HU; Overview document on already existing tools for data harmonization and terminology mapping. Deliverable report for BBMRI-ERIC CS-IT D8.2, to be submitted in 2016.

Merino-Martinez et al. 2016 Merino-Martinez, R; Norlin, L; van Enckevort, D; Anton, G; Schuffenhauer, S; Silander, K; Mook, L; Holub, P; Bild, R; Swertz, M; Litton, JE. Toward Global Biobank Integration by Implementation of the Minimum Information About BIobank Data Sharing (MIABIS 2.0 Core). Biopreservation and Biobanking. August 2016, 14(4): 298-306.

Proynova 2016 Proynova, R. User interface for collection of the colon cancer cases and database. Deliverable report for ADOPT D3.2, submitted September 2016.

Appendices

Appendix 1. Harmonization questionnaire for BBMRI-ERIC Biobanks

Appendix 2. A) Interview questions for biobanks having harmonization tools

B) Interview questions for biobanks not doing data harmonization

Appendix 3. Availability questionnaire for BBMRI National Node Coordinators

Questionnaire regarding data harmonization for Biobanks

BBMRI-ERIC's Common Service IT project aims to develop common IT tools that would aid researchers in finding and accessing samples and data hosted at European biobanks. The work is divided into 8 work packages (WPs), with WP8 responsible for facilitating collaboration on IT harmonization issues.

The aim of this questionnaire is to gather information from biobanks in Europe about their current processes used to harmonize data on biobanked samples.

The questionnaire includes several specific questions about data harmonization for European biobanks.

Harmonization in this questionnaire means the effort to process data in a specific, unified manner so that the data elements or attributes would be comparable over different sample collections.

You can skip any questions in the questionnaire if they don't apply to you.

1. Respondent's information and biobank

Respondent's Name

_____

Title/Position

_____

Biobank's name

_____

City

_____

Country

_____

2. What type of sample related data do you store?

_____

_____

3. Beside the sample related data, do you also have access to sample donor related data, such as:

Medical data

☐ Medical history

☐ Diagnoses

☐ Medication

☐ Medical procedures

☐ Laboratory values

General information

☐ Demographics and other background information

☐ Lifestyle and environmental information

☐ Findings

☐ Laboratory values

☐ Medication

☐ Data from official registries

☐ Other, please describe

_____

4. Are these data entities harmonized in some way (i.e. using ICD-10 for diagnoses or common reference terminologies, such as MeSH, UMLS etc.)?

_____

_____

_____

5. What types of data harmonization processes do you have at your biobank? What is the aim of each harmonization process?

_____

_____

_____

6. Do you use specific IT tools in aiding the harmonization? What tools? Please provide a short description.

_____

_____

_____

7. What steps in the harmonization do you handle manually and what steps have you automatized (and how)? Could you describe the processes briefly?

_____

_____

_____

8. Do you use specific reference terminologies/ontologies/guidelines/standards in the harmonization process?

☐ Standards and/or guidelines, what (i.e. ISO Standards, OECD and ISBER Best Practices)

_____

☐ Reference terminologies and ontologies, what (i.e. SNOMED CT, ICD-10, MeSH)

_____

☐ Other, what

_____

9. Do you have tools to aid researchers in finding specific samples for research purpose (=catalogs, directories, sample locators, metadata repositories, etc.)? Please provide a short description of these tools.

_____

_____

_____

10. Can you provide a name and contact information of an expert whom we can possibly interview to obtain more details of your processes?

Name _____

Lastname _____

Title _____

Mobile _____

Email _____

City _____

Country _____

Organization _____

Department _____

☐ I want to submit my answers

Appendix 2. A) Interview questions for biobanks having harmonization tools

1. You reported that your biobank uses specific IT-tools in aiding the (manual) harmonization. Please try to answer the following questions for each tool:

- How long have you been using these tools?
- Do you know if these tools are in use also by other biobanks?
- What are the advantages and disadvantages of each tool?
- What would be a relevant improvement to these tools?
- What else would you like these tools to support?
- Are these tools in-house developed or (customized) software from commercial partners?

2. Regarding the harmonization tools you are using, will they be usable also to other biobanks (i.e. biobanks in other countries)? Are they open source and can they potentially be further developed by others to accommodate the needs of BBMRI-ERIC?

3. What are/were the main challenges and bottle necks in the harmonization process?

4. What kind of database schema(s) you use in your IT-systems? Can you provide an example of your schemas and/or an example of the (meta)data included in your catalog/availability service? Please, use fake data in the example.

5. Does your biobank employ someone who is in charge of IT-specific issues (like performing the ETL process, installing software, fixing errors, no coding required)?

Use-case specific questions:

What is the process of (manual) data harmonization?

What elements/columns are used in data matching?

How are the harmonized datasets used and stored?

Do you use a specific metadata model in harmonizing/storing the variables?

Appendix 2. B) Interview questions for biobanks not doing data harmonization

1. You reported that you have no specific tools for biobank data harmonization. Do you know the following open source IT harmonization tools or processes that you would like to be able to use?

- Molgenis platform
- Maelstrom platform
- Obiba toolkit
- PhenX toolkit
- Gen2Phen toolkit
- i2b2 framework
- RD-Connect platform
- MIABIS
- Other (e.g. data warehouse solutions), please describe shortly

  What makes the process/software interesting, how it would fit in your data harmonization process?

2. What are the main challenges in implementing harmonization processes at your biobank?

3. What kind of database schema(s) you use in your IT-systems? Can you provide an example of your schemas and/or an example of the (meta)data included in (your catalog/availability service)? Please use fake data.

4. Does your biobank employ someone who is in charge of IT-specific issues (like performing data import into the database, installing software, fixing errors, etc.)?

Availability questionnaire for BBMRI ERIC National Nodes

BBMRI-ERIC's Common Service IT project aims to develop common IT tools that would aid researchers in finding and accessing samples and data hosted at European biobanks. The work is divided into 8 work packages (WPs), with WP8 responsible for facilitating collaboration on IT harmonization issues.

The aim of this questionnaire is to gather information from BBMRI National Nodes in Europe about their current sample availability services.

The questionnaire includes several questions about availability services to be answered by the national node coordinator. Specific questions about data harmonization is sent to European biobanks in another questionnaire.

You can skip any questions in the questionnaire if they don't apply to you.

1. BBMRI National Node

Country

_____

2. Do you have national sample locator/sample finder/catalog services in your country? Please provide link and a short description.

_____

_____

_____

3. Have you used different stakeholders (also companies) to define the requirements of your national services (described in Q1) for biobanked samples? If yes, please provide a short description about the process.

_____

_____

_____

4. What level of availability data (=information about available samples and data) can be shared publicly and what is/should be given to registered users (= self-registration, not registered through an authentication service)?

_____

_____

_____

5. What data is/should be given only after approval of the biobank/national node?

_____

_____

_____