University of Groningen

A Decentralized IT Architecture for Locating and Negotiating Access to Biobank Samples

Proynova, Rumyana; Alexandre, Diogo; Lablans, Martin; Van Enckevort, David; Mate, Sebastian; Eklund, Niina; Silander, Kaisa; Hummel, Michael; Holub, Petr; Ückert, Frank

Published in:
German Medical Data Sciences

DOI:
10.3233/978-1-61499-808-2-75

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2017

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):
Proynova, R., Alexandre, D., Lablans, M., Van Enckevort, D., Mate, S., Eklund, N., Silander, K., Hummel, M., Holub, P., & Ückert, F. (2017). A Decentralized IT Architecture for Locating and Negotiating Access to Biobank Samples. In R. Rohrig, U. Sax, A. Timmer, & H. Binder (Eds.), *German Medical Data Sciences: Visions and Bridges - Proceedings of the 62nd Annual Meeting of the German Association of Medical Informatics, Biometry and Epidemiology (gmds e.V.), GMDS 2017* (pp. 75-79). (Studies in Health Technology and Informatics; Vol. 243). IOS Press. https://doi.org/10.3233/978-1-61499-808-2-75

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license. More information can be found on the University of Groningen website: https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment.

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Download date: 21-01-2023

R. Röhrig et al. (Eds.)

© 2017 German Association for Medical Informatics, Biometry and Epidemiology (gmds) e.V. and IOS Press.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).

doi:10.3233/978-1-61499-808-2-75

A Decentralized IT Architecture for Locating and Negotiating Access to Biobank Samples

Rumyana PROYNOVA [a,1], Diogo ALEXANDRE [a], Martin LABLANS, [a] David VAN ENCKEVORT[b], Sebastian MATE[c], Niina EKLUND[d], Kaisa SILANDER[d], Michael HUMMEL[e], Petr HOLUB[f] and Frank ÜCKERT [a]

[a] *German Cancer Research Center, Heidelberg.&BBMRI.de*
[b] *Universitair Medisch Centrum Groningen, Groningen & BBMRI.nl*
[c] *Friedrich Alexander Universität, Erlangen-Nürnberg & BBMRI.de*
[d] *Terveyden Ja Hyvinvoinnin Laitos, Helsinki & BBMRI.fi*
[e] *Charité, Berlin & BBMRI.de*
[f] *BBMRI-ERIC, Graz*

Abstract. There is a need among researchers for the easy discoverability of biobank samples. Currently, there is no uniform way for finding samples and negotiate access. Instead, researchers have to communicate with each biobank separately. We present the architecture for the BBMRI-CS IT platform, whose goal is to facilitate sample location and access. We chose a decentral approach, which allows for strong data protection and provides the high flexibility needed in the highly heterogeneous landscape of European biobanks. This is the first implementation of a decentral search in the biobank field. With the addition of a Negotiator component, it also allows for easy communication and a follow-through of the lengthy approval process for accessing samples.

Keywords. Biological specimen banks, Information storage and retrieval, Decentral search, European research infrastructure

1. Introduction

Biobanks are an essential part of the research infrastructure in the life sciences. They offer access to human biological samples, as well as the associated epidemiological, clinical, biological, genealogical and molecular information, which is critical for researchers who work on the major challenges that medical science is facing today.

The European biobank landscape is highly heterogeneous. Each biobank uses its own process for sample requests. Researchers have to first locate candidate biobanks, then, if the needed samples are available, negotiate access with each biobank separately. Biobanks desire more external cooperation, but find that few requests reach them.

BBMRI-ERIC is a distributed research infrastructure of biobanks and biomedical resources [1]. One of its aims is to provide the Common Service IT (CS-IT) platform, which supports researchers in locating samples and gaining access to them, providing a

[1] Corresponding Author: Rumyana Proynova, Medical Informatics in Translational Oncology, German Cancer Research Center, Im Neuenheimer Feld 580, 69120 Heidelberg, Germany. r.proynova@dkfz.de.

unified search process, while being flexible enough to account for the differences between biobanks. It allows for focused searches on the sample level, cutting down on the inefficiencies inherent in sending and processing queries to biobanks which might or might not have the desired material.

In work with stakeholders, Lablans [2] identified the major requirements:

- **Technical heterogeneity**. The system should be able to represent sample data from different source systems without the need for manual reentry.
- **Semantic interoperability**. The system has to make the sample descriptions from different organizations comparable to each other.
- **Minimal effort**. Participating in the system should require minimal effort.
- **Data minimization**. The system should avoid saving sample data outside of the biobank where possible.
- **Data sovereignty**. The biobanker should control which data leaves the biobank, and there should be no pressure to justify a decision to deny access.

2. State of the art

Several platforms already facilitate to some degree sample access to biobanks. They all offer partial solution of the access problem, but cannot function as a comprehensive solution on the European level.

Some biobanks provide a sample search on the Internet, for example the Auria biobank [3]. This type of system best represents the dataset of one biobank, and grants perfect data sovereignty, but is not interoperable with other biobanks and requires the researcher to actively seek out a biobank.

A more interconnected approach is seen in biobank registries such as the German biobank registry [4]. The registry approach requires the biobanker to preemptively provide data, which hurts the requirements of minimal effort, data minimization and sovereignty. Its minimal dataset only allows a search with low precision and recall. For situations where this is sufficient, the registry-based approach offers an attractive option for low-friction data access. Centralized search systems such as CRIP [5] work on a similar principle and have similar advantages and disadvantages.

All these options are focused on discovering potentially collaborating biobanks. To our knowledge, none of them supports the subsequent negotiation process.

3. Concept

The BBMRI CS-IT platform allows for the connection of large numbers of biobanks. Its architecture follows the principles of the “decentral search” [6] developed in the German Cancer Consortium [7]. Its main advantage is that the information never leaves the biobank uncontrolled, allowing the biobanks to retain data sovereignty and implement privacy protection, unlike the central solutions discussed in the previous section. Figure 1 shows a diagram of the architecture.

To address the problem of semantic interoperability, BBMRI-ERIC supports a set of data models. The central one is the BBMRI-ERIC core terminology, which encompasses a minimal list of data elements, related to the MIABIS dataset [8, 8]. It is extended by optional purpose-specific data models. The core terminology and the extension data models are defined as data dictionaries contained in a *Metadata repository* (MDR) based on the ISO/IEC 11179 standard [9].

When a biobank joins the platform, its data has to be harmonized. BBMRI-ERIC provides a suite of *Mapping & ETL* tools, with which the biobanker maps the source data structure to one or more of the data dictionaries supported by the platform. After the mapping of the metadata, the data itself is uploaded into a *local data silo*. Unlike the source systems, it contains harmonized data. The biobank also installs a *Connector*. It can execute queries against the dataset in the data silo, but does not actively send information to the outside.

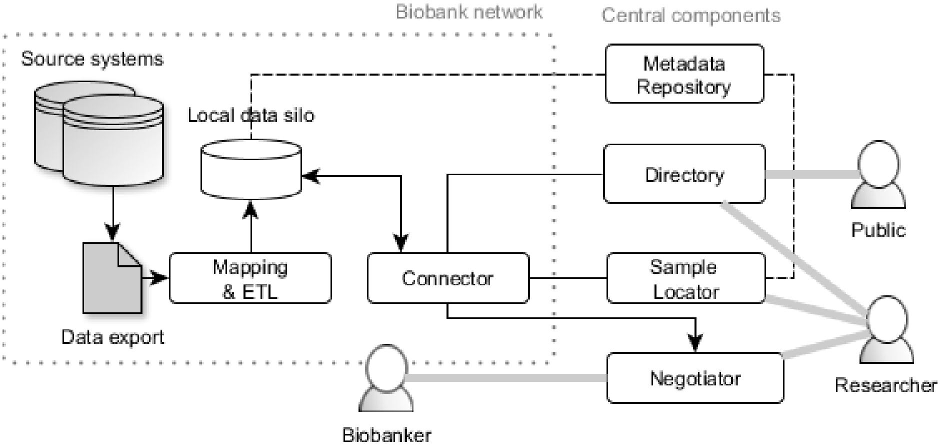

Figure 1. Architecture of the BBMRI-ERIC CS-IT platform

When a researcher needs samples, he or she first interacts with the *Sample Locator* component, which offers an interface for the construction of search queries based on the data elements provided by the MDR. The next step is to enrich the request by composing a short description in the *Negotiator*. The connectors fetch the query from the Sample Locator and execute it locally. If the data silo has matching samples, the Connector notifies the biobanker that a request is waiting. At that point, nobody outside of the biobank knows which biobanks matched the query.

The biobanker visits the Negotiator and uses the freetext description and the query criteria to determine if his or her biobank can actually provide the desired samples. As the request can be underdefined, the Negotiator allows the biobanker to ask for clarification and guide the researcher in improving the query criteria to better match his or her research question. When a biobank is confident that it can provide the samples, the biobanker can start a confidential conversation with the researcher to discuss the logistical, organizational and legal details.

This process offers high data protection, but may become cumbersome to the users. Its complexity is only needed when sample access is desired. Sometimes researchers or members of the public need non-sensitive information only, and the overhead of the sample request process is not necessary. For this use case, the platform also incorporates the *Directory [10]. T*his is a registry-based service, offering low-detail information not violating the donors' privacy. It has a search interface that delivers immediate search results. For the convenience of biobankers, they can update their biobank's information in the Directory through the Connector.

4. Implementation

The implementation of the BBMRI-ERIC CS-IT platform follows an agile approach, stepwise releasing the components in a way that the system can be used to some extent even before the planned scope has been completed. All components are released as free and open source software.

At the time of this writing, the Directory is in active use and contains the data of over 1000 biobanks. The Negotiator is in a rollout phase, and in use by a small number of piloting biobanks, allowing for negotiation after a Directory search. The MDR is completed and contains two datasets: a first version of the BBMRI core terminology, and a colon cancer data dictionary as a first example of an extension data dictionary.

The connector, the local data silo, and ETL tools for a user-friendly mapping process are under active development. The public release is planned in fall 2017.

The Sample Locator is not yet implemented. It is expected to be released in 2018. Until then, users can locate biobanks with candidate samples by using the Directory, and proceed from here to negotiation.

5. Discussion

The architecture presented here meets the five major requirements elicited from stakeholders. Transferring the source data into a data silo using a unified structure solves the problem of technical interoperability. The MDR-driven approach with a core dataset and extensions is a scalable approach to the semantic interoperability problem. As both the data silo and the connector remain within the biobank, the principles of data sovereignty and data minimization are upheld – information is only transmitted with regard to a request, following approval of a biobanker.

The effort needed to setup the system is not trivial, but we argue that it is still small in comparison to the gains from participation. The highest additional effort for the biobank is the one-time process of mapping the source data to the provided data dictionaries. The ETL tools are usable by domain experts without IT background. So while the goal of minimal effort is not completely reached, it is still at a very good level and allows for realistic system adoption.

The decentral search approach is not as convenient for the requester as the immediate serving of search results familiar from other domains. We chose this solution as a response to the need for confidentiality and data protection inherent in biological sample data, since even non-identifying information such as a diagnosis is subject to some confidentiality under current social and legal norms. The non-sensitive data about biobanks and their samples can be found in the Directory, which offers a conventional search with immediate results.

The Negotiator is a novel component, which has not been employed in previous search systems. It lets humans take over the communication after the search has discovered which potential communication partners have relevant samples. This adds flexibility to the system, allowing it to support a large variety of collaboration models, and making it robust to imperfect search queries and missing data sources.

System success is highly dependent on gaining acceptance from biobanks and ensuring high data quality. To achieve this, we are engaged in dialogue with selected biobanks and BBMRI's national nodes since the earliest project phases. We rely on their feedback for creating a system which they are willing and capable of using.

6. Conclusion

In this paper, we present an architecture for a platform which allows researchers to locate and request samples from biobanks. It is an implementation of a decentral search approach, enriched by a Negotiator component which supports a flexible approval and collaboration process between biobankers and researchers.

Some components of the future system are already released and are well-received among users. Close work with stakeholders ensures that the remaining components will also meet the users' needs and allows us to overcome the challenges inherent in the creation of a system of this scale.

7. Conflict of Interest

The authors state that they have no conflict of interest.

8. Acknowledgements

We thank Saher Maqsood, Maximilian Ataian, Paul Weingardt, Polina Litvak, Jonattan Jetten, Fleur Kelpin and Mark de Haan for their work on implementing the platform. The work is part of the ADOPT BBMRI-ERIC project, funded by the EC, topic H2020-INFRADEV-3-2015, Grant Agreement Nr 676550.

References

[1] BBMRI-ERIC. What is BBMRI-ERIC?: BBMRI-ERIC [cited 2017 March 14]

[2] Lablans M. Die dezentrale Suche für die medizinische Verbundforschung. Doctoral thesis, Mathematisch-Naturwissenschaftliche Fakultät, Universität Münster 2015.

[3] Auria Biobank Catalog [cited 2017 March 14] Available from: URL: https://www.auriabiopankki.fi/katalogi/.

[4] Deutsches Biobankregister [cited 2017 March 14] Available from: URL: http://dbr.biobanken.de/de/web/guest/bdb/.

[5] Schroder C, Heidtke KR, Zacherl N, Zatloukal K, Taupitz J. Safeguarding donors' personal rights and biobank autonomy in biobank networks: the CRIP privacy regime. Cell Tissue Bank 2011; 12(3): 233–40 [https://doi.org/10.1007/s10561-010-9190-8][PMID: 20632213]

[6] Lablans M, Kadioglu D, Mate S, Leb I, Prokosch H-U, Uckert F. Strategies for biobank networks. Classification of different approaches for locating samples and an outlook on the future within the BBMRI-ERIC. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2016; 59(3): 373–8 [https://doi.org/10.1007/s00103-015-2299-y][PMID: 26753865]

[7] Lablans M, Kadioglu D, Muscholl M, Uckert F. Exploiting Distributed, Heterogeneous and Sensitive Data Stocks while Maintaining the Owner's Data Sovereignty. Methods Inf Med 2015; 54(4): 346–52 [https://doi.org/10.3414/ME14-01-0137][PMID: 26196653]

[8] Norlin L, Fransson MN, Eriksson M*, et al.* A Minimum Data Set for Sharing Biobank Samples, Information, and Data: MIABIS. Biopreserv Biobank 2012; 10(4): 343–8 [https://doi.org/10.1089/bio.2012.0003][PMID: 24849882]

[9] ISO/IEC. ISO/IEC 11179, Information Technology -- Metadata registries (MDR); 2015 2015 Nov 5 [cited 2017 March 16]

[10] van Enckevort D, Reihs R, Swertz M*, et al.* BBMRI-ERIC directory: Metadata and aggregate data about biobanks and other bioresources 2016.