The International Technology Management Review, Vol. 4 (2014), No. 2, 68-77

Data Sharing Challenges and Recommendations for Human Biorepositories: A Systematic Literature Review

Jitendra Jonnagaddala
Asia-Pacific ubiquitous Healthcare research Centre (APuHC), University of New South Wales, Kensington, Sydney, NSW 2031, Australia
Translational Cancer Research Network, University of New South Wales, Kensington, Sydney, NSW 2031, Australia
E-mail: jitendra.jonnagaddala@unsw.edu.au

Toni Rose Jue
SingHealth Experimental Medicine Centre, Singapore Health Services, Block 9, SGH campus Singapore 169608
E-mail: tonirose_jue@outlook.com

Pradeep Ray
Asia-Pacific Ubiquitous Healthcare Research Centre, University of New South Wales, Kensington Sydney, NSW 2031, Australia
E-mail: p.ray@unsw.edu.au

Amir Talaei-Khoei*
School of Systems, Management and Leadership, University of Technology, Sydney, CB10.04.346, Po. Box 123, Ultimo NSW 2007, Australia
E-mail: amir.talaei@uts.edu.au

Abstract

Biorepositories collect biospecimens like blood and tissue from the patients along with associated clinical and pathological data. Medical research tends to require sharing of biospecimens and associated data at a large scale across multiple biorepositories. The objective of this review is to identify issues faced and recommendations made in sharing biospecimens data. Our review identified that significant amount of work has been done in the past and biorepositories can take advantage of this work by working collectively.

Keywords: biobanks, biorepositories, biobanking informatics, data sharing.

* Dr Amir Talaei-Khoei is also a member of Asia-Pacific Ubiquitous Healthcare Research Centre.

Published by Atlantis Press

1. Introduction

Biorepositories, also called biobanks, are collections of biospecimens for conducting research. The human Biorepositories collect and preserve biospecimens like blood, tissue, bone marrow, etc., from the general population or patients with specific clinical diagnosis. Associated clinical and pathological data are also collected with biospecimens. The collected biospecimens are sometimes stored for more than a decade. Hence, biorepository information systems are used to maintain the data associated with biological specimens. The type of biorepository information systems used vary on biorepository size and type. However, human Biorepositories usually use Microsoft Excel/Access, custom built or open source web applications.

Biomedical research, especially in cancer research, tends to require sharing of biospecimens and associated data at a large scale across hospitals, research institutions and organizations[1]. It has been agreed upon that sharing data and resources between biorepository information systems will contribute to improving our health system as recognized in the Data Sharing directive of the National Institutes of Health[2]. Sharing data between biorepository information systems allows for greater access to data from all health systems and may advance our understanding of human health and diseases. However, biorepository information systems have limitations when it comes to sharing data locally and internationally between various Biorepositories[1, 2]. Biorepository information systems must share a set of common data elements (CDE), or a common standards framework to achieve reproducible, inter-institutional data sharing. The issue with achieving a set of CDEs is that biorepository information systems have been developed from a variety of perspectives and are therefore often difficult to integrate or network within and across organizations[3]. Even within organizations, it has been found that new databases are created for each study as scientific data stored in disparate formats, and/or as systems in multiple sites with different workflows making it difficult to aggregate data across studies for meta-analysis. Issues with the lack in common data elements in data systems include not allowing real time access to data collections, or fine-tuning data collection strategy in the field collection stage[4]. The need to standardize common data elements between institutions is growing as biospecimen collections around the world are increasing as evidenced with the DCEG biospecimen program which is rising at an annual rate of over 15% since 1999[4].

There have been many attempts at achieving inter-institutional collaborative research, successful examples include the Cooperative Prostate Cancer Tissue Resource (CPCTR), Pennsylvania Cancer Alliance for Bioinformatics Consortium (PCABC), Cooperative Breast Cancer Tissue Resource (CBCTR), Cooperative Human Tissue Network (CHTN), Cancer Family Registries (CFR), and the Early Detection Research Network (EDRN), however there still lacks an agreed common set of data elements globally[5-8]. To achieve inter-institutional collaborative research on a global scale, we have to analyze systems that have been successfully integrated with other systems using a defined set of common data elements and understand what factors made it successful as well as the issues that were incurred. Hence, a systematic literature review was conducted on literature published between 1995-2013 to identify, analyze and summarize these data sharing challenges, and also discuss the solutions proposed in literature. The main objective of this review is to highlight the longstanding issues faced by researchers in sharing biospecimens data at an individual biorepository level. As per our knowledge, no such review has been carried out before and hopefully the findings can then be useful in addressing the challenges at a global scale.

2. Methods

A search for literature written in English was conducted using the following databases: Google Scholar, Springerlink, Wiley InterScience, Science Direct, Scirus, IEEE Xplore, PubMED. Keywords were discovered by reading relevant literature and by using the thesaurus to find any words relating to biorepository information management systems. Keywords used to comprehensively search through the databases was sorted into two groups (Table 1) to reduce the chances of missing relevant articles. The first group included words relating to biorepositories, the second group related to information systems. All databases were then

searched for any papers that contained one word from group 1 as well as one word from group 2.

Group	Keywords
Group 1	"tissue banks", "tissue repositories", biobank, biorepositories, biospecimens, "biological specimens", "tumour banks", "tumor banks", "bio-banks", "bio-repositories", "bio-specimens", "bio-logical specimens"
Group 2	DBM, "management systems", "inventory systems", LIS, LIMS, BIMS, app, application, "application software", "application system", "application package", "bundled software", software, "software application", "information system", program, package, DBMS, "database management system", "laboratory information management system", "laboratory information system", " biobank information management system", "enterprise system", informatics

Table 1: Keyword Groups used for literature survey

An example of a search phrase used would be:
"tissue banks" AND (DBM OR "management systems" OR "inventory systems" OR LIS OR LIMS OR BIMS OR app OR application OR "application software" OR "application system" OR "application package" OR "bundled software" OR software OR "software application" OR "information system" OR program OR package OR DBMS OR "database management system" OR "laboratory information management system" OR "laboratory information system" OR " biobank information management system" OR "enterprise system" OR informatics).

In terms of exclusion criteria, papers published before 1995 were excluded. Patents, citations, reference papers, abstracts as well as any paper written in languages other than English were also excluded. Articles were excluded if they did not describe or discuss biorepository information systems data sharing aspects. The research methodology was based on Kitchenham[9] and Kitchenham et al[10]. The screening process is illustrated in Figure 1.

3. Results

Electronic databases were screened using the keywords stated in Table 1, after which 72,443 results were found to include keywords related to Biorepositories and information systems. Of these articles 911 were found to relate to our review based on their title. Of these 911 articles, 474 papers had abstracts describing biorepository information systems and 358 articles were found to have been repeated from previous searches. Finally, 21 papers were found to describe examples and issues of biorepository data sharing after analyzing the full text [1-3, 5, 6, 11-26]. These papers were shortlisted for this literature review. In all the studies identified data sharing challenges are highlighted with examples and at the same time recommendations were made to address those challenges. Our review highlights these challenges and recommendations in the following sections.

3.1. *Challenges*

Our review study identified several challenges documented in literature. At first we noticed that isolated development of databases is a major challenge followed by entry barriers, different regulations between countries, interpretations of Common Data Elements

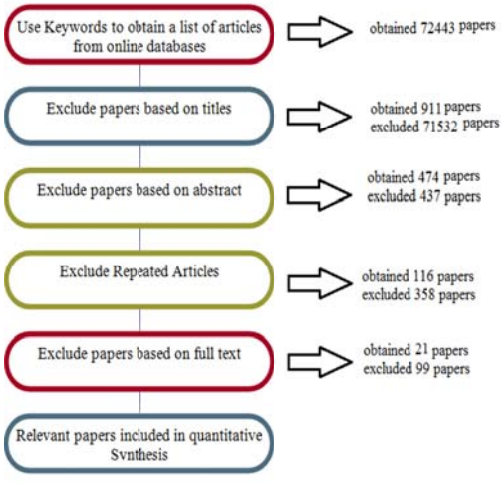

Figure 1. Article screening and selection process.

and free text records. These challenges need to be considered in future biobanking informatics studies.

3.1.1. *Isolated development of databases & Common Semantics*

Databases for biospecimens are often developed in isolation of other databases or research systems, and therefore have challenges in developing a common framework. Challenges in developing in conjunction with other systems include conflicting requirements and design patterns. Common frameworks are needed for biorepostiories to integrate with anatomic and clinical pathology information systems which produces efficient workflows[3]. An issue with common semantics is that differing areas of clinical practice tend to develop their own actual or de facto standards (both terminology and common data elements) somewhat in isolation of other areas. Lack of participation by institutions in the development of these standards can often lead to reluctance to adopt them, even if the standards meet the needs of the new community. The general approach to problems is twofold, harmonization and community involvement[27].

The database TubaFrost is a virtual European human tumor tissue bank designed for the scientific community which is composed of high quality frozen tumor tissue sample collections[23]. All tissue material is stored at the local repositories of the cancer centers and universities, but an online network system links all the local tumor banks together. The TubaFrost network generates common data elements for all specimens by getting collectors to register through the central database using a user name and password to upload specimens from their local storage locations. Initial problems found when implementing this database was with the internet based automatic update system due to the diversity of computerized inventory systems and IT security rules used by the participating institutions as well as the local languages used for patient data description. To combat this problem tissue records were uploaded individually or in batches, which could be done manually or automatically from the local database. For this to be done a universal export file was needed, wherein all databases were compatible with a tab-delimited text file format (*txt), which was used to export tissue records to the central database. Records are checked for errors before being uploaded into the central database[14]. Even after implementing that, the challenge in integrating data from different data sources is not resolved mainly because the information is structured in different data sources. Data integration has three steps: i) data is extracted and harmonized into a common format, ii) data is transferred into the data mart, and iii) data is loaded into a federated database. During these three steps, systems need to know the common semantics for linking the data[19].

3.1.2. *Entry barriers & resistance*

To achieve widespread adoption of sharing data between biorepositories, it is important to keep costs and barriers to use standards and common semantics [3, 20, 28]. A major weakness is the enormous amount of time needed to be invested before information can be retrieved, as information on samples continuously evolve and numbers of samples rapidly increase. The federated biospecimens database project by University of Helsinki showed that it is possible to share complex phenotype/genotype information between Australia and European countries but also highlighted fear of sharing data due to who will access it[29]. Muilu et al. and Colledge et al. also highlighted that biobanks often show resistance to share data because of entry barriers, like lack of resources and technical expertise[29, 30].

3.1.3. *Different regulations between countries & language barriers*

The regulatory requirements in different countries tend to develop in isolation, differing in safety of privacy and intellectual property environment [18]. The language differences between countries may require translations of common standardized terminologies which may not be practical. Common data elements can be difficult to integrate as language differences and variations in coding systems make uploading local data to central database difficult as all local terms need to be converted to English. This may not be practical for all biobanks spread across geographically[31].

3.1.4. *Different interpretations of CDE between institutions*

CDE developed by various institutions might have different objectives and goals for CDE development. This creates an opportunity for bias and confusion. The same common data element can be interpreted very differently by different organizations. Merging CDE

from multiple institutions or organizations can result in the loss of consistency and quality as different institutions have different interpretations of the CDE. Collection and local storage of metadata for each of the CDEs and common quality assurance protocol is needed across institutions to ensure all biorepositories report CDEs in the same manner [6, 22]. However, issues might include keeping up with ongoing annotation of older cases while adding newer cases which become increasingly time- and labor-intensive. Another issue was some PSA values were unidentifiable due to certain patient healthcare habits. It is hard to share data without a common framework. These issues are clearly emphasized by Patel et al. with their Tumor repository project [6, 21, 22] . Common terminologies may solve collaborations from geographically distributed researchers of the same domain, but this needs to be discussed when designing databases[32].

3.1.5. *Protecting patients' privacy when sharing data*

In the recent years, patients' privacy in research has become one of the highly discussed topics in research communities. Researchers accessing biospecimens to conduct studies need access to patient data associated with the specimen. Often this data is distributed by biorepository staff in de-identified format to researchers with valid Institutional Review Board (IRB) approval. However, the biorepository staff have access to identified patient data. Hence, the biorepository staff needs to go through the process of de-identification using honest broker services on ad hoc basis. This is a significant challenge in sharing data with biobanks. In other words, it is difficult to share data automatically without an automated de-identification mechanism. Although there are several automated de-identification systems, the accuracy of such systems are low [33-35]. However, de-identification systems are getting better over time.

3.1.6. *Issues in integrating unstructured data*

Most of individual and small biorepositories store specimen data and clinical data in unstructured formats. The unstructured formats may vary from simple excel sheets to free text word documents. Free text words documents are especially used in storing clinical data associated with specimen. Often these free text records are manually extracted from hospital information systems. This practice of sharing data is not effective if there are multiple biobanks involved. In integrating clinical information from the National biobank of Korea, thirty-one common data elements are associated with all types of diseases. However, five out of thirty-one common data elements could not be integrated automatically as they are in free text format which seems to be a major problem in integrating data from multiple systems[17].

3.2. *Recommendations*

The first step toward ensuring that biorepositories can share data easily — is to adopt best practices in gathering and storing the data. This goes hand in hand with the adoption of standardized bioinformatics solutions like caTissue software. Secondly, we must establish a more flexible common biorepository model. These recommendations would lead to the development of a flexible common biobanking data sharing framework.

3.2.1. *Best practices in managing data*

Recommendations on best practices for biorepostiories, which includes data management methods, was published by the International Society of Biological and Environmental repositories (ISBER) and The National Cancer Institute, Office of Biorepositories and Biospecimens Research (NCI-OBBR)[25, 36, 37]. There are a number of good commercial databases available that meet the ISBER and OBBR recommendations[3]. Some of the databases are certified by ISBER biobanking community. There are also several databases from commercial vendors which participated in best practice implementation initiatives by the National Institute of Health, USA. Ardini et al. clarified that it is easy to share specimen data if the data is gathered, stored and managed using best practices [25, 36, 37]. By following the best practices in data collection, storage and management, the overall performance of biorepository as a unit might also improve[36].

3.2.2. *Adopting caTissue biobanking software*

caTissue software has been developed by the National Cancer Institute as an open source reference system to manage biorepositories and support biospecimens science[15, 16]. caTissue is a web-based application

specifically developed by taking data sharing challenges into consideration. Those principles include role based security, UML driven architecture, and semantically annotated, reusable data elements that leverage standardized vocabularies and ontologies[16]. External systems can be integrated with the caTissue easily. Components that help caTissue interact and exchange data with external systems can also be easily developed. The data models and UML model are the underlying components that provides the basis for connectivity across multiple biobanks. If all biobanks in a network are using caTissue software it is very easy to share the data between biobanks by adding a few additional components on top of the data model & UML model components. The caTissue system also provides a standardized compliant API for accessing domain-object data, which can be utilized by other biorepository systems to access and search on annotations data easily[38].

3.2.3. *Common Biorepository Model (CBM)*

The Common Biorepository Model (CBM) allows existing commercial and open source biorepository systems to share data for research. It also allows a search for specimens within and across repositories. CBM is capable of reducing the time and effort required by a researcher to locate specimens and it only shares summary level data[3]. This model was developed by the community as part of caBIG initiative led by the National Cancer Institute. The model was adopted by several vendors and carried out various pilot studies. The model in its current state doesn't capture all the data of a specimen. However, this model can be easily extended[13, 39]. It is also identified from our literature search that various studies and programs have reused the CBM.

3.2.4. *Use existing vocabularies & terminologies*

In terms of vocabularies or ontologies, existing standards under the UMLS umbrella can be leveraged for biorepositories and biospecimens science (e.g. MeSH, NCIt, LOINC, and SNOMED) [25, 36, 40]. As such, we believe that there is no need to develop new vocabulary standards or ontologies. Well annotated specimens are important for research and annotations made using standard vocabularies play vital role in sharing data between multiple biorepositories. The ambiguity of context in the annotated data can be easily resolved. The usage of existing vocabularies and standards does not only apply to the specimen annotations data but it could also apply to any other supplementary data associated with the specimen. For example, specimens sometimes are also associated with imaging data and by using appropriate standards the imaging data can be easily interpreted by researchers who are not imaging experts.

3.2.5. *Adopt centralized systems*

The solution proposed for standardizing a number of local databases with disparate formats in disparate systems at multiple sites with different workflows, was to develop a centralized scientific data management system. This centralized system would bring together information from disparate sources, transform the data into a uniform format (conform with standardized data element definitions) and load the data into the Central Data Warehouse (CSW) where analysis could be performed[4]. One concern for this is that standardizing data before it is entered into the database is time consuming. The process of extracting, transforming and loading can be very resource intensive. However, adopting centralized systems doesn't address the problem at a global level. Adopting centralized systems is more suitable for biorepositories which are part of a network with rigid structures in management that has no change in scope and function.

3.2.6. *Data and Meta data standards from caBIG program*

The National Cancer Institute (NCI) commissioned the cancer Biomedical Informatics Grid (caBIG) program in 2004 which deals with creating the technical and sociological infrastructure that enables interoperability between health information systems created in a federated environment. caBIG was built upon 4 components to achieve interoperability, interface integration, information models, controlled terminology and common data elements. Common data elements is a semantic component, describing the recorded data and the context in which the data is recorded[18]. As part of this program, several data and metadata standards are developed to define data elements in the biomedical field. Even though these standards are based on a model-driven architecture, they can be easily reused and extended to suit the needs of the biobanking community. For example, caTIES functions within the

context of caBIG to enable interoperability between other cancer research systems[33]. caTIES component extract information from free text pathology reports and load this information into caTissue standards. Similarly, other related components used in sharing data should leverage on existing data and metadata standards to achieve interoperability.

3.2.7. Establish flexible CDE frameworks

It is impossible to create any standard terminology of CDE that meets the requirement of the global biomedical research and care delivery community, but identification of the core data elements for aggregation of specific classes of data and an agreement upon a harmonized means of capturing this data is essential. Involvement of relevant communicates and harmonized standards is needed in order for an appropriate sense of ownership to exist[18]. Establishing a flexible CDE framework, which allows researchers and biorepository staff to customize, can play a vital role in addressing data sharing challenges. The CDE framework should accommodate the growing needs of the biobanking community and researchers accessing the biorepositories.

4. Discussion

There have been an increasing number of national and statewide initiatives aimed at supporting the development of tissue banks promoting the sharing of tissue and data resources. At present, multiple institute participation is a feature of several tissue banks including the Cooperative Prostate Cancer Tissue Resource (CPCTR), Pennsylvania Cancer Alliance for Bioinformatics Consortium (PCABC), Cooperative Breast Cancer Tissue Resource (CBCTR), Cooperative Human Tissue Network (CHTN), Cancer Family Registries (CFR), and the Early Detection Research Network (EDRN). The technique and means of data and tissue collection vary in each biorepository, although many of these multi-institutional collaborations have been compelled to develop principles for sharing data with other groups due to the pressing need for well annotated tissues that can be re-annotated with experimental data[38]. Some of these projects are discussed below from a data sharing perspective.

The NCI Cooperative Prostate Cancer Tissue Resource (CPCTR) is a progressive project that has previously discussed formative and infrastructural issues (i.e. organization issues, procurement issues, IRB issues) related to its initial activities for marketing the large number of annotated prostate cancer specimens to potential investigators[6, 22]. The CPCTR project is a good candidate for discussion about data sharing between biorepositories. The group's work describing the process of developing common data elements for prostate cancer tissues laid the foundation for much of the CPCTRs central database. In addition, by demonstrating how to implement the open access, the Tissue Microarray Data Exchange Specification allowed the Resource to share and merge data with other tissue microarray (TMA) files or link to data contained in external biological databases[6]. CPCTR have several common data elements to annotate tissue samples collected. They include patient-level demographics and clinical history data. They also include pathology details of specimen to describe the cancer staging, grading and other characteristics of individual surgical pathology cases; tissue block-level annotation critical to managing a virtual inventory of cases and facilitating case selection; patient level clinical outcome data including treatment; biochemical (prostate specific antigen values) and clinical recurrence; and vital status.

In the CPCTR project the process of developing CDEs typically involves many individuals and can take up to several months to arrive at a draft that is based on a complete consensus among those involved. In the case of CPCTR there were pathologists, urologists, cancer registrars, data managers, and cancer researchers from five major medical centers and the NCI Cancer Diagnosis Program who provided input and approved changes to the developing CDEs along the process of adopting the initial version. In this process, it was essential to 1) include experts from multiple disciplines, 2) consider the works of others creating similar CDEs, and/or 3) consider established standards when available. This communication describes the significance of cross domain collaboration for successful data sharing standards. However, the same communication process can delay or extend the project timelines.

The GenomEUtwin project linked patients using a de-identified system, assigning a unique randomized

identifier for each subject[29]. This identifier consists of four parts: country, randomized number, twin identification number and a check sum. Each center is responsible for creating and maintaining the EU id numbers for their individuals. Data shared is de-identified. This project harmonized and integrated studies across Europe and Australia [29, 30, 38]. The databases currently contain integrated information for the initial test traits such as weight and height (and BMI) from more than 250,000 individuals, for migraine questionnaire and details of clinical phenotype from 8,000 individuals, and for serum lipid values, insulin and glucose content and other measures of metabolic traits for over 20,000 individuals. Data integration is achieved by first extracting the data from the original database and then harmonizing it after. This was an extremely time consuming exercise due to the differences in design and annotation. Data is then transferred to the site, checked and stored. Similar to CPCTR project this project's data sharing was also very tedious and resource intensive.

The National Mesothelioma Virtual Bank is also another good example which uses common data elements to integrate its databases, and they only include de-identified data entered through a web portal[12]. The problem is incorporating patient data from multiple sources with de-identified data, this is resolved by linking common patient identifiers or health information. Although, errors still occur due to the absence of specific identifiers unique to the patients [11, 12, 38]. It can be inferred from both above discussed projects that researchers are not able to leverage on latest technologies mainly due to the legal, ethical and social issues surrounding collection and distribution of human biospecimens[4, 28, 41]. Patients participating in research are not aware of finer details of how their data is handled and not all patients might understand the technicalities involved in handling data like de-identification. This also encourages few patients to deny consent to share their data because they are confident about data governance policies of biorepositories.

Many institutions currently use labor intensive manual processes to identify cases from archived tissue collections and legacy databases. Various tools are used to integrate data. First, de-identified data is electronically recorded. Reports are then parsed into fields or chucks specified in to a preferred XML scheme. Third, text is auto coded so all medical concepts in text receive a code derived from a standard unencumbered vocabulary [6, 22]. When the same tools and processes are not used by another partner biorepository, it is very hard to share data. Local data collection methods vary at each institution based on personnel. Some have cancer registers, while others have registry systems. Also, CDE's definitions and associated metadata need to be clearly understood: i) what the fundamental definition of the data element is (i.e., date of diagnosis), ii) how that data element will be collected (e.g. 11/2003 vs. Nov. 2003 vs. 11/03, etc), iii) what is the consensus on acceptable values or codes are for the data element (e.g., precise date of birth, not calculated from clinical records where the "patient appears to be a well-developed 75 year old"), and iv) what the acceptable data format is for inclusion into the central database (e.g., dates as integers not character strings). A shared biorepositories informatics approach avoids these issues in a consortium of biorepositories[28].

Most tissue bank databases which contain clinical data generally lack the ability to exchange information in a common format (syntactic interoperability), and the ability to understand and use the information once it is received by other systems (semantic interoperability). Another problem is the panoply of ways that similar or identical concepts or data are described by different users even within the same institution as well as across institutions. For example, a data element called "grade of tumor" can be collected using various formats (e.g. some collect "grade-1, grade-2, and grade-3" vs. others who collect "low grade, intermediate grade, and high grade"). Such inconsistency in data descriptors makes it nearly impossible to aggregate and manage even modest-sized data sets in order to perform basic queries[3, 38]. As a result, these systems are neither uniform nor flexible. They are incapable of performing exchange or sharing of information, as unambiguous interpretation of the information is not possible without semantic and syntactic interoperability.

5. Conclusion

Over the past decade, researchers have developed or initiated various projects to implement and improve existing biorepository data standards. The vocabularies, data exchange formats, ontologies and object models for

biospecimens are developed. Given the strength and coverage of existing standards and community efforts, there is no apparent need to develop new data models, data exchange formats, vocabularies, ontologies, or guidelines for systems to support biospecimens science. Rather, existing standards from public databases, biological and biomedical investigations, programs like caBIG and CDISC can be leveraged, extended, integrated and in some cases refactored. Data should be exchangeable across multiple national and international sites in a secure and reliable manner. Worldwide biobanking community indicated that biobanks should agree on minimum data sets that are interchangeable between biobanks and identify the complete ontology and multi-lingual definitions of the data set. Common data elements must only include data that is de-identified to protect patient privacy[24]. And it must be built in a way to accommodate the ever-evolving requirements of participants and users of biobanks [7, 26, 41].

Acknowledgements

This project was supported by a Cancer Institute NSW's translational cancer research centre program grant. The views expressed herein are those of the authors and are not necessarily those of the Cancer Institute NSW.

References

1. Lemke, A., et al., *Public and biobank participant attitudes toward genetic research participation and data sharing.* Public Health Genomics, 2010. **13**(6): p. 368-377.
2. London, J.W. and D. Chatterjee. *Using the semantically interoperable biospecimen repository application, caTissue: End user deployment lessons learned*. in *BioInformatics and BioEngineering (BIBE), 2010 IEEE International Conference on*. 2010. IEEE.
3. Fearn, P.A., et al., *Towards a Common Informatics Framework for Biorepositories.*
4. Henderson, M., et al., *Challenges of Scientific Data Management for Large Epidemiologic Studies.* Cell Preservation Technology, 2005. **3**(1): p. 49-53.
5. Dhir, R., et al., *A multidisciplinary approach to honest broker services for tissue banks and clinical data.* Cancer, 2008. **113**(7): p. 1705-1715.
6. Patel, A.A., et al., *The development of common data elements for a multi-institute prostate cancer tissue bank: the Co operative Prostate Cancer Tiss ue Resource (CPCTR) experience.* BMC cancer, 2005. **5**(1): p. 108.
7. Qualman, S., et al ., *Establishing a tumou r bank: banking, informatics and ethics.* British journal of cancer, 2004. **90**(6): p. 1115-1119.
8. Qualman, S.J., et al., *The role of tumor banking and related informatics*, in *Expression Profiling of Human Tumors*. 2003, Springer. p. 103-117.
9. Kitchenham, B., *Procedures for performing systematic reviews.* Keele, UK, Keele University, 2004. **33**: p. 2004.
10. Kitchenham, B., et al., *Systematic literature reviews in software engineering–A systematic literature review.* Information and software technology, 2009. **51**(1): p. 7-15.
11. Amin, W., et al., *An informatics supported web-based data annotation and query tool to expedite translational research for head and neck malignancies.* BMC cancer, 2009. **9**(1): p. 396.
12. Amin, W., et a l., *National Mesothelioma Virtual Bank: a standard based biospecimen and clinical data resource to en hance translational research.* BMC cancer, 2008. **8**(1): p. 236.
13. Cuticchia, A.J., et al., *NIDDK data repository: a central collection of clinical trial data.* BMC medical informatics and decision making, 2006. **6**(1): p. 19.
14. Isabelle, M., et al., *TuBaFrost 5: multifunctional central database application for a European tumor bank.* European Journal of Cancer, 2006. **42**(18): p. 3103-3109.
15. Jonnagaddala, J., *An open source biospecimen informatics application for translational research*, 2012, University of New South Wales.
16. Jonnagaddala, J., J. L i, and P. Ray . *Evaluation of caBIG® caTissue Software*. in *World Congress on Medical Physics and Biomedical Engineering May 26-31, 2012, Beijing, China*. 2013. Springer.
17. Kim, H., et al., *Integrating clinical information in National Biobank of Korea.* Journal of m edical systems, 2011. **35**(4): p. 647-656.
18. Komatsoulis, G.A. *caBIG™: Opportunities and challenges to creating a federated global network of interoperable information systems*. in *BioInformatics and BioEngineering, 2008. BIBE 2008. 8th IEEE International Conference on*. 2008. IEEE.
19. Litton, J.-E., *Biobank informatics: connecting genotypes and phenotypes*, in *Methods in Biobanking*. 2011, Springer. p. 343-361.
20. Manley, G.T., et al., *Common data elements for traumatic brain injury: recommendations from the biospecimens and biomarkers working group.* Archives of physical medicine and r ehabilitation, 2010. **91**(11): p. 1667-1672.
21. Patel, A.A., et al., *An informatics model for tissue banks–Lessons learned from t he Cooperative Prostate Cancer Tissue Resource.* BMC cancer, 2006. **6**(1): p. 120.
22. Patel, A.A., et al., *Availability and quality of paraffin blocks identified in pathology archives: a multi-institutional study by the Shared Pat hology*

Informatics Network (SPIN). BMC cancer, 2007. **7**(1): p. 37.
23. Riegman, P., et al., *TuBaFrost: European virtual tumor tissue banking*, in *New trends in cancer for the 21st century*. 2006, Springer. p. 65-74.
24. Thasler, W.E., et al., *Human tissue for in v itro research as an alternative to animal experiments: a charitable" honest broker" model to fulfil ethical and legal regulations and to protect research participants.* Alternatives to laboratory animals: ATLA, 2006. **34**(4): p. 387-392.
25. Vaught, J., et al., *An NCI per spective on creating sustainable biospecimen resources.* JNCI Monographs, 2011. **2011**(42): p. 1-7.
26. Watson, R.W.G., E.W. Kay, and D. Smith, *Integrating biobanks: addressing the practical and ethical issues to deliver a valuable tool for cancer research.* Nature Reviews Cancer, 2010. **10**(9): p. 646-651.
27. Harris, J.R., et al., *Toward a roadmap in global biobanking for health.* European Journal of Human Genetics, 2012. **20**(11): p. 1105-1111.
28. Jonnagaddala, J. and D. Sue, *A report on the Workshop on Biobanking Informatics*, 2013.
29. Muilu, J., L. Peltonen, and J.-E. Litton, *The federated database–a basis for biobank-based post-genome studies, integrating phenome and genome data from 600 000 twin pairs in Eu rope.* European Journal of Human Genetics, 2007. **15**(7): p. 718-723.
30. Colledge, F., B. El ger, and H.C. Howard, *A Review of the Barriers to S haring in Biobanking.* Biopreservation and Biobanking, 2013. **11**(6): p. 339-346.
31. Karlsen, J.R., J.H. Solbakk, and R. Strand, *In the Ruins of Babel: Should Biobank Regulations be Harmonized?*, in *The Ethics of R esearch Biobanking*. 2009, Springer. p. 331-343.
32. Myneni, S. and V.L. Patel, *Organization of biomedical data for collaborative scientific research: A res earch information management system.* International journal of info rmation management, 2010. **30**(3): p. 256-264.
33. Crowley, R.S., et al., *caTIES: a g rid based system for coding and retrieval of surgical pathology reports and tissue specimens in support of translational research.* Journal of the American Medical Informatics Association, 2010. **17**(3): p. 253-264.
34. Garla, V., et al ., *The Yale cTAKES extensions for document classification: architecture and application.* Journal of the American Medical Informatics Association, 2011. **18**(5): p. 614-620.
35. Savova, G.K., et al., *Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications.* Journal of the American Medical Informatics Association, 2010. **17**(5): p. 507-513.
36. Ardini, M.-A., et al., *Sample and data sharing: Observations from a cent ral data repository.* Clinical biochemistry, 2013.
37. Wichmann, E., *Need for guidelines for standardized biobanking.* Biopreservation and Biobanking, 2010. **8**(1): p. 1-1.
38. Mohanty, S.K., et al., *The development and deployment of Co mmon Data Elements for tis sue banks for translational research in cancer–An emerging standard based approach for the Mesothelioma Virtual Tissue Bank.* BMC cancer, 2008. **8**(1): p. 91.
39. Freimuth, R.R., et al., *Life sciences domain analysis model.* Journal of the American Medical Informatics Association, 2012. **19**(6): p. 1095-1102.
40. Pan, H., et al., *'What's in the NIDDK C DR?'—public query tools for the NIDDK central data repository.* Database: the jou rnal of biological databases and curation, 2013. **2013**.
41. Eiseman, E., et al., *Case studies of exis ting human tissue repositories:" best practices" for a biospecimen resource for the g enomic and proteomic era*. 2003: Rand Corporation.