BBMRI-ERIC
Biobanking and
BioMolecular resources
Research Infrastructure

RDA Data Fabric IG (DFIG): BBMRI-ERIC IT

Petr Holub, Jan-Eric Litton, and Common Service IT contributors

July 31, 2015
Revision: 59

This document has been written as a BBMRI-ERIC use case description for Research Data Alliance (RDA) Data Fabric IG (DFIG). References to existing RDA work are minimized on purpose, as it focuses on description of the use case itself. The structure of section complies with the RDA DFIG. The document is intended to be published by RDA among other use cases.

Document Overview

1 Scientific Motivation and Outcomes

Biobanks have become a major source of biosamples as well as data for the biomedical and bioinformatics research. Data collection, harmonization and processing has been part of the biobanks since their inception, as biosamples without the data is of little use. The data collection started with the phenotype, clinical, and lifestyle data (with focus on specific data types given by the type of the biobanks, such as population biobanks or clinical biobanks). Unprecedented growth of omics data generation in recent 15 years have brought biobanks into the domain of big data, processing and storing genomics, proteomics, metabolomics and other types of data.

After about ten years of preparations, BBMRI-ERIC as become one of the first European Research Infrastructure Consortia, with the mission of providing high-quality samples, data, and biomolecular resources from biobanks to support healthcare advancement in Europe and beyond. The major goals of BBMRI-ERIC are:

- to *increase use of material and data* stored in European biobanks, while adhering to strong *privacy protection* of patients and donors contributing the material and data,
- to *improve quality and traceability* of the material and data in European biobanks, referring to the infamous recent publications demonstrating that large portions of biomedical reasearch are not reproducible [1, 2, 3, 4, 5] and this has been even demonstrated specifically for the process of generating data from samples [6],
- to *improve data harmonization* and contribute to the standardization processes,
- to *contribute to the ethical, legal, and social issues*, with particular focus on cross-border exchanges of human biological resources and data attached for research use.

Although biomedical, bioinformatics researchers (coming from both academia and industry), and biobankers are mostly seen as the primary users of BBMRI-ERIC, other users are also embraced and supported, such as patients/donors and their organizations, data protection agencies and research funding agencies are also part of the target users. Furthermore, even for the researchers, the use cases go beyond well-known sample/data request use case: recent investigations by BBMRI.uk[1] have shown that sample/data storage and curation requests may be as frequent, and industry is specifically known for joint prospective studies with biobanks instead of requesting existing samples[2].

The IT infrastructure of BBMRI-ERIC will be developed and opearated using Common Service IT instrument, to which all the full-member countries of BBMRI-ERIC contribute. It follows up on experience from the BBMRI Preparatory Phase[3] as well as collaboration within

[1] Results have not been published yet.

[2] The reasons for this range from the informed consent signed by the patients/donors to tighter control over the sample collection/processing/storage requirements.

[3] Material from BBMRI Preparatory Phase can be found at `http://bbmri-eric.eu/reports`

other projects in the BBMRI ecosystem, such as BBMRI-LPC[4], BioSHaRE[5], BioMedBridges[6], or BiobankCloud[7].

2 Functional Description

BBMRI-ERIC relies on a component-based software stack with well-defined components of reasonable size (preferably not excessively large), interconnected using well-defined and well-documented APIs. The component diagram is shown in Figure 1 and the components are described in further detail in Sections 3 and 4. Architecture of the system is fully distributed, following distributed architecture of BBMRI-ERIC itself, where it is called "hub and spokes" with central level, national nodes level, and individual biobanks level. This architecture is applied to all the aspects including the long-term data storage and curation, querying data, migration of computations to data, etc. The architecture, however, must support temporary data caching for performance reasons. From this perspective, BBMRI-ERIC has no ambition to setup large central storage facilities, although some members or specific BBMRI-ERIC-related projects may opt for aggregation of data into highly secure storage systems.

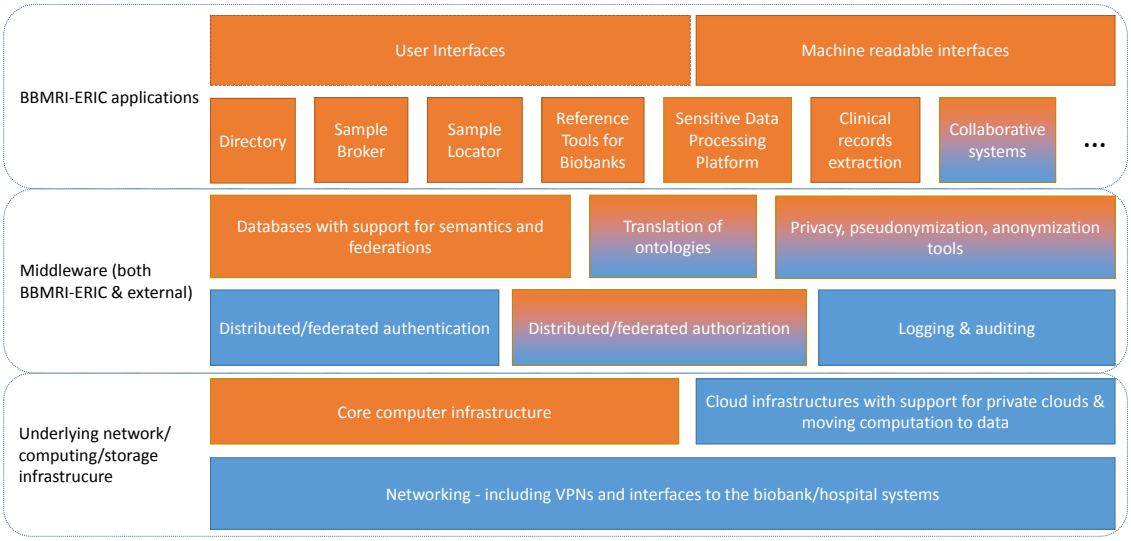

Figure 1: Software stack of BBMRI-ERIC IT system. Orange components are assumed to be build by BBMRI-ERIC, blue components are expected from other e-Infrastructures. Orange-blue components are assumed to be developed jointly with other e-Infrastructures.

From the data exchange perspective, BBMRI-ERIC is committed to FAIR principles[8] (Findable, Accessible, Interoperable, Reusable), with accessibility the limited by privacy protection of

[4] http://www.bbmri-lpc.org/
[5] https://www.bioshare.eu/
[6] http://www.biomedbridges.eu/
[7] http://www.biobankcloud.com/
[8] Data FAIRport, http://datafairport.org/

patients and donors given the nature of data in BBMRI-ERIC infrastructure. This implies that access is only provided to the authorized people, i.e., typically researchers who work on ethically approved research projects.

First step toward these goals is identifiability and traceability, which requires identifiers with time-stamping support for both data and samples, as well as ability to identify subsets generated by queries on data in specific time. This is well aligned with one of the major lines of RDA focus.

Typical workflow for the user starts with authenticated user[9] searching for the samples and/or data, or trying to identify biobanks to start collaboration with (see the Directory and Sample Broker/Locator components described in Section 3). Before accessing samples and/or actual data, the user must submit a project that undergoes ethical evaluation, and only users with approved projects may be allowed any further. The users then request the samples and/or data and negotiates with biobankers. At this step, the user's request may still be rejected for several reasons: the samples or data may not be fit for the intended purposes, the sample may be reserved for another project with higher priority or for another purpose (e.g., biobanks make certain samples reserved for quality management purposes including verification of previous experiments in case of dispute). Once user's request is approved, the user signs Material Transfer Agreement (MTA) and the sample/data is given to the user.

When processing data, the sensitive nature of the data may require that raw data never leaves biobank and only the aggregate anonymized data is sent out, as has been previously described and demonstrated, e.g., using DataSHIELD[10] [7, 8, 9]. Both size of the data and its nature will be helped by the moving computations to data paradigm that has been promoted in last 10 years and that has been strongly pushed forward by the availability of clouds that can be deployed also within the perimeter of a biobank; use of private clouds for processing of biobank data has been developed and demonstrated by the BiobankCloud project[11]. An extended version of this scenario is targeted by the Sensitive Data Processing Platform component in the software stack diagram.

Another specific aspect of BBMRI-ERIC infrastructure is the heterogeniety of data that are coming into the biobanks and that need to be harmonized into consistent data sets. Therefore BBMRI-ERIC works with the federated databases with semantic data support (triple store systems) and translation of ontologies, which has been being worked upon, e.g., in the BioMedBridges project[12]. Specific issue for the clinical biobanks is the unstructure clinical records that are on one hand one of the most valuable sources of information, but on the other hand that in many cases require reliable extraction from unstructured records written in the natural language.

[9] Strong authentication is needed, preferably multi-factor, because of the privacy and security aspects.
[10] http://www.p3g.org/biobank-toolkit/datashaper
[11] http://www.biobankcloud.com/
[12] http://www.biomedbridges.eu/

3 Describe Essential Components and Their Services

BBMRI-ERIC Directory A distributed tool to provide highly aggregated information about biobanks, biobank networks, sample and data collections, and studies. This tool is primarily intended for the researchers to identify biobanks that might potentially have samples/data of their interest. The data is typically collected from the local biobanks via national nodes to the central level of BBMRI-ERIC, while national nodes utilize this structure to also run their national directories. This tool is used to assign identifiers to all the entities (biobanks, biobank networks, sample and data collections, studies), which can be further used not only for reproducibility and traceability, but also to assess their impact[13].

Sample Broker This tool is intended for the researchers who already have their research intent/project and need samples or data to implement it. Inquiries by the researchers for the samples often span multiple biobanks and they are subject to iterative refinement. As a part of this process, the biobankers must understand various aspects of the expected methods to be used in the planned research, in order to evaluate whether their samples are fit for the particular purpose (e.g., analytical method). This is by its nature a M:N communication between researchers and biobankers, generating large overhead that can be simplified by employing efficient tools for group communication.

Sample Locator If there were no privacy concerns (e.g., in case of non-human biosamples), the researchers could easily look up individual samples of their interest based on parametric search. For BBMRI-ERIC, the situation is, however, more complicated because and various strategies related to differential privacy [12, 13, 14] need to be in place. Approaches such as k-anonymity, l-diversity, and t-closeness together with generalization and suppression may result in substantial "hidden black matter" because in practice the high-dimensional data is sparse [15]. An alternative solution to avoid too much supression is by reducing dimensionality, which may in turn result in users being unable to ask as specific queries as they need. Another aspect is competing interests of biobankers and researchers, which results in biobankers being reluctant to put all of their samples into a system that can identify individual samples. Despite the fact that only subset of samples and data is assumed to be available through this tool, it will still be part of the overall system because of its unique capability to support generation of novel research ideas.

Ontology Translation Service With distributed nature of BBMRI-ERIC, the data come in many different ontologies even in a single domain[15]. As data harmonization and ontology translation is an extremely important service for many other tools, we define it as a separate component with well-defined interface to be incorporated into other applications.

[13] See, e.g., BioResource Impact Factor (BRIF)[14] [10, 11].

[15] A nice illustration is simple diagnosis coding, where not all the European countries use standard ICD-10 system and some use nationally customized variants of it of or customized variants of SNOMED CT.

Sensitive Data Processing and Sharing Platform This component is composed of two parts: one is the private cloud-based tools for biobanks and the other is a platform where sensitive data can be collected and shared, such as TSD[16] or MOSLER[17].

4 Describe Optional/Discipline-Specific Components and Their Services

Clinical Records Extraction Clinical records are valuable source of information especially for the clinical biobanks, which take biosamples from the clinical practice. Typical clinical records, however, contain only limited structured information and large portions are written as free text in natural language, often with some particular domain specifics. In many cases, there is further complication for the biobanks that they are detached from the hospital information systems and may not access this data online. While very important and characteristic for BBMRI-ERIC, reliable extraction from the unstructured clinical records is still an open basic research problem to a large extent and therefore it is in the optional components list.

Reference Tools for National Nodes and Biobanks Because biobanks and BBMRI-ERIC national nodes have often very limited IT personnel capacity, BBMRI-ERIC is committed to provide reference tools for both of these levels. These tools are assumed to be distributed either as software packages or even as pre-installed and mostly pre-configured virtual machines.

An important aspect of the reference tools will be documentation of APIs and file formats used for the data exchange, as biobanks and national nodes will be free to replace any of the components of the reference toolset by the tools of their preference, only retaining the API interoperability.

5 Describe Essentials of the Underlying Data Organization

The schema below tries to provide an overview of data organization. Please note there are two major types of biobanks that differ in how they store and access data in most cases: (a) population biobanks, which typically store all the relevant data inside the biobank together with the biosamples, (b) clinical biobanks, which rely on their connection to the clinical source of biosamples/data (hospital or other healthcare provider) and which typically need to query that source for more detailed data beyond very basic data structure that is transferred initially together with the biosample.

(1) **Data stored inside a biobank.**

This is data that is stored within physical or at least logical perimeter of the biobank. Typically comprises several subtypes:

[16] https://www.uio.no/tjenester/it/forskning/sensitiv/
[17] https://wiki.bils.se/wiki/Mosler_user_documentation

(2a) **Data generated inside a biobank.**

Typically operational data related to the biosamples, such as their position in storage systems. In some cases, biobanks also perform further biosample analysis on their own, such as sequencing.

Example data: location information of biosamples (in storage system).

(2b) **Data received together with the biosample and stored in a biobank.**

This is the data the comes into the biobank as a part of ingestion of the biosample into the biobank storage system. For clinical biobanks, it may consist of a subset of structured clinical data, while for population biobanks it may contain complete data set collected in the research/study about the donor.

Example data: (a) description of the sample (information on how and when the sample was taken and processed), (b) excerpt of structured patient's clinical data (pre-approved structure – typical for the clinical biobanks), (c) donor-related information related to the purpose of the research or biobank, such as life-style data, phenotype data, etc. (typical for the population biobanks).

(3c) **Data generated outside biobank and stored in a biobank.**

Example data: omics data generated by a user of a biobank, which is returned back to the biobank.

(2) **Data used by biobanks but stored outside the biobank.**

This category is typical for clinical biobanks detached from the hospital on technical or administrative basis[18]. For any data access that is not part of the initial data transfer with the biosample (Item (2b)), the biobank needs to apply for the data to the hospital information system managers.

Example data: clinical records of patients.

(3) **Data stored at national level.**

Amount and types of the data stored on this level varies largely based on the type of the national node. Typically consists of administrative/operational data and data linking to the biobanks. For some (typically smaller) national nodes, it may also store some data on behalf of the biobanks.

Example data: (a) Lists of interfaces to the biobanks, (b) authorization data for the services on the national level, (c) access/usage logs, (d) data query caches, (e) registry data on behalf of biobanks (if there is no on-line interface for the biobank).

(4) **Data stored at central BBMRI-ERIC level.**

This typically consists of administrative/operational data and data linking national nodes to the central BBMRI-ERIC level. BBMRI-ERIC intentionally avoid storing any privacy-sensitive data on the central level.

Example data: (a) Lists of interfaces to the national node services and service discovery, (b) authorization data for the services on the central BBMRI-ERIC level, (c) access/usage logs, (d) data query caches.

[18]This happens often that biobanks are considered research instructures and as a part of their institutionalization, they become detached from the clinical network in the hospital and from the hospital information systems, even though they may still reside in the same hospital premise.

(5) **Data stored outside of EU.**
This data may consist of any of the previously described data types (Items (1)–(4)), but regulations of other countries as well as European Union apply, if integrated into BBMRI-ERIC.

As one can see from the list above, BBMRI-ERIC features fully federated distributed architecture with distributed databases in autonomous organizations and organizational units (working under same umbrella of BBMRI-ERIC allowing for the federated operations) and distributed querying.

Data life cycle and traceability. An important aspect for traceability is data modifications/updates, which are an inherent part of the data life cycle in the BBMRI-ERIC ecosystem. This aspect is particularly critical for the clinical biobanks, where the data coming from the clinical practice may come in largely varying quality and may require several rounds of refinement before they become usable for further research. The issue of data improvements and fixes should not be underestimated, however, even for other types of biobanks. The primary data can be only edited on the level where they are stored, see the Items (1)–(5). All the changes must result in a traceable and identifiable changes that can be used, e.g., in the provenance graphs [16, 17].

6 Indicate the Type of APIs behind Used

The most common interfaces in the BBMRI-ERIC community are REST interfaces. For linked data, JSON-LD and less frequently RDF is being used with Virtuoso[19] used as triple store database.

Other interfaces are used as appropriate for given applications. For example Directory 1.0 relies on hierarchy of LDAP servers (national nodes can run their own LDAP servers, or can upload LDIF/JSON data directly to the central server) and LDIF data format for distributed data queries and JSON translators are available in/out for the LDAP.

When dealing with the clinical data, hospital information systems rely on HL7 (Health Level 7)[20] as well as custom interfaces. Data often utilize PACS formats (if relevant for given data type, e.g., imaging). There is ongoing work on harmonization of Electronic Health Records (EHR) within HL7 called Fast Healthcare Interoperability Resources (FHIR)[21], which in turn relies again on REST.

National nodes and local biobanks run variety of systems and APIs and it is one of the major goals of BBMRI-ERIC to simplify the situation by providing reference tools for the national nodes and biobanks.

[19] http://virtuoso.openlinksw.com/
[20] http://www.hl7.org/
[21] Pronounced “fire”, http://hl7.org/implement/standards/fhir/ .

As a part of the efforts to improve quality and interoperability of APIs and data formats, BBMRI-ERIC actively participates in ISO TC 276[22] Working Group 5 (WG5) "Data processing and integration", which aims at (a) definition of data and model formats and their interfaces; (b) definition of metadata and relations of data and models; (c) quality management of processed data and models. In order to provide consistent input, BBMRI-ERIC also participates in ISO TC 276 WG1 (terminology) and WG2 (biobanking).

7 Achieved Results

At the time of writing, BBMRI-ERIC is running collaborative tools to support interaction of its community, released Directory 1.0 covering more than 500 biobanks and standalone collections with overall estimated size between 34,000,000 and 46,000,000 samples[23] and Common Service IT is under setup process with expected start in Fall 2015. National nodes are running their own infrastructures of highly varying extent and quality, as do also local biobanks. BBMRI-ERIC also benefits from other related activities such as operation Catalogue of BBMRI-LPC providing data warehouse capabilities.

[22] http://www.iso.org/iso/home/standards_development/list_of_iso_technical_committees/iso_technical_committee.htm?commid=4514241

[23] Only an order of magnitude of biobanks and standalone collections has been collected for the first release of Directory (1.0), in order to avoid frequent data updates on the biobankers side, as for many biobanks the amount of samples is constantly changing and there is so far no automatic link between the biobank and the national/central levels of the Directory.

References

[1] John PA Ioannidis, David B Allison, Catherine A Ball, Issa Coulibaly, Xiangqin Cui, Aedín C Culhane, Mario Falchi, Cesare Furlanello, Laurence Game, Giuseppe Jurman, et al. Repeatability of published microarray gene expression analyses. *Nature genetics*, 41(2):149–155, 2009.

[2] Florian Prinz, Thomas Schlange, and Khusru Asadullah. Believe it or not: how much can we rely on published data on potential drug targets? *Nature reviews Drug discovery*, 10(9):712, 2011.

[3] C Glenn Begley and Lee M Ellis. Drug development: Raise standards for preclinical cancer research. *Nature*, 483(7391):531–533, 2012.

[4] Mina Bissell. Reproducibility: The risks of the replication drive. *Nature*, 503:333–334, 2013.

[5] Sean J Morrison. Time to do something about reproducibility. *eLife*, 3:e03981, 2014.

[6] Peter AC't Hoen, Marc R Friedländer, Jonas Almlöf, Michael Sammeth, Irina Pulyakhina, Seyed Yahya Anvar, Jeroen FJ Laros, Henk PJ Buermans, Olof Karlberg, Mathias Brännvall, et al. Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. *Nature biotechnology*, 31(11):1015–1022, 2013.

[7] Michael Wolfson, Susan E Wallace, Nicholas Masca, Geoff Rowe, Nuala A Sheehan, Vincent Ferretti, Philippe LaFlamme, Martin D Tobin, John Macleod, Julian Little, et al. Datashield: resolving a conflict in contemporary bioscience—performing a pooled analysis of individual-level data without sharing the data. *International journal of epidemiology*, page dyq111, 2010.

[8] EM Jones, NA Sheehan, N Masca, SE Wallace, MJ Murtagh, and PR Burton. Datashield–shared individual-level analysis without sharing the data: a biostatistical perspective. *Norsk epidemiologi*, 21(2), 2012.

[9] Amadou Gaye, Yannick Marcon, Julia Isaeva, Philippe LaFlamme, Andrew Turner, Elinor M Jones, Joel Minion, Andrew W Boyd, Christopher J Newby, Marja-Liisa Nuotio, et al. Datashield: taking the analysis to the data, not the data to the analysis. *International journal of epidemiology*, 43(6):1929–1944, 2014.

[10] Anne Cambon-Thomsen, Gudmundur A Thorisson, Laurence Mabile, Sandrine Andrieu, Gabrielle Bertier, Martin Boeckhout, Jane Carpenter, Georges Dagher, Raymond Dalgleish, Mylène Deschênes, et al. The role of a bioresource research impact factor as an incentive to share human bioresources. *Nature Genetics*, 43(6):503–504, 2011.

[11] Laurence Mabile, Raymond Dalgleish, Gudmundur A Thorisson, Mylène Deschênes, Robert Hewitt, Jane Carpenter, Elena Bravo, Mirella Filocamo, Pierre Antoine Gourraud,

Jennifer R Harris, et al. Quantifying the use of bioresources for promoting their sharing in scientific research. *Gigascience*, 2(1):7, 2013.

[12] Cynthia Dwork. Differential privacy: A survey of results. In *Theory and applications of models of computation*, pages 1–19. Springer, 2008.

[13] Ninghui Li, Wahbeh H Qardaji, and Dong Su. Provably private data anonymization: Or, k-anonymity meets differential privacy. *CoRR, abs/1101.2604*, 49:55, 2011.

[14] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Theoretical Computer Science*, 9(3-4):211–407, 2013.

[15] Charu C Aggarwal. On k-anonymity and the curse of dimensionality. In *Proceedings of the 31st international conference on Very large data bases*, pages 901–909. VLDB Endowment, 2005.

[16] Gianmauro Cuccuru, Simone Leo, Luca Lianas, Michele Muggiri, Andrea Pinna, Luca Pireddu, Paolo Uva, Alessio Angius, Giorgio Fotia, and Gianluigi Zanetti. An automated infrastructure to support high-throughput bioinformatics. In *High Performance Computing & Simulation (HPCS), 2014 International Conference on*, pages 600–607. IEEE, 2014.

[17] Gianluigi Zanetti. Data intensive biology and data provenance graphs, February 2015. BiobankCloud Project. URL: `http://www.biobankcloud.com/sites/default/files/ngshadoop/hadoop_ngs_zanetti.pdf`.