Computers in Biology and Medicine 180 (2024) 108941

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

The sample locator: A federated search tool for biosamples and associated data in Europe using HL7 FHIR

Cecilia Engels [a,b,c,*], Jori Kern [b,d,e,l], Zdenka Dudová [a,c], Noemi Deppenwiese [f], Alexander Kiel [b,d,e,g,l], Björn Kroll [h], Tobias Kussel [b,d,e,l], Christina Schüttler [f,i], Radovan Tomášik [j], Michael Hummel [a,b,c], Martin Lablans [b,d,e,k,l], on behalf of theGerman Biobank Alliance (GBA) IT development team[1]

[a] German Biobank Node (GBN), Charité – Universitätsmedizin Berlin, Berlin, Germany
[b] German Cancer Consortium (DKTK), DKFZ, Heidelberg, Germany
[c] Charité University Hospital Berlin, Berlin, Germany
[d] Federated Information Systems, German Cancer Research Centre (DKFZ), Heidelberg, Germany
[e] Complex Medical Informatics, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany
[f] Medical Centre for Information and Communication Technology, Universitätsklinikum Erlangen, Erlangen, Germany
[g] Leipzig Research Centre for Civilisation Diseases, University of Leipzig, Leipzig, Germany
[h] IT Centre for Clinical Research, University of Lübeck, Lübeck, Germany
[i] Medical Informatics, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany
[j] Masaryk Memorial Cancer Institute, Brno, Czech Republic
[k] DKFZ Hector Cancer Institute at the University Medical Center Mannheim, Germany
[l] Mannheim Institute for Intelligent Systems in Medicine, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany

ARTICLE INFO

Keywords:
Biobanks
Biomedical research
Biosamples
FAIR data sharing
Feasibility search
Federated system
IT infrastructure
Software tools
Technology development

ABSTRACT

Background: This study outlines the development of a highly interoperable federated IT infrastructure for academic biobanks located at the major university hospital sites across Germany. High-quality biosamples linked to clinical data, stored in biobanks are essential for biomedical research. We aimed to facilitate the findability of these biosamples and their associated data. Networks of biobanks provide access to even larger pools of samples and data even from rare diseases and small disease subgroups. The German Biobank Alliance (GBA) established in 2017 under the umbrella of the German Biobank Node (GBN), has taken on the mission of a federated data discovery service to make biosamples and associated data available to researchers across Germany and Europe.
Methods: In this context, we identified the requirements of researchers seeking human biosamples from biobanks and the needs of biobanks for data sovereignty over their samples and data in conjunction with the sample donor's consent. Based on this, we developed a highly interoperable federated IT infrastructure using standards such as Fast Healthcare Interoperability Resources (HL7 FHIR) and Clinical Quality Language (CQL).
Results: The infrastructure comprises two major components enabling federated real-time access to biosample metadata, allowing privacy-compliant queries and subsequent project requests. It has been in use since 2019, connecting 16 German academic biobanks, with additional European biobanks joining. In production since 2019 it has run 4941 queries over the span of one year on more than 900,000 biosamples collected from more than 170,000 donors.
Conclusion: This infrastructure enhances the visibility and accessibility of biosamples for research, addressing the growing demand for human biosamples and associated data in research. It also underscores the need for improvements in processes beyond IT infrastructure, aiming to advance biomedical research and similar infrastructure development in other fields.

* Corresponding author. German Biobank Node (GBN), Charité – Universitätsmedizin Berlin, Berlin, Germany.
E-mail address: caecilia.engels@charite.de (C. Engels).
[1] The acknowledgements section provides the members of the GBA IT development team by name.

https://doi.org/10.1016/j.compbiomed.2024.108941
Received 12 March 2024; Received in revised form 28 June 2024; Accepted 23 July 2024
Available online 5 August 2024

1. Introduction

Modern biomedical research relies heavily on large amounts of high-quality biosamples and associated clinical data. The central biobanks of the German Biobank Alliance (GBA) [1] collect, process and store such samples as part of a health research infrastructure [2]. Human samples stored in the GBA biobanks can be retrieved upon request based on the consent of the donor. These can serve for research from developing new treatments to biomarkers for personalised therapy identification [3]. The professional biobanks ensure long-term sample security, robust documentation, and linkage of biosample data with clinical information, resulting in high-quality, well-annotated samples [4]. Based on a recent stakeholder survey among researchers who use biosamples in their projects, over half of the 354 respondents were not aware of a central biobank at their own institution. The goal of our development was, to address the issues of low biobank visibility, inhomogeneous use and access rules, non-interoperable data sets, and the inaccessibility of valuable samples scattered across collections and locations These issues impede external researchers' visibility and accessibility.

GBN serves as the German hub for the European Research Infrastructure for Biobanking BBMRI-ERIC [5]. It facilitates access to a large number of biosamples of comparable quality for research purposes. This is important for rare diseases where individual institutions may not have enough cases to conduct a statistically sound study[6]. Under BBMRI-ERIC supervision, the biobanking community has co-developed and operates an Authentication and Authorisation Infrastructure (AAI), a catalogue listing 423 biobanks in the BBMRI-ERIC Directory [7], and a communication tool called BBMRI-ERIC Negotiator [8].

Despite significant advances in IT infrastructures to support biomedical research, the visibility and accessibility of biosamples across Germany remains a motivational driver and several research gaps needed to be addressed. The German Biobank Alliance (GBA) aimed to develop a system that would allow real-time discovery of biosamples at the donor level through a user-friendly graphical interface (GUI). This development followed the FAIR principles of data sharing [9]: discoverability, accessibility, interoperability and reusability. These principles were applied not only to the biosamples, but also to their descriptive data, ensuring comprehensive data sharing and integration. During the pilot phase of the German Biobank Node (GBN) [1], a thorough requirements analysis [10] was conducted with input from various stakeholders, leading to the identification of key needs [11]. These included providing value to researchers through an intuitive query interface, securely connecting biobank data to a central search service without transferring sensitive non-aggregated data, while maintaining data sovereignty so that biobanks control their data and access requests. In addition, technical interoperability was ensured by adhering to established standards to support different biobank systems, while syntactic and semantic interoperability was achieved through common data formats and semantic codes. Sustainability was a priority, with the infrastructure being open source and well documented to facilitate widespread access and co-development. Finally, the infrastructure was designed to support existing biobank systems and workflows without imposing significant additional burdens. These measures highlight both the progress that has been made and the continuing need to address the remaining challenges in optimising biobank sample and data sharing and accessibility.

2. Methods

2.1. Access control and data protection

Access control and data protection are crucial in safeguarding clinical data used in research. Patient consent is the primary condition for the utilization of clinical data, collected from sources like the biobank-information-management-system (BIMS) and clinical information systems. These datasets contain medical data (MDAT) and identifying data (IDAT). Pseudonymization, managed by the biobank, is used to protect patient identities [12]. Only MDAT is processed and stored in the data warehouse of the local component, known as Bridgehead.

Access to local data warehouses is restricted to authorized IT administrators, who may access it only when necessary and with a clear understanding of their confidentiality obligations. Biosample usage adheres to informed consent protocols and restrictive use and access processes managed by the biobank. All user activity, including access details, is logged and retained for one year. To maintain data integrity and security, direct communication between local components across different institutions is not facilitated.

Communication within the federated IT infrastructure occurs through encrypted HTTPS connections, complying with security requirements outlined in the GBA data protection concept [13]. Firewalls are deployed to safeguard central components hosted on organizational servers, permitting only essential protocols and designated ports for communication. Processing of personal data adheres to GDPR regulations, focusing on confidentiality, integrity, availability, and system resilience. Data protection impact assessments (DPIA) are conducted to identify potential risks and implement appropriate technical and organizational controls, ensuring confidentiality and security within the infrastructure.

2.2. Data harmonisation & FHIR integration

The initial step in establishing a shared infrastructure for biobanks is achieving mutual understanding and compatibility of biobank-related data among participants via semantic and syntactic interoperability. Partner biobanks of the GBA agreed upon a unified dataset comprising biosample and clinical donor data, aligned with national and international biobanking standards such as SPREC v2.0 [14], MIABIS Core 2.0 [15], the MIABIS sample and donor component [16,17], and the BBMRI-ERIC Directory [18]. The responsibility for transforming the data from the local sources into this harmonised dataset lies with the partner biobanks.

During this project, we adopted HL7®FHIR® (Health Level Seven Fast Healthcare Interoperability Resources) [19,20], a standard for transferring healthcare data between different medical systems. We selected the following FHIR resources: *Specimen, Observation, Patient, Condition,* and *Organisation*, ensuring alignment with the standard's fields and values [21] as closely as possible. Fig. 1 illustrates the selected resources in a Unified Modeling Language (UML) diagram. Two extensions were added to the FHIR standard: Firstly, one was added directly to the Specimen resource containing a diagnosis with the ICD-10 code. This was necessary to reflect the process of histology on the sample as the direct basis for a patient's diagnosis in pathology, in addition to the diagnosis (Condition resource) indirectly associated with the donor (Patient resource). Secondly, the storage temperature is now linked directly to the *Specimen resource* (instead of to the container).

Validate of the resulting dataset involved executing sample requests collected from participating biobanks (refer to Table 1 in the appendix) on generated test datasets using the FHIR® standard search mechanism (FHIR® Search). Despite our efforts to harmonize the data, hidden semantics posed challenges, particularly in complex queries: Finding patients with a specific diagnosis and sample, such as a blood sample, can be a complex task. It is necessary to connect a sample to a particular diagnosis to conduct research on a given disease. To execute such a query, the *Patient* resource is used as the context for the sample. This will display any patient with the matching diagnosis and sample type. Depending on the intent of the query, this could be a false positive. The researcher may have intended for the sample to be taken in the context of the diagnosis, rather than months or years before or after the diagnosis. This complexity led to the de-novo development of the Samply. Blaze FHIR store (described under the "local component") which enables the necessary querying of the data harmonised via the adapted FHIR profiles.

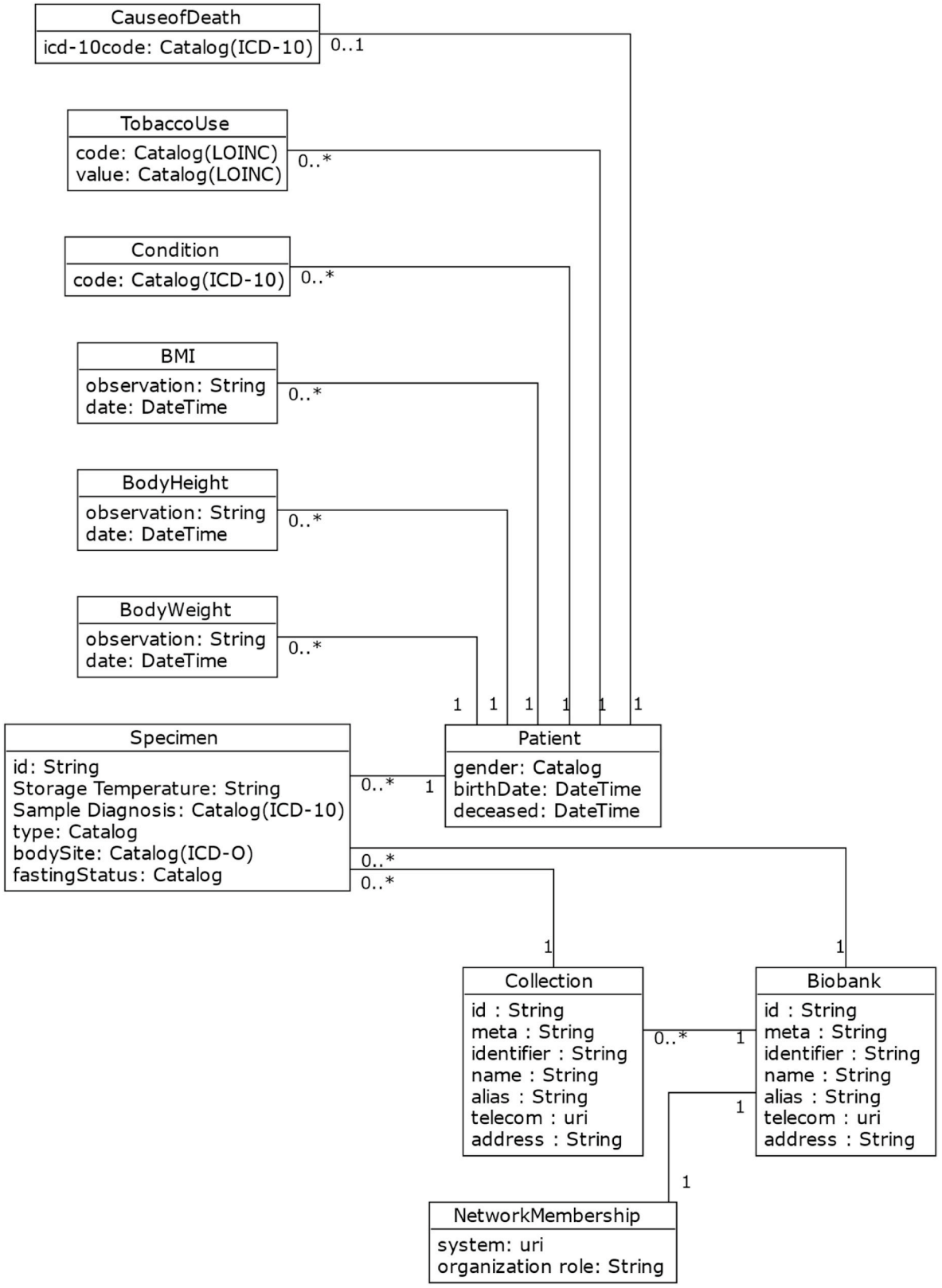

Fig. 1. Shows the UML diagram of the selected and adapted FHIR resources for the federated biosamples search network. These resources include Specimen (sample), Patient (donor), Organisation (Biobank, Sample Collection), Condition (Diagnosis), and Observations (Cause of Death, Tabacco Use, BMI, Body Height, Body Weight).

2.3. The federated search network

Local data processors known as Bridgeheads are implemented by participating institutions to connect to the federated search network. These Bridgeheads, located within the secure network of each biobank and its institution, serve as the hub for the federated query infrastructure. Search requests initiated from the central component are processed locally at each site, with aggregated results sent to the Sample Locator

for display. For visualisation of the data flow, please refer to the schema in Fig. 2a and b in the results section.

2.3.1. Local components

The Bridgehead is a turnkey solution, connecting sites to the central components of the infrastructure. Docker is used for deployment, ensuring ease of installation and updates. The site's non-sensitive configuration is stored in a Git repository [22], which only accessible to local administrators and known developers. Sensitive data is never publicly available and Docker Secrets are used for sensitive configuration. Network components are monitored using Icinga [23], which promptly identifies any issues. The Bridgehead contains a local data warehouse (Samply.Blaze HL7 FHIR Store) for data storage and a secure communication Connector.

The Samply.Blaze FHIR store is responsible for data persistence and querying according to the defined HL7 FHIR-based data model. We use HL7 Clinical Quality Language (CQL) [24] as a query language for FHIR data. At the time of developing the system architecture in the GBA, no software implementation using CQL was feasible. Therefore, we developed Samply.Blaze *de-novo*, combining the functionality for searches based on FHIR Search [25] and CQL, while supporting semantic and technical interoperability. Local IT administrators build specific Extract-Transform-Load (ETL)-pipelines at each site due to the heterogeneity of the data sources at the biobanks. The data is extracted from various local data sources (see Fig. 2a, A), transformed to the semantically and syntactically interoperable HL7 FHIR profiles, and loaded into the local Samply.Blaze store (see Fig. 2a, B). The Samply.Blaze FHIR store, developed for GBA, has since been refined further and successfully implemented in many Data Integration Centres (DIZ) of the Germany-wide Medical Informatics Initiative (MII).

The Connector of the Bridgehead facilitates secure communication with the central components through a RESTful API, ensuring outbound connections and employing differential privacy techniques to enhance data security. The Connector only accesses the pre-prepared inventory of the Samply.Blaze store using defined interfaces without holding any medical data except for caching purposes. To enhance data security, the data is obscured using differential privacy [26] with a privacy parameter ϵ for sensitive, medical data [27] (ϵ) of 0.28. This parameter results in a symmetric, exponential (i.e. Laplace) permutation with a mean of zero and a standard deviation of approximately 5. Results are also rounded to the nearest ten to mitigate potential privacy attacks.

2.3.2. Central components

The central components cater to researchers initiating sample searches via the Sample Locator, spanning multiple biobanks across Germany. The time required for a search depends on the size of the queried data set and the hardware at the sites, typically ranging from 5 to 20 s. Authentication, utilizing LifeScience AAI [28] (LS AAI) infrastructure, offers various identity providers such as the German DFN-AAI, Life Science Hostel registration with BBMRI-ERIC, and ORC ID for user authentication. After logging in and viewing the different sites that match their request, researchers can select which sites to contact. The Sample Locator's graphical user interface (GUI) has been evaluated for its intuitive usability [29]. The BBMRI-ERIC Negotiator [8] provides a communication platform for researchers and biobanks facilitating data exchange and collaboration.

3. Results

3.1. Federated IT infrastructure across biobanks

The federated IT infrastructure developed and implemented for

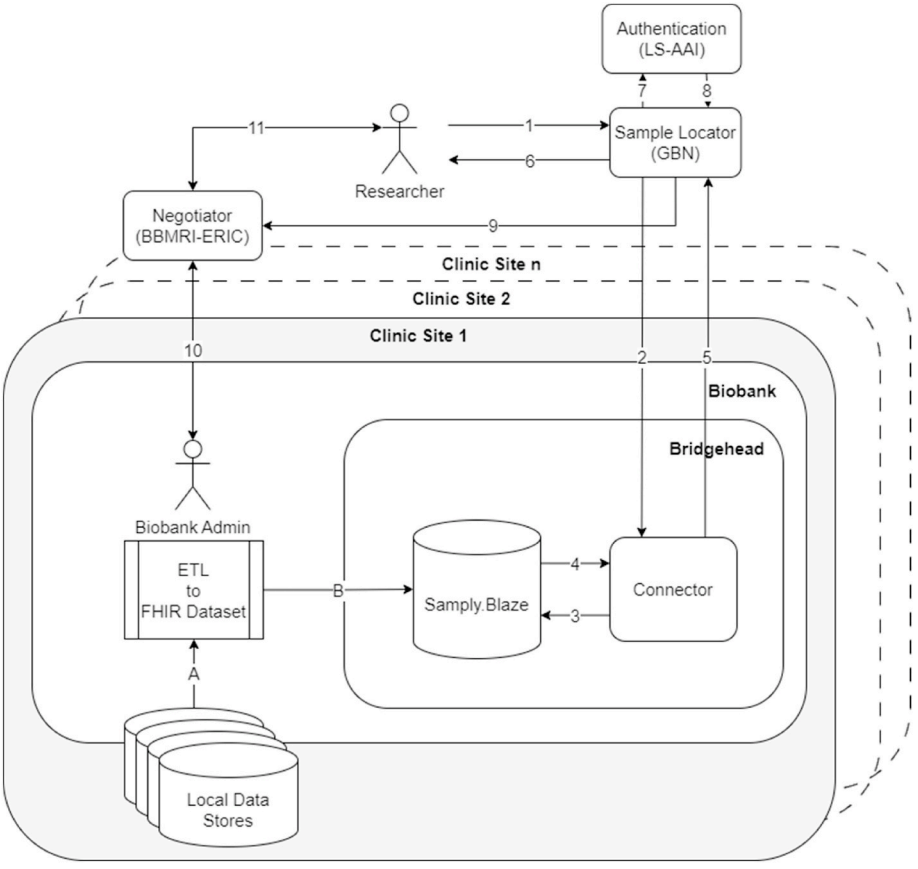

Fig. 2a. shows the data flow and interaction between the local component (Bridgehead) and central components (Sample Locator, LS-AAI and BBMRI-ERIC Negotiator).

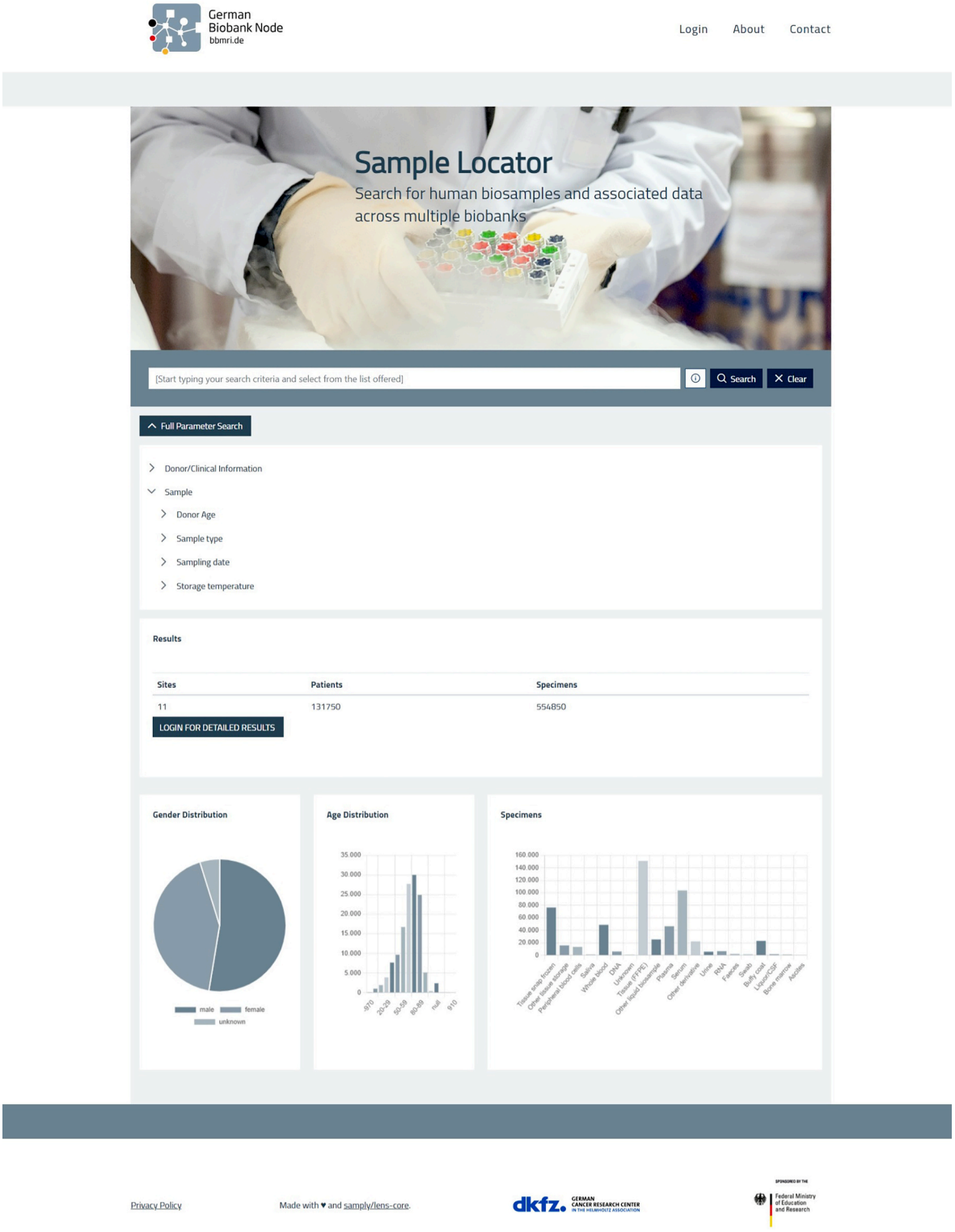

Fig. 2b. illustrates the Sample Locator Graphical User Interface (GUI), which enables researchers to select the parameters for their sample searches and obtain a stratified overview of the results. Once registration and login are complete, the number of samples at each site is displayed, along with the option to contact the relevant biobank directly [www.samplelocator.bbmri.de].

biobanks provides findability, accessibility, and interoperability for already collected biosamples and their associated data sets. This supports biomedical research and precision medicine while ensuring data security and regulatory compliance. Fig. 2a depicts the resulting data flow, with referring enumerations (A-B and 1–10) described as follows: When a researcher performs a search through the Sample Locator's graphical user interface, a query is automatically initiated (1). The Connector within each local Bridgehead (2) periodically retrieves queries relevant to its location from the central Sample Locator. The local Connector sends the polled query to the Samply.Blaze store that contains the data according to the FHIR profiles (3). The Samply.Blaze FHIR store then executes the query and returns the results to the

Connector (4). The Connector internally aggregates the data to maintain data security before returning patient and sample counts to the Sample Locator (5). The Sample Locator presents the reported counts to the researcher, aggregated across all sites or, if viewed by an authenticated researcher, aggregated for each individual site (6). On the user interface, the researcher is informed that the query results represent only potentially available biospecimens and data. For further information on access, the researcher can communicate directly with the respective biobanks after activating the checkboxes and clicking on 'Ask sites to negotiate'. To aid the researcher in comprehending the structure of the search results, additional histograms are displayed on the user interface. These histograms show the distribution of age, sex, and sample type of the patient, as well as presented sample counts. To identify which biobanks, present the requested samples, researchers are directed to authenticate via LS AAI (7, 8). To request the displayed samples, the Sample Locator forwards the query results, including parameters and detected collections within the biobank, to the BBMRI-ERIC Negotiator (9). The Negotiator allows the researcher to contact multiple biobanks simultaneously (10) to request the relevant samples and negotiate access to them (11). By establishing this communication, biobanks and researchers can clarify the eligibility and consent conditions related to the researcher's project. The BBMRI-ERIC Negotiator software, tool and process are described in detail by Holub et al. [8].

3.2. How biobanks can join the sample locator

Biobanks that want to take part in the GBA IT infrastructure and offer their samples through the Sample Locator can access detailed technical instructions from the BBMRI.de/GBA Implementation Guide [30]. The instructions cover the deployment and installation of the Bridgehead, which includes of the Samply.Blaze store and the Connector. For specific installation guidance, they can refer to the Bridgehead repository on GitHub [31]. Local biobanks must ensure that their server or virtual machine meets the minimum hardware requirements and supports Docker. Network communication requires outgoing HTTPS and proxy support, along with specific firewall configuration recommendations [32]. The local IT staff in the biobank or the institution then must then ensure data harmonisation via the ETL-process described above. Fig. 2a illustrates the process of *extracting* donor and sample data from primary sources, *transforming* it (A) according to the bbmri.de FHIR profile [33], and *loading* it (B) into the Samply.Blaze store within the local Bridgehead. The software's development, deployment, and operation are thoroughly documented [32], and the source code is published open source under the Apache-2.0 license [34]. In the following years, additional European biobanks from the BBMRI-ERIC will be joining the network, along with other German biobanks.

3.3. Utilization of the federated search network for biobanks

The Sample Locator was launched in October 2019 and has been in production since. Statistical monitoring of queries via the Sample Locator began in early May 2020. This report provides a statistical overview of the system's usage from May 2020 until the project's conclusion in April 2021. During this period, the central component sent 4941 queries to the Bridgeheads of the connected biobanks. The queries occasionally used a maximum of 6–8 criteria for a single query, although most queries only contained one criterion.

Diagnosis was the most frequently chosen attribute for a request, accounting for around 60 % of all requests, followed by gender, liquid sample, or tissue sample. Less frequently selected attributes included age at diagnosis, smoking status, and date of diagnosis. The least chosen attributes were storage temperature, donor BMI, body weight, fasting status, and date of collection. Table 1 shows the proportion of queries carried out for specific criteria such as diagnosis or sample type.

Of all the available attributes to select, the least requested were the sampling date, fasting status, and storage temperature. We included these in the dataset to allow the requester to assess the sample quality and suitability for their purposes. It might have been reasonable to assume that the ability for researchers to choose a qualitative attribute would be significant. However, without additional information, it is challenging to fully comprehend this outcome.

4. Discussion

Biobanks are facilities that store human biosamples and associated data for researches. In Germany, each university clinic has independently developed its own biobank. The German Biobank Node and Alliance (GBN/GBA) have developed and integrated a federated search infrastructure that enables real-time feasibility queries of available biosamples and data within the connected heterogeneous biobanks. Currently, the IT infrastructure consists of 16 secure local Bridgeheads situated within the academic biobanks of German university medical hospitals. These Bridgeheads are connected efficiently and securely connected to central components of the Sample Locator for federated searches, and the BBMRI-ERIC Negotiator used for follow-up communication to request sample access. The results at the sample level provide an overview over the interconnected biobanks. This system allows for efficient search and access to over 900,000 high-quality samples from more than 190,000 donors. The system's functionality is well-described and user-friendly, making it easily accessible to researchers.

We have established and executed an IT framework for federated data inquiries across German biobanks in alignment with the FAIR data sharing principles [9]. The data handling standard also adheres to the central goals of biobanking and is valuable to researchers. The utilization of high-quality biosamples retained in biobanks strengthens research integrity, reduces variance between results, and warrants traceability of the sample sources. The federated search Sample Locator provides an opportunity for extensive sample and data analysis and reanalysis. This infrastructure the *findability* and *accessibility* of biosamples and enables *interoperability* across multicentre studies. This leads to the *reuse* of information, promoting fairness, and driving high-quality research, thereby paving the way for improved personalised medicine.

The infrastructure enhances the visibility, discoverability, and accessibility of biosamples within academic biobanks while upholding the security and sovereignty of biobanks' data. Feasibility queries ensure compliance with data protection regulations, such as the European General Data Protection Regulation (GDPR), by transmitting only aggregated data from sites. This approach minimises the disclosed information relevant to the query objective and avoids transmitting sensitive patient data through locally-based, federated patient data processing, as required by Article 44 of GDPR. To maintain data privacy in the local components with access control mechanisms and privacy-enhancing technologies are in place, following the principles of differential privacy. The relevant sites negotiate all the terms before specific samples can be obtained.

Harmonising a dataset across all participating sites, based on stakeholder input incorporating national and international standards ensures

Table 1
Shows the percentage of time each criterion was selected for a search.

Parameters	Diagnosis	Donor gender	Liquid sample	Tissue sample	Donor age	Age at Diagnosis	Smoker	Date of Diagnosis	Storage temperature	BMI	Donor weight	Fasting status	Sampling date
%	60	15	13	8	5	3	3	2	2	2	2	1	0

technical, syntactic, and semantic interoperability. To achieve this, we developed a new FHIR store that facilitates querying of the harmonised and profiled FHIR dataset for complex and realistic sample queries. Our entire infrastructure's code is open source and documented on GitHub. Centralised maintenance of infrastructure operations improves sustainability and minimises the burden on the IT resources of each participating site by automatically deploying updates or changes by the central IT-Team in consultation with the sites.

4.1. Related work and research gaps

The research presented in this article has already influenced further initiatives and infrastructures beyond its own research domain. For instance, ABIDE_MI (Aligning Biobanking and Data Integration Centres Efficiently) integrates the developments in this study with the German Medical Informatics Initiative (MII). The aim was to synchronise IT infrastructures between biobanks and MII Data Integration Centres at German university hospitals [35], with the goal of improving data accessibility while minimising integration efforts. The datasets were profiled in HL7 FHIR from the onset, and the IT architecture relies on, among other software, the Samply.Blaze FHIR store [36], which is being developed further. The project intends to connect routine care data, from German university hospitals, as opposed to research data. The infrastructure is limited to registered users, with registration facilitated by the TMF – Technologie-und Methodenplattform für die vernetzte medizinische Forschung e.V.

The Clinical Communication Platform (CCP) [37] of the German Cancer Consortium (DKTK) is a research network that utilises federated IT infrastructures to facilitate biomedical research through data sharing and analysis. The DKTK CCP-IT [38] software paved the way for the development of federated search tools and established a technology-based system architecture comprising a Bridgehead linked to central components. The CCP currently comprises of 14 Comprehensive Cancer Centres (CCC) affiliated with German university hospitals. In comparison to the GBA use-case, the DKTK provides comprehensive information from a specific research domain, allowing for searches of cancer entities or specific clinical contexts in the realm of personalised medicine [6]. However, the infrastructure described here has a significant advantage over the CCP search, which is not available for public access but only for DKTK-affiliated researchers. The biosample FHIR profile and the FHIR store developed by GBA were later backported to the CCP. These contributions exemplify successful inter-project cooperation and open sourcing of research projects. The scalability and adaptability [39] of the system are demonstrated by these two networks.

4.2. Strengths, limitations and future work

One of the key design principles of this work is to avoid centralised storage of sensitive data. Instead, local and decentralised storage using Bridgeheads [40] within the secure network of the host institution (e.g., university hospital) is preferred. While this approach minimises the installation, deployment, and maintenance of the developed software, it also means that the relevant data are heterogeneously stored across each institution. Local IT administrators must consistently make efforts to integrate data. Currently, some German biobanks lack IT personnel, which hinders their ability to participate in the described infrastructure.

The system is being developed further to simplify local administration and maintenance, enhance search performance and increase flexibility in adding attributes. In collaboration with the German Cancer Consortium DKTK, and the German MII we are extending the FHIR data model to allow disease-specific searches for oncology queries. Through BBMRI-ERIC, the infrastructure is expanding to include additional academic biobanks in Germany and Europe. In the future, we will develop concepts to ensure the quality of the data transformed to fit the FHIR-profiles. We will also create a feedback tool for researchers who have requested samples via this federated network, which will connect metadata back to the residual samples in the biobank. Additionally, we have developed a new component, Samply.Beam [40], which improves communication, simplifies networking logic, facilitates maintenance, boosts performance, and enables end-to-end encrypted communication. This enhancement is being introduced to our infrastructure. Result and randomisation caching mitigate repeated query attacks and enhance the obfuscation implementation. The graphical user interface has been completely redesigned to better utilise CQL's stratification capabilities, improving the scalability of searchable criteria and enhancing the system's usability and responsiveness.

5. Conclusion

The Sample Locator, powered by HL7 FHIR, is a significant advancement in biosample and data discovery for healthcare and research in Europe.

Enhanced Collaboration: The tool promotes collaboration among research institutions and biobanks by simplifying the discovery and sharing of biosamples and related data.

Data Quality and Standardization: The use of HL7 FHIR and standardized terminologies ensures data quality and consistency, reducing errors and enhancing research reliability.

Compliance: The Sample Locator helps institutions adhere to data privacy regulations by implementing stringent access controls and ensuring GDPR compliance.

Streamlined Workflow: Researchers can efficiently locate and access the biosamples they need, reducing lengthy search processes or even preventing parallel sample collection and speeding up research projects.

By providing a federated search tool with standardized data and robust privacy controls, this tool facilitates collaboration and accelerates scientific advancements in the fields of genetics, genomics, personalised medicine, and beyond. As biobanks and research institutions continue to adopt this innovative tool, the potential for ground-breaking discoveries in the life sciences becomes even more promising.

Funding

The German Federal Ministry of Education (https://www.bmbf.de/) funded the German Biobank Alliance project (grant numbers 16EY1701-14).

Ethics approval and consent to participate

The manuscript does not report on or involve the use of any animal or human data or tissue. Thus, this section is not applicable.

Consent for publication

The manuscript does not contain data from any individual person. Thus, this section is not applicable.

Availability of data and materials

The datasets generated during the current study are available on simplifier.net.

- Table: Requests to biobanks - > listed in the appendix below
- FHIR-dataset generated: https://simplifier.net/bbmri.de
- Mapping tables: https://samply.github.io/bbmri-fhir-ig/mappings.html

All software generated during this project are included in this published article and additionally listed here as GitHub links.

- https://github.com/samply/bridgehead
- https://github.com/samply/beam
- https://github.com/samply/lens
- https://github.com/samply/spot
- https://github.com/samply/transfair
- Implementation Guide: https://samply.github.io/bbmri-fhir-ig/howtoJoin.html

CRediT authorship contribution statement

Cecilia Engels: Writing – original draft, Project administration, Conceptualization. **Jori Kern:** Writing – review & editing. **Zdenka Dudová:** Writing – review & editing. **Noemi Deppenwiese:** Methodology. **Alexander Kiel:** Software, Methodology. **Björn Kroll:** Software. **Tobias Kussel:** Writing – review & editing, Software. **Christina Schüttler:** Project administration. **Radovan Tomášik:** Software. **Michael Hummel:** Supervision. **Martin Lablans:** Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We would like to thank all active and former members of the IT development team and the local IT team members of the German Biobank Alliance and BBMRI-ERIC CS IT. We would also like to thank our anonymous reviewers and the editor for their constructive feedback on the manuscript.

List of abbreviations

AAI	Authentication and Authorisation Infrastructure
ABIDE_MI	Aligning Biobanking and DIC Efficiently
BBMRI-ERIC	Biobanking and Biomolecular Resources Research Infrastructure – European Research Infrastructure Consortium
BMBF	Bundesministerium für Bildung und Forschung (Federal Ministry of Education and Research)
CCP	Clinical Communication Platform
CQL:	Clinical Quality Language
DKTK	Deutsches Konsortium für Translationale Krebsforschung (German Cancer Consortium)
DWH	Data warehouse
ETL:	extract, transform, load
FAIR data	data meeting the principles of findability, accessibility, interoperability, and reusability
FHIR	Fast Healthcare Interoperability Resources
GBA	German Biobank Alliance
GBN	German Biobank Node
Git	Global information tracker
GUI	graphical user interface
HL7	Health Level 7
ICD	International Classification of Diseases
MIABIS	Minimum Information About Biobank data Sharing
MII	Medical Informatics Initiative
ORC ID	Open Researcher and Contributor ID
SPREC	Standard PREanalytical Code
UML:	Unified Modeling Language
VPN	Virtual private network

Appendix

Table 1

	Requests to biobanks
1.	Search for patients for a register of patients with fistula carcinoma in M. Crohn's disease. The idea of this registry is to collect clinical courses of these patients prospectively and retrospectively throughout Germany in order to be able to make statements about the risk for the occurrence of fistula carcinoma and to observe oncological courses. In parallel, the molecular characterisation of these carcinomas will be carried out using FFPE samples.
2.	Lung cancer: The aim is to comprehensively characterise typical and atypical PCa to reveal the molecular changes associated with the more aggressive phenotype of atypical PCa.
3.	Search for FFPE tumour samples from routine
4.	Search for tissue and blood samples for Crohn's disease and ulcerative colitis. Both tumour tissue and healthy tissue (terminal ileum and colon) are required
5.	We are looking for patients with AML of whom there is biomaterial before and after therapy with azacitidine.
6.	Number of patients with serum and tissue samples diagnosed with C50* who underwent surgery in 2018
7.	Request for samples on the subject of thyroid nodules (Thyroid nodule)
8.	Request for blood samples from patients with latent tuberculosis infection for validation of biomarkers
9.	Request for a technical performance evaluation requiring the following urine samples with concentration data for 6-acetymorphins (6AM), hydrocodones (HYD), EDDP and BUP (preferably not hydrolysed). The values should be determined with GCMS or LCMS, approx. 60 positive samples per parameter. It would be great if you could provide us with the chromatograms as well.
10.	For a project in XYZ blood samples (cellular components) from healthy samples: father, mother and child (trio). Of course, it must be the biological parents and the "child" can/should be adult.

(*continued on next page*)

Table 1 (*continued*)

	Requests to biobanks
11.	Evaluating miRNAs as circulating biomarkers for Glioma and Meningioma, Requesting blood or serum samples for both, 10 samples for each of glioma stage I, II, III and glioblastoma and 10 for each of meningioma stage I, II and III.
12.	We need 300 number of archived invasive breast cancer tissues (paraffin preserved) from patients diagnosed anytime between 2009 and 2014 that meet the following criteria: The paraffin blocks should have minimum 30 % tumour content for us to perform IHC staining.
13.	Ebstein Barr Virus (EBV): samples either urine or plasma samples with specific cut-off titres (determined with the CE-marked pcr kit) both positive and negative samples. BK Virus: samples either urine or plasma samples with specific cut-off titres (determined with the CE marked pcr kit) both positive and negative samples
14.	NMR metabolomics data + phenotype data
15.	GWAS data (chip or imputed) + phenotype data
16.	Any genetic data (eg. WES, WGS, SNP array) data in any format + phenotype data
17.	GWAS chip data in vcf-format, all sequencing data available (VCF-files and bam-files), signal intensity data for CNV calling (B-allele frequency and Log-R-Ratio for each probe) + phenotype data
18.	NMR metabolomics data and genome variation data (sequencing or SNP genotyping data preferable in imputed variant call format) + phenotype data
19.	GWAS data, WES data, WGS data, NMR metabolomics data, MS metabolomics data, expression data + phenotype data + stool and plasma samples
20.	GWAS data + phenotype data + DNA samples
21.	Recall study using following criteria for selecting participants: certain genotype in defined genetic locus/loci + phenotype data
22.	Inflammatory bowel disease including ulcerative colitis (K51), Crohn's disease (K50) and colorectal cancer (C18, C20). Exclusion criteria: Microscopic colitis Patients age range >18 yrs. FFPE tissue TMA slides/block, Fresh frozen tissue, Whole blood, Serum/Plasma, Buffy coat, Digitised/scanned slides Tissue sections should be collected on poly-l-lysine coated slides, thickness of the slides is 4 or 5 μm. Clinical data linked to the sample: Date of birth, gender, height, weight, medication, dates of colonoscopy, endoscopy findings, pathology findings, diagnoses

References

[1] German Biobank Node, Webseite des German Biobank Node (GBN). http://bbmri.de/. (Accessed 30 July 2015).

[2] T. Castillo-Pelayo, S. Babinszky, J. LeBlanc, P.H. Watson, The importance of biobanking in cancer research, Biopreserv. Biobanking 13 (3) (2015) 172–177, https://doi.org/10.1089/bio.2014.0061.

[3] H.A. Massett, N.L. Atkinson, D. Weber, et al., Assessing the need for a standardized cancer HUman Biobank (caHUB): findings from a national survey with cancer researchers, J. Natl. Cancer Inst. Monogr. 2011 (42) (2011) 8–15, https://doi.org/10.1093/jncimonographs/lgr007.

[4] C. Klingler, M von Jagwitz-Biegnitz, R. Baber, et al., Stakeholder engagement to ensure the sustainability of biobanks: a survey of potential users of biobank services, Eur. J. Hum. Genet. (2021), https://doi.org/10.1038/s41431-021-00905-x.

[5] BBMRI-ERIC. http://www.bbmri-eric.eu/. (Accessed 30 July 2015).

[6] R.O. Barnes, P.H. Watson, Precision medicine: driving the evolution of biobanking quality, Healthc. Manag. Forum 33 (3) (2020) 102–106, https://doi.org/10.1177/0840470419898874.

[7] K.J. van der Velde, F. Imhann, B. Charbon, et al., MOLGENIS research: advanced bioinformatics data software for non-bioinformaticians, Bioinformatics 35 (6) (2019) 1076–1078, https://doi.org/10.1093/bioinformatics/bty742.

[8] R. Reihs, R. Proynova, S. Maqsood, et al., BBMRI-ERIC negotiator: implementing efficient access to biobanks, Biopreserv. Biobanking (2021), https://doi.org/10.1089/bio.2020.0144.

[9] M.D. Wilkinson, M. Dumontier, I.J.J. Aalbersberg, et al., The FAIR Guiding Principles for scientific data management and stewardship, Sci. Data 3 (2016) 160018, https://doi.org/10.1038/sdata.2016.18.

[10] C. Schüttler, N. Buschhüter, C. Döllinger, et al., Requirements for a cross-location biobank IT infrastructure: survey of stakeholder input on the establishment of a biobank network of the German Biobank Alliance (GBA), Pathologe 39 (4) (2018) 289–296, https://doi.org/10.1007/s00292-018-0435-9.

[11] C. Schüttler, H.-U. Prokosch, M. Hummel, M. Lablans, B. Kroll, C. Engels, The journey to establishing an IT-infrastructure within the German Biobank Alliance, PLoS One 16 (9) (2021) e0257632, https://doi.org/10.1371/journal.pone.0257632.

[12] T.H. Müller, R. Thasler, Separation of personal data in a biobank information system, Stud. Health Technol. Inf. 205 (2014) 388–392.

[13] Lablans M, Schmidt EE, Duhm-Harbeck P, Prokosch H-U, Hummel M. German Biobank Alliance (GBA) Data Protection Concept. Accessed July 28, 2023.

[14] S. Lehmann, F. Guadagni, H. Moore, et al., Standard preanalytical coding for biospecimens: review and implementation of the Sample PREanalytical Code (SPREC), Biopreserv. Biobanking 10 (4) (2012) 366–374, https://doi.org/10.1089/bio.2012.0012.

[15] N. Eklund, N.H. Andrianarisoa, E. van Enckevort, et al., Extending the minimum information about BIobank data sharing terminology to describe samples, sample donors, and events, Biopreserv. Biobanking 18 (3) (2020) 155–164, https://doi.org/10.1089/bio.2019.0129.

[16] R. Merino-Martinez, L. Norlin, D. van Enckevort, et al., Toward global biobank integration by implementation of the minimum information about BIobank data sharing (MIABIS 2.0 Core), Biopreserv. Biobanking 14 (4) (2016) 298–306, https://doi.org/10.1089/bio.2015.0070.

[17] D. Bender, K. Sartipi, HL7 FHIR: an Agile and RESTful approach to healthcare information exchange, in: P.P. Rodrigues (Ed.), IEEE 26th International Symposium on Computer-Based Medical Systems (CBMS), 2013: 20 - 22 June 2013, IEEE, University of Porto, Portugal. Piscataway, NJ, 2013, pp. 326–331.

[18] P. Holub, M. Swertz, R. Reihs, D. van Enckevort, H. Müller, J.-E. Litton, BBMRI-ERIC directory: 515 biobanks with over 60 million biological samples, Biopreserv. Biobanking 14 (6) (2016) 559–562, https://doi.org/10.1089/bio.2016.0088.

[19] Http://hl7.org/fhir. Index - FHIR v5.0.0. https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf. Updated March 31, 2023.000Z. Accessed July 27, 2023.163Z.

[20] H. Leroux, A. Metke-Jimenez, M.J. Lawley, Towards achieving semantic interoperability of clinical study data with FHIR, J. Biomed. Semant. 8 (1) (2017) 41, https://doi.org/10.1186/s13326-017-0148-7.

[21] BBMRI.de. FHIR Ressource: Specimen. Accessed June 10, 2022.

[22] Samply Open Source Community. Samply/Bridgehead-config. https://github.com/samply/bridgehead-config Accessed October 19, 2023.

[23] Icinga GmbH. Monitor Your Entire Infrastructure with Icinga. https://icinga.com. Updated October 19, 2023.

[24] HL7 International. HL7 Cross-Paradigm Specification: Clinical Quality Language (CQL), Release 1. http://www.hl7.org/implement/standards/product_brief.cfm?product_id=400 Accessed December 20, 2022.

[25] C. Gulden, S. Mate, H.-U. Prokosch, S. Kraus, Investigating the capabilities of FHIR search for clinical trial phenotyping, Stud. Health Technol. Inf. 253 (2018) 3–7.

[26] C. Dwork, F. McSherry, K. Nissim, A. Smith, Calibrating noise to sensitivity in private data analysis, in: D. Hutchison, T. Kanade, J. Kittler, et al. (Eds.), Theory of Cryptography, vol. 3876, Springer Berlin Heidelberg, Berlin, Heidelberg, 2006, pp. 265–284.

[27] J. Hsu, M. Gaboardi, A. Haeberlen, et al., Differential Privacy: an Economic Method for Choosing Epsilon, 2014, https://doi.org/10.48550/arXiv.1402.3329.

[28] BBMRI-ERIC, BBMRI-ERIC AAI privacy policy: version 1.1, 2017-08-23. https://web.bbmri-eric.eu/Policies/BBMRI-ERIC-AAI-Privacy-Policy.pdf. (Accessed 6 February 2023).

[29] C. Schüttler, V. Huth, M von Jagwitz-Biegnitz, M. Lablans, H.-U. Prokosch, L. Griebel, A federated online search tool for biospecimens (sample locator): usability study, J. Med. Internet Res. 22 (8) (2020) e17739, https://doi.org/10.2196/17739.

[30] Samply Open Source Community, BBMRI.de/GBA Implementation Guide. https://samply.github.io/bbmri-fhir-ig/. (Accessed 19 October 2023).

[31] Samply Open Source Community, Bridgehead deployment. https://github.com/samply/bridgehead-deployment. (Accessed 19 October 2023).

[32] Samply Open Source Community, BBMRI.de/GBA Implementation Guide. https://samply.github.io/bbmri-fhir-ig. (Accessed 19 October 2023).

[33] BBMRI-ERIC. Profiles for the GBA/BBMRI.de project. https://simplifier.net/bbmri.de/de.bbmri.fhir/.

[34] Samply Open Source Community, Apache licence. https://github.com/samply/share-client/blob/master/LICENSE. (Accessed 13 October 2022).

[35] H.-U. Prokosch, R. Baber, P. Bollmann, M. Gebhardt, J. Gruendner, M. Hummel, Aligning biobanks and data integration centers efficiently (ABIDE_MI), Stud. Health Technol. Inf. 292 (2022) 37–42, https://doi.org/10.3233/SHTI220317.

[36] J. Gruendner, N. Deppenwiese, M. Folz, et al., The architecture of a feasibility query portal for distributed COVID-19 Fast healthcare interoperability resources (FHIR) patient data repositories: design and implementation study, JMIR Med Inform 10 (5) (2022) e36709, https://doi.org/10.2196/36709.

[37] M. Lablans, E.E. Schmidt, F. Ückert, An architecture for translational cancer research as exemplified by the German cancer Consortium, JCO Clin Cancer Inform (1) (2017) 1–8, https://doi.org/10.1200/CCI.17.00062.

[38] M Lablans. CCP-IT. https://dktk.dkfz.de/de/klinische-plattformen/arbeitsgruppen-partner/ccp-it.
[39] D. Maier, J.J. Vehreschild, B. Uhl, et al., Profile of the multicenter cohort of the German cancer consortium's clinical communication platform, Eur. J. Epidemiol. 38 (5) (2023) 573–586, https://doi.org/10.1007/s10654-023-00990-w.
[40] Samply Open Source Community, Bridgehead deployment. https://github.com/samply/bridgehead-deployment. (Accessed 14 July 2021).