doi:10.3233/978-1-61499-808-2-100

Proof-of-Concept Integration of Heterogeneous Biobank IT Infrastructures into a Hybrid Biobanking Network

Sebastian MATE [a,1], Dennis KADIOGLU [a,b], Raphael W. MAJEED [c], Mark R. STÖHR [c], Michael FOLZ [d], Patric VORMSTEIN [d,e], Holger STORF [d,e], Daniel P. BRUCKER [e,f], Dietmar KEUNE [g], Norman ZERBE [h], Michael HUMMEL [h,i], Karsten SENGHAS [j], Hans-Ulrich PROKOSCH [a] and Martin LABLANS [j]

[a] *Medical Informatics, Univ. of Erlangen-Nürnberg, Erlangen, Germany*
[b] *Institute of Medical Biostatistics, Epidemiology and Informatics, University Medical Center of the Johannes Gutenberg University Mainz, Mainz, Germany*
[c] *UGMLC, German Center for Lung Research (DZL), Justus-Liebig-University, Giessen, Germany*
[d] *Medical Informatics Group, University Hospital Frankfurt, Frankfurt, Germany*
[e] *German Cancer Consortium (DKTK), partner site Frankfurt; and German Cancer Research Center (DKFZ), Heidelberg, Germany*
[f] *University Cancer Center (UCT) Frankfurt, University Hospital Frankfurt, Frankfurt, Germany*
[g] *Clinical Cancer Registry, Charité Comprehensive Cancer Center, Berlin, Germany*
[h] *Institute of Pathology, Charité - Universitätsmedizin Berlin, Berlin, Germany; and Central Biobank Charité (ZeBanC), Berlin, Germany*
[i] *German Biobank Node*
[j] *Medical Informatics in Translational Oncology, German Cancer Research Center (DKFZ), Heidelberg, Germany*

Abstract. Cross-institutional biobank networks hold the promise of supporting medicine by enabling the exchange of associated samples for research purposes. Various initiatives, such as BBMRI-ERIC and German Biobank Node (GBN), aim to interconnect biobanks for enabling the compilation of joint biomaterial collections. However, building software platforms to facilitate such collaboration is challenging due to the heterogeneity of existing biobank IT infrastructures and the necessary efforts for installing and maintaining additional software components. As a remedy, this paper presents the concept of a hybrid network for interconnecting already existing software components commonly found in biobanks and a proof-of-concept implementation of two prototypes involving four biobanks of the German Biobank Node. Here we demonstrate the successful bridging of two IT systems found in many German biobanks – Samply and i2b2.

Keywords. Translational medical research, biobank network, federated queries, cohort identification, German Biobank Node

[1] Corresponding Author: Sebastian Mate, Wetterkreuz 13, 91058 Erlangen-Tennenlohe, Germany; E-mail: Sebastian.Mate@fau.de.

1. Introduction

Biobanks play a pivotal role in facilitating biomedical research in the era of personalized medicine. As research infrastructures, they effectively support the discovery and validation of molecular disease mechanisms and biomarker detection, which will ultimately lead to deep insights into disease pathogenesis and allow the development of innovative treatment options. Such new knowledge can be incorporated into the assessment and stratification of risk factors, new diagnostic methods, pharmacogenomics, and drug development [1]. However, access to samples and associated clinical data of suitable quality is still one of the major challenges (see e.g. [2]). Since almost all diseases are composed of highly diverse molecular subgroups, it becomes more and more difficult, even for large biobanks, to provide sufficient samples and data of a certain molecular subgroup to provide statistical rigor to a study [3,4]. Researchers increasingly require large and sufficiently characterized data sets to uncover the subtle statistical associations between phenotypes and diseases. As stated in [5], “biobanking is required to change strategic focus from a sample dominated perspective to a data-centric strategy.” Thus, it is desirable to merge data from multiple biobanks for further analysis [6].

BBMRI-ERIC is a European research infrastructure aiming to interconnect high quality biobanks all over Europe via a federation of national nodes. The planned software platform will enable researchers to identify samples by running queries across participating biobanks. The German Biobank Node (GBN) has been established as one of the BBMRI-ERIC national nodes [7]. In the first funding phase of GBN, it was the goal to design and evaluate the concept of an architecture for integrating biobanks with different local data warehouse (DWH) implementations into one network, and allowing queries from a single consistent user interface. The aim of this paper is to describe the development and lessons learned from two prototypical cross-biobank query implementations within a hybrid IT infrastructure.

2. Methods

The motivation within GBN was to develop a platform that is easily adoptable for the majority of biobanks. Hence, we intended to base our approach on technology that is already available in many biobanks. To this end, we analyzed the status quo at the five GBN-coordinated centralized biobanks (cBMBs) as well as the six biobanks of the BMBF-funded German Centers for Health Research (Deutsche Zentren der Gesundheitsforschung, DZGs), and the m4 Biobank Alliance in Munich, which were considered to be representative for German biobanks in terms of IT infrastructures. We recognized that only a few already had the infrastructure to identify samples or patient cohorts. Some of these were based on the commercial software CentraXX® provided by Kairos GmbH. Many of these CentraXX®-based biobanks, such as the Charité ZeBanC in Berlin and the University Cancer Center's biobank at UCT Frankfurt also participated in the research network of the German Cancer Consortium (DKTK), for which they established “bridgeheads”, a combination of CentraXX® and the open-source software Samply [8]. Other sites implemented local research DWHs based on i2b2 [9], e.g. the DZL biobank at Giessen University Hospital and the one at the Comprehensive Cancer Center in Erlangen (CCC Erlangen-EMN). Starting from this initial analysis of available tools, we developed two prototypical GBN architectures. Finally, we evaluated both architectures at the above-mentioned four biobanks.

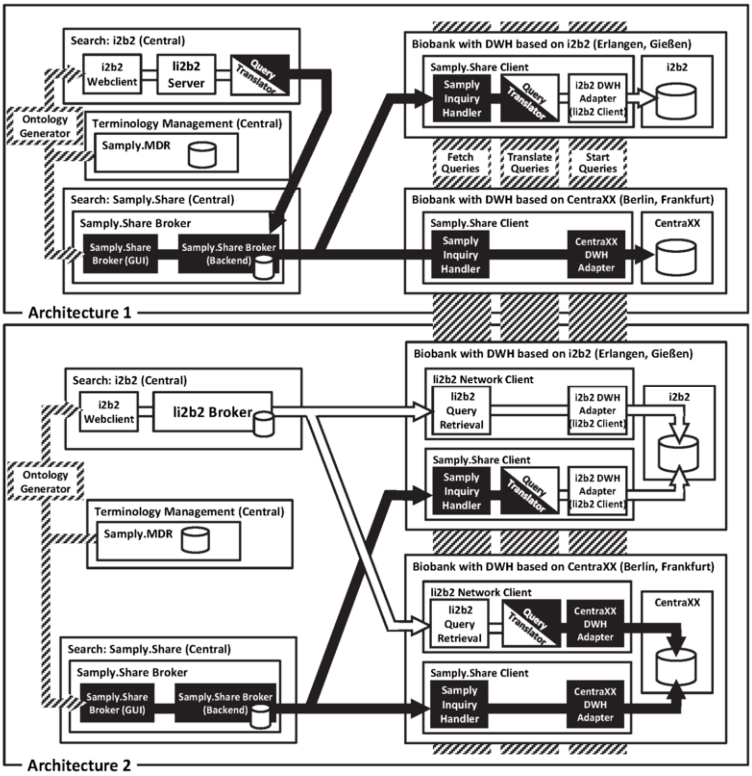

Figure 1. The two prototypical architectures that have been developed within GBN.

3. Results

The pilot was based on data stored in CentraXX® and i2b2, due to their availability at the selected sites. Two prototypical GBN architectures (as shown in Figure 1) have been developed. To enable networking with the i2b2 sites, we utilize li2b2 (https://github.com/li2b2) and not the better-known SHRINE [10], which uses a query push mechanism that is incompatible with firewalls usually installed at university hospitals. The central components, which are used for central terminology management and for issuing queries, are depicted on the left, local biobank components on the right side. i2b2/li2b2-related components and messages are shown in white, Samply/CentraXX®-based ones in black. The main difference between the architectures is that in architecture 1, outgoing queries are managed in a single queue (the *Samply.Share Broker*), whereas in architecture 2, a second queue (the *li2b2 Broker*) exists. Architecture 2 is in fact based on two networks that are not interconnected. This is also apparent on the right side, where for the second architecture, two clients are utilized – one for each technology (*li2b2 Network Client* and *Samply.Share Client*). As indicated by the three vertical stripes in the figure, the clients perform two to three tasks. The first one is to fetch new queries from a broker (*Samply Inquiry Handler* and *li2b2 Query Retrieval*). If necessary, a *Query Translator*, based on the *OmniQuery library* transforms the query into the syntax of the other platform, on the fly. Additionally, a mapping between the common GBN data elements from the central *Samply.MDR* to the respective local ones had to be implemented. The *Samply.Share Client* uses a mapping table, whereas the *li2b2 Network Client* uses an XSL transformation to replace concept identifiers. Finally, *DWH Adapters* issue the query to the DWH (i2b2 or CentraXX®), which reports the patient count back to the initial search component after execution.

We demonstrated both approaches to the GBN consortium in February 2017. We were able to successfully generate queries with both search clients (*i2b2 Webclient, Samply.Share Broker GUI*), automatically distribute them across our hybrid Samply/li2b2 network and execute them on the DWHs based on CentraXX® and i2b2. We currently support querying for patient phenotypes and sample characteristics (e.g. gender, type of sample). Furthermore, numeric data elements, such as laboratory measurements, can be constrained by comparator and value. As a result, the network returns the aggregated patient count for each site.

4. Discussion

The current literature presents different examples for integrating diverse data sources with information regarding biospecimens and their respective donors. These can be implemented at a single institution [11-13], in networks of institutions with identical local DWHs uploaded into central repositories [14], or they can remain in federated systems [8]. In contrast to central repositories, federated networks enable sites to retain more control of their data [15,16]. For example, the Breast Cancer Campaign Tissue Bank is a collaboration of four individual biobanks in the UK, and a platform has been developed to address the challenges of running a distributed network [17].

Our approach integrates already available local components into a federated network, thereby avoiding the establishment of new IT systems. For biobanks without existing local DWHs, this can be addressed by IT tools that have been developed recently (e.g. [18]). Our approach is also open for the integration of further platforms by implementing additional query translators.

To the best of our knowledge, our proof-of-concept is the first successful attempt to connect different DWH and federated search technologies into a single hybrid network. Architecture 1 grants biobanks a high degree of flexibility in their choice of a DWH. In our example, Frankfurt and Berlin employed their existing DKTK "bridgeheads", while Giessen and Erlangen set up an installation based on i2b2. On top of that, architecture 2 allows entire networks that had already settled on different federated search technologies to be interconnected, enabling inter-consortia queries.

However, our prototype also revealed inevitable limitations of hybrid networks. First, they introduce more complexity, as they require additional components to translate between different technologies. A certain degree of resources and technical expertise is required to design, develop, deploy, maintain and troubleshoot these additional components, especially since DWHs operate in secure clinical environments. Second, combining several technologies in consistent and comprehensible user interfaces requires determining and settling on their "lowest common denominator" in terms of features. For example, our prototype removed support for both i2b2's temporal relations in queries (as they are unsupported in CentraXX®' SQL-based query syntax) and Samply's fine-grained access control and privacy-preserving "decentral search" paradigm [8] (as i2b2 only supports *queries* and not Samply *inquiries*).

In conclusion, the high degree of flexibility of hybrid networks comes at a cost of increased complexity and reduced functionality. Obviously, biobank networks should build on a harmonized architecture and interoperable software whenever possible. Our proof-of-concept provides an important contribution by demonstrating that biobank databases and even whole existing networks can be federated across technological boundaries, such as different DWHs or query paradigms.

5. Acknowledgements

The present work has been funded by the Federal Ministry of Education and Research of Germany within the project *German Biobank Node* (project number 01EY1301). It was performed in (partial) fulfillment of the requirements for obtaining the degree “Dr. rer. biol. hum.” from the Friedrich-Alexander-Universität Erlangen-Nürnberg (SM).

References

[1] Olson JE, Bielinski SJ, Ryu E, *et al.,* Biobanks and Personalized Medicine, *Clin Genet* **86** (2014), 50.

[2] Mabile L, Dalgleish R, Thorisson GA, *et al.,* Quantifying the Use of Bioresources for Promoting their Sharing in Scientific Research, *Gigascience* **2.1** (2013), 7.

[3] Laffert von M, Penzel R, Schirmacher P, *et al.,* Multicenter ALK Testing in Non–Small-Cell Lung Cancer: Results of a Round Robin Test, *J Thorac Oncol* **9.10** (2014), 1464–1469.

[4] Quinlan PR, Mistry G, Bullbeck H, *et al.,* A Data Standard for Sourcing Fit-for-Purpose Biological Samples in an Integrated Virtual Network of Biobanks, *Biopreserv Biobank* **12** (2014), 184–91.

[5] Quinlan PR, Gardner S, Groves M, *et al.,* A Data-Centric Strategy for Modern Biobanking, *Adv Exp Med Biol* **864 (**2015), 165–169.

[6] Pang C, Hendriksen D, Dijkstra M, *et al.,* BiobankConnect: Software to Rapidly Connect Data Elements for Pooled Analysis Across Biobanks Using Ontological and Lexical Indexing, *J Am Med Inform Assoc* **22.1** (2015), 65–75.

[7] Hummel M, Rufenach C, Biomaterial Banks Are Crucial to Developing Genetically-Based Prevention Concepts, *Bundesgesundheitsblatt* **58** (2014), 127–130.

[8] Lablans M, Kadioglu D, Muscholl M, *et al.,* Exploiting Distributed, Heterogeneous and Sensitive Data Stocks While Maintaining the Owner’s Data Sovereignty, *Methods Inf Med* **54.4** (2015), 346–352.

[9] Murphy SN, Weber GM, Mendis ME, *et al.,* Serving the Enterprise and Beyond with Informatics for Integrating Biology and the Bedside (i2b2), *J Am Med Inform Assoc* **17** (2010), 124–130.

[10] McMurry AJ, Murphy SN, MacFadden D, *et al.,* SHRINE: Enabling Nationally Scalable Multi-Site Disease Studies, *PLoS ONE* **8.3** (2013), e55811.

[11] Gainer VS, Cagan A, Castro VM, *et al.,* The Biobank Portal for Partners Personalized Medicine: A Query Tool for Working with Consented Biobank Samples, Genotypes, and Phenotypes Using i2b2, *JPM* **6.1** (2016), 11.

[12] McIntosh LD, Sharma MK, *et al.,* caTissue Suite to OpenSpecimen: Developing an Extensible, Open Source, Web-Based Biobanking Management System. *J Biomed Inform* **57** (2015), 456–464.

[13] Eminaga O, Özgür E, Semjonow A, *et al.,* Linkage of Data From Diverse Data Sources (LDS): A Data Combination Model Provides Clinical Data of Corresponding Specimens in Biobanking Information System, *J Med Syst* **37.5** (2013), 9975.

[14] Oberländer M, Linnebacher M, König A, *et al.,* The “North German Tumor Bank of Colorectal Cancer”: Status Report After the First 2 Years of Support by the German Cancer Aid Foundation, *Langenbecks Arch Surg* **398.2** (2013), 251-258.

[15] Lablans M, Bartholomäus S, Ückert F, Providing Trust and Interoperability to Federate Distributed Biobanks. *Stud Health Technol Inform* 169 (2011), 644–648.

[16] Lablans M, Kadioglu D, Mate S, et al., Strategies for Biobank Networks, *Bundesgesundheitsblatt* **59.3** (2016), 373-378.

[17] Quinlan PR, Groves M, Jordan LB, *et al.,* The Informatics Challenges Facing Biobanks: A Perspective from a United Kingdom Biobanking Network, *Biopreserv Biobank* **13** (2015), 363–370.

[18] Bauer CRKD, Ganslandt T, Baum B, *et al.,* Integrated Data Repository Toolkit (IDRT). *Methods Inf Med* **55** (2016), 125–135.