Architecture Enabling Service-oriented Digital Biobanks

Jarkko Hyysalo
University of Oulu
jarkko.hyysalo@oulu.fi

Anja Keskinarkaus
University of Oulu
anjakes@ee.oulu.fi

Gavin Harper
University of Oulu
gavin.harper@student.oulu.fi

Jaakko Sauvola
University of Oulu
jaakko.sauvola@oulu.fi

Abstract

In Finland, the Biobank Act entered into effect in 2013. The primary motivation for the act is to enable the utilization of collected biological sample material for medical research. However, in order to effectively utilize this data, there exists a need to develop new technological solutions to support the collection and management of potentially large sets of sensitive data through multiple stages of processing. The cumulative data stored within biobanks will enable multi-disciplinary research and new innovations. We propose an architecture that addresses several challenges involved in defining and deploying a biobank infrastructure including consent management, data management and data transfer. Our architecture expedites the development of this important area within the research and industrial communities, and enables the deployment of service-oriented biobanks.

1. Introduction

Biobanks are becoming an increasingly common method for storing large collections of biological samples and relevant medical data collected from donors who have provided consent. There exist many benefits in creating a locally centralized repository for including but not limited to allowing researchers access vast quantities of data that would be otherwise difficult to obtain from many different sources.

Globally, many hundreds of biobanks exist. In the European Union, there are currently 315 active biobanks. Large biobanks currently in operation include the UK biobank, German National biobank, The Estonian Genome Center/The Estonian Biobank cohort and Danish National biobank. Currently, most scientific publications come from North America, Great Britain and Central and Southern Europe. Europe has a common infrastructure BBMRI-ERIC (Biobanking and BioMolecular resources Research Infrastructure-The European Research Infrastructure Consortium) for co-operation. [16, 21]

The guiding principles of the Biobank Act in Finland (established 1.9.2013) are promotion of trust, equal access to data and samples, protection of privacy, acceleration of innovation activities and exposing biobank activities to public scrutiny. The primary motivation for this act coming to pass is to develop a strong biobank research infrastructure, by overcoming hurdles like complex collaboration networks, sensitive data (hard identifiers, sensitive attributes), anonymized data conflicting with the need to identify subjects and additionally to identify and satisfy the needs of multiple stakeholders [19].

In 2015 nine biobanks have been granted the right to operate by Valvira (National Supervisory Authority for Welfare and Health) in Finland: Auria biobank, THL biobank, Finnish Hematological Registry and Biobank, Helsinki Urological Biobank, Academic Medical Center Helsinki Biobank, Northern Finland Biobank Borealis, Finnish Clinical Biobank Tampere, Central Finland biobank and Biobank of Eastern Finland. The profiles of the biobanks differ from each other to avoid redundant effort. There are currently no commercial biobanks in Finland, and the existing infrastructures primarily offer sample management and consultation in marketing and sales [16].

Some research goals would not be feasible without large biobank collections and the collaboration between biobanks. Moreover biobanking will require a consistent, harmonious collection of samples and data. Without harmonization, the advantage of the collaboration and data integration to enlarge dataset from several biobanks is highly non-trivial [10]. This puts pressure also for updating the representation of data in already existing collections. This is not a simple task as biobanks can store data of many distinct types and employ different data models and formats. Yet, the various information systems must be integrated with other information systems to allow both internal and external interoperability while maintaining and efficient and secure environment for the data to reside.

The purpose of this paper is to define an architecture for biobanks that corresponds to the requirements of a national architecture for biobanks and address the requirements for harmonization, national needs and regulations. Hence our research question: *How to define an architecture for a biobank?*. This work is based on extensive

URI: http://hdl.handle.net/10125/41578
ISBN: 978-0-9981331-0-2

collaboration between national actors in biobank domain; Finnish biobanks, software and systems providers, standardization organizations, hospitals and several other actors in the health care domain.

The remainder of this paper is organized as follows. Section 2 discusses the background for this work. Section 3 presents our research process. Section 4 outlines the proposed biobank architecture and Section 5 discusses the results. Section 6 concludes this work summarizing the key findings.

2. Background

The biobank shall store collected samples (DNA, cell tissue etc.) and associated clinical data about the donors [13] in a locally centralized repository. The information contained within the biobank is to be made available for the purposes of medical research. Management of the data is initiated by the granting of consent by a donor [4]. Currently the consent or revocation of consent is provided in written form and stored [7]. Samples and relevant data are stored within the biobank in an encoded form. The platform should handle the management of the data contained within, so that it can be easily queried and combining results with auxiliary information is possible. Such data may include but is not limited to digitized sample material in the form of whole slide images that may each occupy multiple gigabytes on disc [8]. When a researcher or commercial organization expresses interest in the biobank data (through e.g. a national catalog), the data is retrieved and passed through a secondary round of encoding prior to being released. Auxiliary material is collected from other sources, if required. The biobank should keep various registers including coding registry, registry of removed codes, registers of samples, registers of the utilization of data, registers for releases and returns of data. The biobank information management system should also handle the results of operations upon this data by researchers, which may of an arbitrary type.

2.1. National consensus

The operating model of the biobanks requires several stakeholders to give consent, to supervise, to utilize services, to complement data, to provide finance and support, to provide services and to provide common methods. The biobank itself will report these activities, provide sample and other data, provide services, and give expertise to BBMRI work to find and establish common national methods. International work goes through BBMRI in BBMRI-ERIC. National Institute for Health and Welfare in Finland (THL) is a national coordinator of BBMRI work, but also acts in many other roles. [22]

The Ministry of Health and Social Affairs (STM) has initiated a national effort on describing the architecture for biobanks in Finland. The purpose is to describe the operational environment as well as to define requirements for efficient operation. The short-term goal is to enable uniform operating guidelines for biobanks so that there may exist a national-level common infrastructure. This is an important window of opportunity for Finnish research and for international research networking. [22]

The framework suggests that certain services be handled in a common way. The management of consents that are granted or revoked by donors should reside in a single service. The transfers and material transfer agreements should also be similarly handled. Furthermore, it is important that the researcher can see the available data through an availability catalog database and in the second level through availability search service. In this regard, the complexity of the biobank should be abstracted away behind a single user-facing interface. Additionally, a coding service (method to separate actual identity from the stored data) is planned to be common. Biobanks shall thus integrate these national services into their own operational environment. [22]

In the national framework, certain parties have been suggested to assume control over these services. The consent register and event register could be part of Kanta (National Archive of Health Information) and OmaKanta (portal to personal medical records and electronic prescriptions). Responsibility of the coding service is not yet clear. It is important that the coding service is consistent across several biobanks when data is combined. The possibility to search all available data is a necessity to establish a national biobank operation and could arise from work in BBMRI. Consequently there should exist a common or at minimum interoperable information system. This is preceded by a definition of a reliable method to identify researchers and to manage access rights to the services. [22, 15]

The biobank is to store data indefinitely or until consent for the data is revoked. The structure of the basic data and logic for linking should therefore be simple. Moreover, the interfaces for data collection and release should be well defined and easy to use.

2.2. Core components of the biobank IT infrastructure

Functional components of the biobank infrastructure may be divided into four core components (cf. [1]): 1) Consent tracking; 2) Sample management; 3) Laboratory processing; 4) Facilitating

access to data and samples. Forthcoming European General Data Protection Regulation (GDPR) requires that informed, explicit consent is necessary [4].

When the samples and related metadata are processed for inclusion into the biobank, coding, anonymization and pseudonymization are required. Sample management includes sample processing, storage and inventory management. High-speed, large storage solutions are required (up to 100 TB per year) [17, 11]. An integral part of the biobank IT infrastructure is the database for samples and related metadata. Laboratory processing includes the analysis and processing of biological samples. Powerful processing capabilities are required for potential concurrent analysis of large image files that may be up to several gigabytes in size. Finally, the distribution of data and samples to researchers has to be managed. This requires sufficient information capabilities [15] including availability services like catalogues to deliver the data to users. This occurs either through online viewing or other data distribution methods. Further components are e.g. firewalls and other security measures that increase the security, integrity and redundancy of data. The definition of public and private APIs is also necessary to enable applications and services. Standardization and harmonization are also necessary, as there are differences in operating procedures between biobanks [2]. This can be due to local policies or technologies, and transparency is needed to make the data re-usable [2].

3. Research process

The research consisted of studying several organizations in related fields with the aim of defining the architecture. Several workshops with various organizations were arranged. Approaching our research question in an iterative manner was viewed as optimal as there were multiple actors and different domains involved. The concept was still evolving such that there were no clear views on how a biobank could be defined or constructed. Additionally, the requirements and environment experienced a constant flux as the laws and regulations were being crafted and refined.

Table 1 presents the organizations and different sources that participated the definition of the biobank architecture. The organizations were chosen such that they all could provide relevant data related to the research question. Experts and managers from different organizational levels were involved. Examples of input defined by the sources are presented in Table 1.

Table 1. Biobank actors.

Source	Input
Valvira	-Biobank permission -Supervision
Sample donor, person	-Consent -Samples
KELA (The Social Insurance Institution of Finland)	-Information systems service
THL	-Architecture -BBMRI-ERIC: Obligations
BBMRI	-Common methods
Registry	-Source data
Service provider	-Service
Health care units	-Sample and data -Support services
Research	-Sample and data
National ethical board	-Reports
STM	-National biobank overall architecture

Building the outcome iteratively allowed us to see the results, gather feedback and further improve our design. This was continued until we were able to draw the architecture presented in this paper. The intention was to define an architecture that is not dependent on specific vendors or components and may be modified and scaled to meet the future demands. The research process (Fig. 1) followed the guidelines presented by Runeson and Höst [18], in an iterative manner.

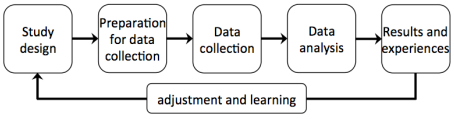

Figure 1. Research process [18].

In the design phase, the objectives and research problem in addition to the theoretical background were determined. Following this, the research methods and sources were chosen. The preparation phase included defining the topics to discuss with data sources, defining the data sources, and agreeing on the procedures. During the data collection phase, data were gathered from data sources via numerous workshops and discussion (first-degree data). This included collecting archive material such as process descriptions and workflows provided by the data sources (second-degree data) and the literature (third-degree data). Memorandums were made from workshops and discussions and summarized for all researchers. The main discussion points, relevant questions and interesting themes were then extracted from the data and analyzed further. Data analysis provided further topics for the following workshops and discussions. An understanding was gradually built as the process continued. Then, based on our understanding, the initial architecture was sketched. This was then discussed between representatives of participating organizations. Several discussions and workshops were

arranged at the companies for the participating organizations to go through the findings and results. Feedback from the workshops was incorporated in the analysis. The results of this process are reported here.

4. Proposed biobank architecture

The complete topology has been separated into multiple independent domains, grouping related services. It is possible to assign ownership responsibilities to each domain and modify the physical or logical mechanics to meet real-world requirements while ensuring topological coherence. This is achieved by strictly defining the input and output interfaces described for each domain. This architecture presents the topology of the biobank infrastructure and does not specify specific implementation details. The overall architecture is designed such that each component exists in an isolated environment communicating with other components through encrypted sockets. The overall architecture with an example on service scenario for both donor and recipient is shown in the Fig. 2.

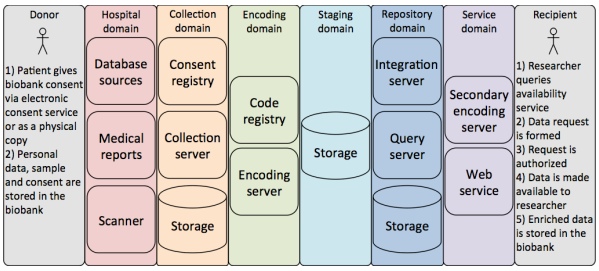

Figure 2. Overall architecture for the biobank.

Isolation is achieved at the component level through dedicated virtual machines and at the domain level by physical hardware. This approach allows for simplified scalability and allows new services to be included in the system. Additionally, by isolating components, the impact of security breaches may be localized and better contained. Similarly, isolating components allows for stricter definitions of valid traffic and thus potentially malicious or erroneous traffic can be better identified. Furthermore, this approach enables a heterogeneous collection of components from multiple vendors to coexist, communicating using open standard protocols and data formats. This aids in ensuring the future relevance of the biobank in constantly evolving medical and technological environments.

The hospital domain (Fig. 3) represents a collection of systems and devices upon which the data from consenting donors is stored. New data sources may be added within this domain as permission is granted or access is deemed feasible. Access to each data source is granted independently and that access is governed through a formal request procedure.

Additionally, the success of the biobank infrastructure may encourage the inclusion of data sets that are of a format that have not yet been considered, such new data inserted into the biobank from the research community. As a consequence, the biobank infrastructure has been designed to be agnostic of data formats and instead seeks to allow data from multiple sources to co-exist through a harmonization process.

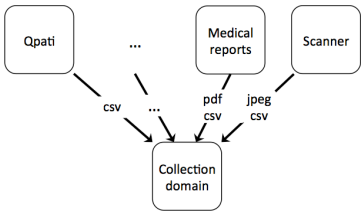

Figure 3. The hospital domain.

The collection domain (Fig. 4) represents a series of systems that are utilized in the amalgamation of the various data sources in the hospital domain in addition to consent information. The data contained within this domain is a complete representation of the original data. An additional benefit of maintaining this collection of data is that currently the requirements for encoding and obfuscating the datasets in addition to what may be included is not currently standardized and may change over time. By having an exact representation of the original data it is possible to ensure the continued compliance to data safety regulations over the lifetime of the biobank.

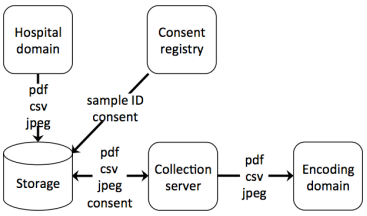

Figure 4. The collection domain.

This is designed as such so that in the event of a catastrophic failure, the biobank may utilize the data stored within this section to re-build the data set, bypassing production hospital systems. This is required as access to the hospital domain systems is limited. Prior to forwarding a data set to the encoding server for pseudonymization, a transfer delta is computed of the data contained within the collection server. This transfer delta is obtained by combining the unique, relevant fields from each table in the database that have

been created or modified after the last batch transfer and for which consent exits.

The collection domain requires the use of a database running on a dedicated hardware instance that is directly connected only to the encoding domain and will expose an encrypted network attached storage device for each hospital domain system with unique keys that is then pulled into the collection server. This database shall mirror the table structure of each database that exists as a valid source for the biobank. Data will undergo a harmonization process to ensure that incoming data is in a format suitable for inclusion into the biobank while maintaining the original structure. By mirroring the table structure of the source data, it is possible to ensure that data is not discarded unnecessarily and all transformations upon the source data may be computed non-destructively.

Consent data is transmitted to the consent register, which is stored within the collection domain. After consent has been registered, a sample ID created from the original data is stored within the collection domain with a Boolean value indicating the status of consent.

The encoding domain (Fig. 5) is responsible for ensuring that any personally identifying information is sanitized from the original data set prior to inclusion in the biobank primary repository.

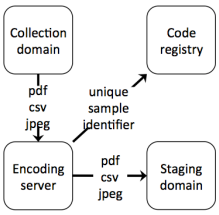

Figure 5. The encoding domain.

The encoding domain is additionally responsible for building up a code registry, which keeps a record of personally identifying information and primary codes. For anonymizing data a variety of methods are available, like encryption, hashing, substitution or removal of personally identifiable information. We approach this as follows. Data entering the encoding server from the collection domain is passed through a sanitization process generating a unique sample identifier through a one-way mapping process. Additionally, fields within the data set may be altered to prevent exposing identifying information. Upon completion of the sanitization process, a unique identifier for each sample processed is transmitted to the code registry over a secure connection. This identifier enables authorized personnel to map between the original samples and the pseudonymized biobank samples.

The staging domain (Fig. 6) facilitates the physical separation of the systems containing sensitive, personally identifying information and the sanitized data within the biobank primary repository. This domain comprises only a network attached storage device that may be mounted by only a single device at any point in time.

Since it is required that the regions containing sensitive information be entirely separated from the regions that can be considered general-access, it is necessary for there to exist an additional separating region. In this staging region, the encoding service briefly mounts a remote storage drive and deposits the encoded data prior to disconnecting. The repository domain then mounts the drive to fetch the deposited data. Data deposited in the staging area by the encoding domain shall be encrypted with the public key of the repository domain and shall have execution permissions disabled on the data. Upon successful mounting, the device shall lock access to the drive until such time that the connected client disconnects. It is required that only one device be connected to the staging domain at a given time to ensure that there does not exist a possibility for malicious traffic to flow between the regions.

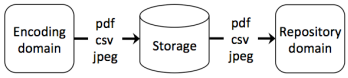

Figure 6. The staging domain.

The repository domain (Fig. 7) serves as the primary repository for the data that will be utilized for research purposes. The primary components of the biobank are a query server upon which resides a relational database allowing for queries of arbitrary complexity. Additionally, there exists a storage allocation containing non-textual data sets that may incur a significant storage, bandwidth or computational cost. Within this domain, there exists a requirement for the primary database and storage server to have a high-availability and be reasonably fault-tolerant. In Fig. 7 the structure of the repository domain is described, showing the path of data from the staging domain into the primary biobank database. The access mechanisms through which the service domain may communicate with the biobank primary repository are also detailed.

For data received from the staging domain, it is necessary to verify the source of the data and determine that the data is well formed. To do so, an integration server is utilized to read the retrieved data and prepare it for inclusion into the primary biobank database. To effectively serve arbitrary queries into the database, it is necessary that within the service domain, a structured request is constructed using prepared statements and have any potentially malicious

statements nullified prior to entering the repository domain. Upon receiving a well-formed request statement, the query server shall collect the requested data with textual data represented as a JSON or similarly trivially readable file for effective parsing by a web client in the service domain. Additionally, other non-textual data shall be packaged and transmitted over a secure connection to complete the request.

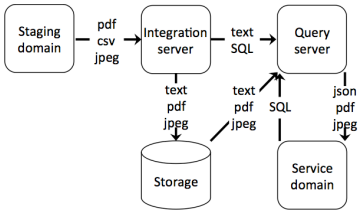

Figure 7. The repository domain.

The service domain (Fig. 8) comprises the set of systems that serve as a boundary layer between the biobank systems and the formation of requests for biobank data. In the service domain, an extensible set of services can be developed facilitating multiple use-cases for accessing the biobank allowing for scalable research efforts. In the basic case, there shall exist an interface into the biobank served over a secure HTTPS connection that shall facilitate requesting science data from within the biobank to the recipient. Prior to releasing data to external systems, a second round of data encoding is performed to fully sanitize any data that is made available from within the biobank.

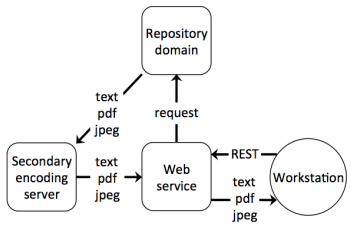

Figure 8. The service domain.

The use of an interface served over HTTPS, utilizing modern and open standards allows for there to not exist a dependency on any one particular type of environment from which the request is formed. This platform-agnostic approach ensures that the biobank is accessible from many types of research environments. Additionally, there is a possibility for a rapid deployment of security updates in the event that vulnerability is discovered. The safety of the data contained within the primary repository and preceding domains is paramount and thus there exists a requirement that modifications can be rapidly deployed to all data request origin points.

5. Discussion

Identified challenges can be divided into three categories: 1) Consent management and anonymization; 2) Sample management and harmonization; 3) Distribution of data and samples. In addition, there are technological challenges that exist in each category. Large-scale infrastructures have an inherent complexity in their design, construction and management [21]. Consequently, implementation costs for large-scale digital pathology environment may be high [11]. Moreover, technological limitations and insufficient performance of commercial solutions are identified [11]. With our architecture we aim to address all these challenges.

Data transfer requirements are considerable in digital biobanks. For example a typical tissue sample of 12x20 mm in size with 20x magnification can result an image many gigabytes in size. Thus, the transfer of images will require sufficient bandwidth. It is however possible to compress the images without introducing artifacts that measurably impact the quality of the image. The impact of compression must be validated prior to conducting an analysis on the data, as different compression methods may have different effects on image quality. [5]

Furthermore, the number of large files induces a storage burden. Rojo [17] estimated that about 100 TB of storage per year is needed to store the data generated in a pathology department in a 500 beds hospital. Hence, the storage investment might be substantial. In any case, there must be carefully planned storage and backup systems. Tiered storage systems are suggested to mitigate the high costs of storage [11].

Interoperability and integration issues are likely to arise. There are several components and services in a biobank environment that must be interoperable and there are various standards and specifications for enabling an interface [12]. Thus, data harmonization must be addressed carefully.

5.1. Consent management and anonymization

Given the sensitive nature of the data contained within a biobank, it is paramount that biobanks protect the confidentiality of the data according to national and international laws and regulations. Currently there is no real consensus regarding managing consents in biobanks [3]. However, the forthcoming GDPR covers e.g. consents and data protection.

The permission of the biobank to handle samples and related information is based upon the existence of consent. The Biobank Act requires written consent, with a signature. There is a right for the sample donor to deny and revoke the consent to the use of samples at

any time. Despite the national work integrating the consent management as part of the Kanta system, the biobank itself has to maintain a register of the consents. The Biobank Act defines that the donor has a right to know about the information stored in the biobank and any usage of said data (log files). Additionally, donors may elect to get information about the research done with their data. The initial requirements considering biobank consent management with OmaKanta have been defined in Kääriä et al. [9]. One technical issue is the integration of the local system to the forthcoming implementation of the national system as well as the integration to basic processes of the particular processing environment.

Protecting the privacy of genetic information requires that the data or samples are coded or anonymized. One issue influencing the information systems is the coding used to pseudonymize the social security numbers. According to the law, the sample and related information has to be coded. The information must be kept separate from the coding key and the information system must make possible the secure storage, usage and follow up of the information. The coding service is defined to include services related to the separate storage of the identifying information and samples and the management of the related coded information. One portion of this is code registry, which enables associating sample and information registry data to the information in the consent registry. The coding service is utilized for generating secondary codes when releasing data. If systems do not agree on a compatible coding space, and hinder or make it more difficult to collect data from multiple biobanks for research there shall exist a greater complexity in attaining a coherent and unified national biobank infrastructure [22].

5.1.1. Our solution to consent management and anonymization. Input from multiple sources is amalgamated and sanitized prior to inclusion in the biobank database. Consent for a data set is verified, the data is encoded to remove personally identifying information and any file format specific data is disassembled into raw meta data and image data prior to being sent into the biobank database. The collection domain receives data from sources including, but not limited to, hospital databases providing data in a trivially readable format by means of either a direct connection or by manual file transfer, medical scanners producing large image files, existing image data retrieved from PACS servers within the hospital or data that is manually entered. The sanitization process for images shall require significant memory if multiple requests are to be handled simultaneously. However, since the encoding domain functions as a filter in the data path, it is not required to store additional copies of the data passed to it. This data shall exist in its original form in the collection domain and in its encoded form in the repository domain.

5.2. Sample management and harmonization

Qpati is one example of a hospital information management system. It specially designed for the utilization in pathology departments. The system stores the sample number, sample types and other relevant information concerning the samples and has several device interfaces. For example, it has the capability to produce barcodes or QR codes for physical samples. The Qpati system enables analysis, management of information, storing diagnoses and has in addition connections to other patient information systems [14]. The Qpati system has been in use in several hospitals in Finland and updates on the system have gradually been added, consequently different hospitals may have a different version in use. Different hospitals may also use different clinical terms [15].

Harmonization is one of the critical enablers for achieving a greater effective utilization from biobanks. Harmonization and sustainability are the two key strategic priorities to increase interoperability among international biobank infrastructures [6]. Harmonization at a national level is also necessary and it is the first step towards global interoperability. Standard operating procedures and harmonization are required during sample collection, processing and storage. Access to comprehensive and well-organized collections of samples is needed as well as the associated clinical and research data which in turn require harmonization of operational procedures and best practices [6]. Interoperability is essential in enabling the exchange of data and samples. BBMRI is also one effort towards harmonization and development of common research infrastructures.

There exist multiple choices to implement an infrastructure that facilities these constraints. Currently there exist commercial products like BC Platforms (http://bcplatforms.com/), which handle basic biobank functionality (code, consent, transfer, log registers sample data management, links to availability database, data integration), and enable application development through API interfaces. Similarly, basic functionality could be achieved with non-commercial alternatives like Samwise, which is an in house sample management and information management system used at FIMM (the Institute for Molecular Medicine Finland) and THL biobank or Core (Code, Consent and Transfer Registry application by FIMM). This option is under consideration by several biobanks. However,

differences in storing blood sample information and the needs for pathological sample data vary. Similarly, the operational environments vary so the system is far from compliance with existing interfaces and hospital information systems. However, there is an option to construct a system from first principles relevant to the requirements of a particular biobank. The advantage of this is the possibility to devise innovative solutions and not rely on a single vendor.

5.2.1. Our solution to information management. We elected to derive a system based upon the requirements of a biobank storing pathological samples. In our solution, the primary database in the biobank will contain all the relevant data for which a patient has consented to the inclusion of in addition to all data generated during use of the biobank. Images and other non-textual data will be referenced by a URI, which can be retrieved from the data storage. It is intended that this database be scalable to accommodate future growth. Thus, efforts will be made from the commencement of this research project to facilitate distributed database schemas.

The hardware for the primary database must have a high availability and fault tolerance as time in which the system is not functioning is to be minimized. The data stored in the primary biobank database shall be primarily textual data, which will not incur significant storage cost relative to the image data stored separately. However, due to the vast number of records that may be queried, in addition to requests including image data, there will be a high memory utilization on this system. It is necessary to include solid-state storage for optimal disc access times when components may not be fully retained in memory.

The storage of large images and additional data sets that may be of an arbitrary type are to be stored in a scalable array of attached storage devices. This separation from the primary database is beneficial as it permits a faster storage mechanism on the primary database server for many queries that do not contain image data as it is cost prohibitive to create a large storage array of solid-state drives. The consequence of this approach is that there will exist latency for the real-time visualization of images. However there exist benefits in the trivial scalability of the storage system and the ease of creating a distributed file system to aid in redundancy. To effectively store the large data sets present within the biobank in addition to facilitating the scaling of storage capacities to meet future demand, external storage servers are required. The servers will be accessible only by the primary database and will appear as a single unified file system. Effective redundancy strategies and data protection strategies must be determined carefully.

Considering harmonization, the information stored in the Qpati system plays an important role. The aim is to harmonize usage of terminology by creating a common term catalog. Accordingly, using Qpati data provided by different hospitals, all data is converted into the same code spaces. Automatic conversion using UMLS can provide conversion of SNOMED (The Systematized Nomenclature of Medicine) medical terms into the same code space to some extent. However, the process needs manual verification for ensuring the quality of the output. Frequency lists are created representing the usage of terminology (parsed from Qpati data) to evaluate how the terms are used and which variables are more important for automatic analysis. This harmonization of terminology shall act as the basis for the build-up of a common availability service enabling queries over multiple biobanks. Moreover, tools to provide frequency lists can be useful also in the actual availability service.

5.3. Facilitating access to data and samples

There is a strong need to ensure confidentiality of medical data in biobanks. Such requirements are set by the European Data Protection Directive (Directive 95/46/EC) and in the forthcoming GDPR. Use of samples and data are subject to strict coding requirements unless otherwise consented to [19]. BBMRI.fi (Finnish national node of BBMRI) attempts to address these issues with an infrastructure that provides various IT tools; such as tools for ensuring security like KITE availability service (https://kite.fimm.fi/), which supports authentication and REMS (Data request and access approval system).

For facilitating access to data and samples, a national availability service shall be set up. The primary goal of a national availability service is to provide researchers a method to easily locate appropriate data/samples for research [20].

The definition of biobank research is broader than the definition of scientific research in other Finnish laws. Biobank research is defined as research, which takes advantage of samples and the related information to improve health, to understand disease mechanisms and the development of services and practices in health and patient care. The law additionally includes applied research targeting products and services [20].

Furthermore, facilitating access to cumulative digitized data enables e.g. identifying trajectories of healthy aging and developing new strategies for public health, including new tools for diagnostics, treatment and intervention. It will also create possibilities for new service and business models.

5.3.1 Our solution to data distribution. Our solution offers a web server–a boundary layer that facilitates interaction with the biobank database. In this layer, a web interface hosted on a physical server will allow users e.g. to query the biobank database, visualize image data and obtain statistics on the available data set. It is recommended that this layer not receive incoming connections from the internet. This boundary layer consists of a single physical server that requires the ability to handle multiple concurrent connections and service multiple requests. This layer will experience significant bandwidth consumption during the visualization of the image data and so there exists a large memory requirement. All communications between any two parts of the system in addition to all storage is to be fully encrypted.

The data catalog is utilized for finding information about the data that is currently available. This should be presented to the researchers in commonly agreed format and with appropriate links to the biobank for further information if available and methods to search over included metadata. In order to provide efficient services, an information pool with appropriate information collected by biobanks that could be queried in one location would be optimal. Information stored in Qpati is just one example of existing information. In addition hospitals store information in various other systems. The three-step process of forming a national availability service is described in [20]. Work on Qpati data harmonization will open up the issue by providing scripts and tools for the future. The catalog and availability service is planned to have a web-interface, and the already available KITE system is a good candidate as a platform.

The biobanks have defined basic protocols for handling research permits in their applications to Valvira. A scientific board, potentially involving an ethics committee and external consultants, processes each research application. Then a written contract is made with the background organization of the applicant. The researcher receives the samples and related information in a coded format. Challenges exist regarding the confidentiality (what kind of register information can be obtained about the donor) and on practices and policies of research results returning to the biobank as this data should be made available for future research as well as to the donors.

6. Conclusions

A biobank is a system that amalgamates data from multiple sources to facilitate multidisciplinary research by providing a repository for a large number of samples and any associated data. Data stored within the biobank is considered to exist in perpetuity and as a consequence, it is necessary to define an architecture that utilizes the knowledge of multiple fields spanning science, technology and medicine to ensure the relevance of the system in the future. The architecture presented in this document strives to define a simple architecture that can be easily scaled to meet the future demands of the biobank.

The primary considerations in this architecture are toward the safety and security of the data and services in addition to ensuring that the internal details of the system may be abstracted away when the system is utilized in production. The correctness of the system and the data stored within it is paramount.

For data to be included into the biobank, consent must be obtained and stored. Upon such time that consent is received, the data for which consent is granted is stored within the biobank and becomes available to be queried. The data may contain textual data only, or it may contain image data that may be multiple gigabytes in size.

Furthermore, there exist many challenges in implementing a biobank architecture. The system shall contain highly sensitive data with strict legal guidelines on how it may be stored and accessed. Implementation must necessarily enforce the separation of components both through the use of virtual machines hosting each component and the use of physical hardware boundaries to delimit regions of varying data sensitivity. Additionally, all communication between any two components and any storage upon which data resides are to be fully encrypted. While currently many of these challenges remain as open questions with regards to specific implementation details, this architecture attempts to account for a volatile set of requirements.

The results should interest both academics and practitioners as they provide an architecture for a biobank infrastructure. The study also lays the groundwork for further scholarly inquiry including validating the findings in practice. For practitioners, this work provides a better understanding of the key challenges to be addressed and it defines real needs of various stakeholders.

There is still a need for further work. For example the quantity of data potentially handled in biobanks creates many research problems regarding how to best index and access the data in a multi-user environment. Additionally, there exists an opportunity for further study into the effective integration and interoperability of biobanks at both a national and international level. Solutions to these problems will require further studies. Finally, implementing our architecture and deploying a functioning biobank environment should follow in order to better validate our proposed architecture.

7. References

[1] N. Boutin, A. Holzbach, L. Mahanta, J. Aldama, X. Cerretani, K. Embree, I. Leon, N. Rathi, and M. Vickers, The Information Technology Infrastructure for the Translational Genomics Core and the Partners Biobank at Partners Personalized Medicine. Journal of personalized medicine, 6(1), 6, 2016

[2] I. Demir, and M.J. Murtagh. Data sharing across biobanks: Epistemic values, data mutability and data incommensurability. New Genetics and Society, 32(4), 350-365, 2013

[3] R.A.M. El-Desouki, Best practices for human biorepositories. Biobanks. Critical Review in Pharmaceutical Science, 3(2), 2014, 6-28

[4] European Commission. Protection of personal data, http://ec.europa.eu/justice/data-protection/

[5] P.W. Hamilton, P. Bankhead, Y. Wang, R. Hutchinson, D. Kieran, D.G. McArt, J. James, and M. Salto-Tellez, Digital pathology and image analysis in tissue biomarker research. Methods, 70(1), 2014, 59-73

[6] J.R. Harris, P. Burton, B.M. Knoppers, K. Lindpaintner, M. Bledsoe, A.J. Brookes, ... and K. Zatloukal, Toward a roadmap in global biobanking for health. European Journal of Human Genetics, 20(11), 2012, 1105-1111

[7] J. Kaye, E.A. Whitley, D. Lund, M. Morrison, H. Teare, and K. Melham. Dynamic consent: A patient interface for twenty-first century research networks. European Journal of Human Genetics, 23(2), 141-146, 2014.

[8] M. Khushi, G. Edwards, D.A. de Marcos, J.E. Carpenter, J.D. Graham, and C.L. Clarke. Open source tools for management and archiving of digital microscopy data to allow integration with pathology and treatment information. Diagnostic Pathology, 8(1), 2013

[9] K. Kääriä, S. Soini, S. Kouki, and J. Suhonen, Biopankkisuostumusten hallinta Kanta-palvelujen avulla – Toiminnallinen määrittely, 2016, http://urn.fi/URN:ISBN: 978-952-302-609-4, retrieved 14.4.2016

[10] H. Kääriäinen, Benefits of Biobank research. Conference at Uppsala Konsert & Kongress, 3-4 May, 2001

[11] C. Lundström, S. Thorstenson, M. Waltersson, A. Persson, and D. Treanor, Summary of 2nd Nordic symposium on digital pathology. Journal of pathology informatics, 6, 2015

[12] A. Miettinen, M. Suhonen, J. Mykkänen, H. Virkanen, and M. Tuomainen, Needs for Open Interfaces in Personal Health Record Systems and Citizen eServices–Results from a National Survey. Finnish Journal of eHealth and eWelfare, 6(2-3), 2014, 89-102

[13] D. Mitchell, J. Geissler, A. Parry-Jones, H Keulen, D.C. Schmitt, R. Vavassori, and B. Matharoo-Ball, Biobanking from the patient perspective. Research Involvement and Engagement, 1(1), 1, 2015

[14] J. Moilanen, Potilas- ja näytetietojen siirto sairaalan tietojärjestelmistä biopankin tietojärjestelmään. Oulun ammattikorkeakoulu, 2014, http://urn.fi/URN:NBN:fi:amk-201404285147, retrieved 14.4.2016

[15] P.R. Quinlan, M. Groves, L.B. Jordan, H. Stobart, C.A. Purdie, and A.M. Thompson, The Informatics Challenges Facing Biobanks: A Perspective from a United Kingdom Biobanking Network. Biopreservation and biobanking, 13(5), 2015, 363-370

[16] M. Ranki-Pesonen, and E. Soppi, Biopankkien liiketoimintamahdollisuudet, Tekes 2014, http://www.tekes.fi/globalassets/global/nyt/uutiset/2014/tekes_biobanks_ _13_11_2014_pdf.pdf, retrieved 13.4.2016

[17] M-G. Rojo, State of the art and trends for digital pathology. Stud. Health Technol. Inform, 179, 2012, 15-28

[18] P. Runeson, and M. Höst, Guidelines for conducting and reporting case study research in software engineering. Empirical software engineering, 14(2), 2009, 131-164

[19] S. Soini, Finland on a road towards a modern legal biobanking infrastructure. European Journal of Health Law 2013, (3), 2013, 289-294

[20] STM: Biopankkilainsäädännön ohjausryhmän väliraportti 2015, 2015, http://urn.fi/URN:ISBN:978-952-00-3588-4, retrieved 13.4.2016

[21] S. Tamminen, Bio-objectifying European bodies: standardisation of biobanks in the Biobanking and Biomolecular Resources Research Infrastructure. Life sciences, society and policy, 11(1), 2015, 1-21

[22] J. Viitanen, T. Martti, and P. Kortekangas, Valtakunnallinen biopankkien kokonaisarkkitehtuuri. STM 2014, https://yhteistyotilat.fi/wiki08/download/attachments/27558004/Valtakunnallinen%20biopankki%2020140206.docx?version=1&modificationDate=1391679587390&api=v2, retrieved 13.4.2016