e-Health – For Continuity of Care
C. Lovis et al. (Eds.)

doi:10.3233/978-1-61499-432-9-388

Separation of Personal Data in a Biobank Information System

Thomas H. MÜLLER[a,1] and Reinhard THASLER[b]
[a] *IBE – Institute for Medical Informatics, Biometry and Epidemiology, University of Munich, Marchioninistr. 15, 81377, Munich, Germany*
[b] *Biobank under administration of HTCR, Department of General, Visceral, Transplantation, Vascular and Thoracic Surgery at Munich University Medical Centre, Marchioninistr. 15, 81377, Munich, Germany*

Abstract. Separation of different types of personal data has been introduced as an effective measure to improve data protection in the context of medical research. In particular, research associated with human biomaterials requires not only secure technologies but also trustworthy processing of personal data on a need-to-know basis. Web-based information systems make use of a technological infrastructure that is well suited to distributed data repositories and remote processing systems. This approach was successfully applied to develop an information system supporting acquisition, processing and storage of remnant biomaterial from surgical treatment, as well as its allocation to research projects. In order to enhance data protection, the contents of the originally unified database were divided into identification data and medical data. A web application was created for each part and appropriate functionality to maintain and access corresponding data was developed. It is concluded that a distribution of biobanking data across separate databases can be achieved if workflows and staff roles are redesigned accordingly.

Keywords. Biobanking, Data Protection

Introduction

Routine, quality managed sample and data collection in a clinical context, frequently referred to as “biobanking”, can be seen as a prerequisite in advancing translational research [1]. Critical factors, especially for biomarker validation, are not only the molecular quality of the physical sample, but include data annotation and management [2]. Therefore, appropriate IT-support is essential. In cooperation with the Clinic for General, Visceral, Transplantation, Vascular and Thoracic Surgery at Munich University Medical Centre and the Human Tissue & Cell Research (HTCR) Foundation, the Institute for Medical Informatics, Biometry and Epidemiology (IBE) at the University of Munich has developed a web based application for standardized data entry, central data management, sample storage and allocation as well as documentation, including reports, subsequently termed “HTCR Web Application”.

The HTCR Foundation was established in 2000 in Regensburg, Germany with the intent to provide a “honest broker model” [3] to foster in-vitro-research with human tissues, with the foundation acting as the donors' trustee across institutions involved in

[1] Corresponding Author.

sample and data collection as well as research. HTCR asks patients in cooperating clinics for their consent and transfer of ownership of remnant tissues and blood samples to the foundation for use by researchers under the terms of the HTCR-framework of rules and regulations. [3,4]

This special role as an external governance body is particularly important in context of "Biobanking for Research in Surgery" [5] and its project related workflow, where "starting from a clinical or scientific hypothesis, donor identification and selection is the first step in a long chain of generating information towards biobanking."[5, p. 492]. In its support of research based on human specimen, HTCR bears responsibility to its donors for full transparency of sample storage and use in line with data protection considerations. This commitment can only be fulfilled by generating samples with high quality documentation while at the same time making the information available on a strict need-to-know basis with appropriate safeguards.

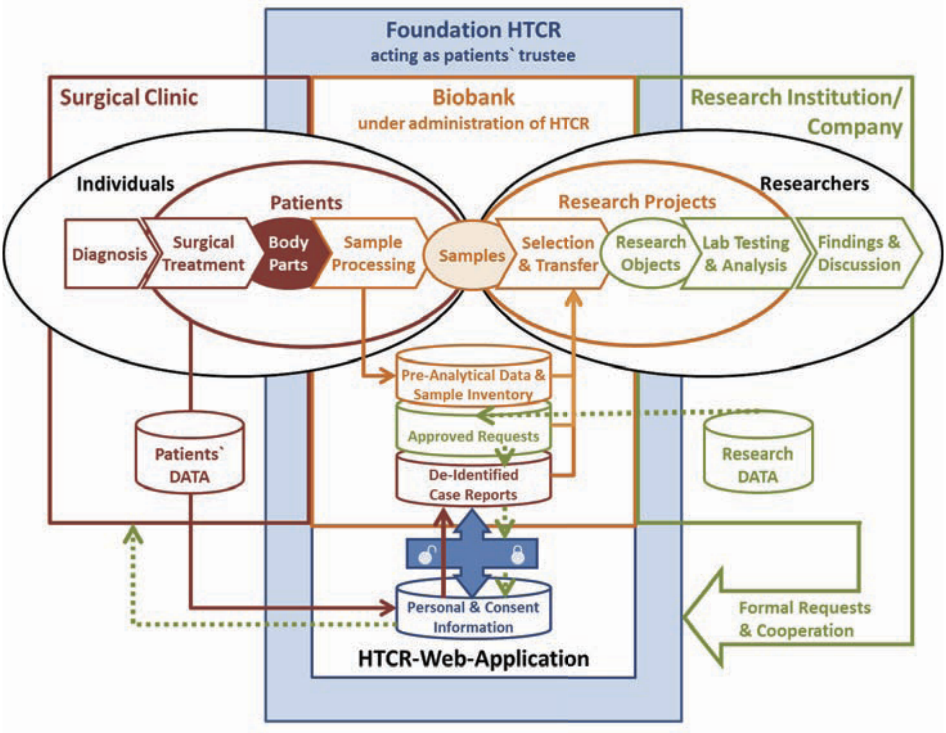

Figure 1. Overview of biobank workflow.

Fig. 1 illustrates the HTCR Web application's role within the general context of the biobanking process it is meant to support. The overall challenge for this process is to transform tissue removed from identified individuals entering the clinical context for surgical treatment into adequately annotated and at the same time anonymized samples to be used in the context of research projects.

In this paper we describe an enhancement of the HTCR Web Application that effectively alters it into a distributed information system, using separate storage for 1) personal identification data and 2) de-identified medical data as well as sample and project related data in two different, physically separated data bases as a key safeguard.

The experiences gained during this transformation process may serve to improve data protection in other kinds of research data repositories or even provide design hints for a generic implementation of similar distributed web applications.

1. Methods

The HTCR Web Application is implemented using a web forms generator and development kit called “dbform” that has been developed for electronic data capture in clinical research projects at the IBE. The base system has already been used in a number of projects as described previously [6]. In brief, it is based on standard open source components widely used in web applications, including the operating system Linux, the database management system PostgreSQL and the Apache web server. The web forms generator and the associated tools are implemented in object-oriented Perl. The base system is instantiated for each data capture project. The bulk of the data capture characteristics contained in a tabulated data dictionary. This includes role-based access control allowing the definition of arbitrary roles. In addition, presentation and functionality can be changed or extended using HTML/JavaScript templates and a variety of programming interfaces.

Additional functionality to support a hidden linkage of two separate databases, or more precisely, two web-application instances has recently been developed. This feature allows corresponding structures in both databases to be created and maintained in a consistent manner. Interlocking of information is maintained by a common reference key stored in both databases, but not circulated elsewhere. [7]

In the HTCR Web Application this feature is now used to separate donor identification data from associated clinical data and sample data. Implementing this separation in a manner that would improve data protection in turn involved changes in the biobank operational processes. Co-ordination of application development and changes in the biobank standard operations was accomplished through regular meetings with biobank staff and the HTCR data protection officer.

2. Results

The basic requirement to store donor identification data (IDAT) such as name or date of birth separate from information related to the biomaterial sample(s) and the donor's clinical condition (medical data; MDAT) stems from the data protection rationale that access to both should be strictly limited to need-to-know individuals and that a sensitive and long-term data set should not be kept all in one place.

Implementation of the need-to-know principle was found to have a profound impact on biobank operation. As a direct consequence, tasks could no longer be performed ad hoc by any staff member. Instead, specific tasks had to be assigned to specific roles that in turn were assigned to available staff on a rotational schedule. For example, no single user should be able to trace an allocated sample back to the donor identity. Two mutually exclusive core roles were designed so that simultaneous access to these data would be limited as much as possible. Combined with personnel rotation, substitution and the need to fill certain supporting roles, this reorganisation ultimately resulted in increased minimum staff requirements.

However, for an initial period of about 60 days during donor and sample documentation, simultaneous access to both, IDAT and MDAT databases was found to be necessary, because required information is available, either electronically or in hardcopy, only via the patient's name or case number. Both these identifiers are deserving of protection in the context of biobanking. Therefore, the linkage mechanism, which is fully integrated into the system's role-based access control, needed to be extended by a functionality that allows temporary access to corresponding data in both databases for a limited, possibly renewable time period.

Certain details of the database separation paradigm are also subject to other requirements related to biobank operation or to the research to be conducted with samples. Patients undergoing several surgeries for related or unrelated medical conditions may donate biomaterial on multiple occasions. To carry out research projects in a valid manner, it must be known whether samples are from the same or from different donors. Hence, both databases must support common structures that reflect both, donors and surgeries (in a 1-n relation) albeit with different sets of attributes. Table 1 lists the distribution of some of the attributes and attribute groups among the two databases.

Table 1. Distribution of Data between IDAT and MDAT Databases

Documented Object	Database 1 (IDAT)	Database 2 (MDAT)
DONOR	name, gender, date of birth, case identifier	gender, year of birth
SURGERY	date of consent and consent details, date of surgery, date of revocation	date of surgery, consent status, medical history, diagnosis, surgical procedure, laboratory test results, serology, pathology findings
SAMPLE	(not present)	organ, tissue type (e.g., tumor vs. normal), sample type, aliquots, processing data (e.g., ischemia times), allocation to research projects
RESEARCH PROJECT	(not present)	sample selection criteria (organ, tissue type, sample type,...), project information (aims and methods), recipient of sample & data, etc.

Table 1 also shows that some of the attributes are actually needed in both databases. The donor's gender is a typical example. While it is part of the patient's identification data, it is also needed in most research projects. In other cases, information derived from an identification attribute may be needed to determine a parameter of interest to research projects. While the donor's exact date of birth is normally irrelevant for research projects, it is necessary to determine the donor's age at the time the surgical treatment was performed. This determination can be made with sufficient accuracy using the year of birth. Since the latter is much less precise, it is generally not considered to be an identifying attribute in itself. In order to avoid redundant entry, an automatic update of a configurable set of attributes was implemented.

3. Discussion

Separating identification data and medical of the HTCR Web Application has shown that this type of effort needs to be accompanied by a careful redesign of workflows and roles. Obviously, there are practical limits to such redesigns. For example, there is little benefit to defining a large number of roles in order to "minimise" individual access to critical items of information, if ultimately not enough staff will be available to actually comply with a highly differentiated authorization scheme.

Likewise, the design of generic functionality – tentatively referred to as "database link" – that would facilitate the design and implementation of distributed database systems within the scope suggested by this paper is equally difficult. In particular, no obvious communication standard that is currently available offers itself for this purpose. Nevertheless, some, admittedly elementary conclusions may be drawn for a generic interoperability design:

- The linkage mechanism should distinguish between command data, identification data and medical data.
- For identification and medical data, communication should allow for attribute-value pairs and provide specific cryptographic encryption in addition to security mechanisms associated with the network transport layer. It is also practical to include some form of attribute mapping so that differences in attribute names in both databases can be resolved.
- Necessary link commands include remote object creation, update, search and an interactive session launch focused on a specified remote object. Some of these have already been specified in somewhat more detail in [7]

As a complementary measure, the data protection policy calls for relabeling of samples upon allocation with a randomly generated sample- and project-specific allocation number. This raises the level of anonymization of the allocated samples, thereby controlling the risk of unauthorized identification of donors.

References

[1] Patel A.: Tissue banking for research--bench to bedside and back--myth, reality or fast fading reality at the dawn of a personalised healthcare era. *Cell Tissue Bank.* **12** (2011) ,19-21.

[2] Moore HM, Kelly AB, Jewell SD, McShane LM, Clark DP, Greenspan R, Hayes DF, Hainaut P, Kim P, Mansfield E, Potapova O, Riegman P, Rubinstein Y, Seijo E, Somiari S, Watson P, Weier HU, Zhu C, Vaught J.: Biospecimen reporting for improved study quality (BRISQ). *J Proteome Res.* **10** (2011), 3429-38.

[3] Thasler WE, Schlott T, Kalkuhl A, Plän T, Irrgang B, Jauch KW, Weiss TS: Human tissue for in vitro research as an alternative to animal experiments: a charitable "honest broker" model to fulfil ethical and legal regulations and to protect research participants. *Altern Lab Anim.*. **34** (2006), 387-92.

[4] Thasler WE, Weiss TS, Schillhorn K, Stoll PT, Irrgang B, Jauch KW: Charitable State-Controlled Foundation Human Tissue and Cell Research: Ethic and Legal Aspects in the Supply of Surgically Removed Human Tissue For Research in the Academic and Commercial Sector in Germany. *Cell Tissue Bank.* **4** (2003), 49-56.

[5] Thasler WE, Thasler RM, Schelcher C, Jauch KW: Biobanking for research in surgery: are surgeons in charge for advancing translational research or mere assistants in biomaterial and data preservation? *Langenbecks Arch Surg.* **398** (2013); 487-99.

[6] Müller TH: Central IT-Structures for Integrated Medical Research and Health Care of Viral Hepatitis – Hep-Net. *Stud Health Technol Inform.* **116** (2005), 1016-20.

[7] Müller TH: Key-linked on-line databases for clinical research. *Stud Health Technol Inform.* **180** (2012), 524-8.