BIOPRESERVATION AND BIOBANKING
Volume 10, Number 4, 2012

DOI: 10.1089/bio.2012.0003

A Minimum Data Set for Sharing Biobank Samples, Information, and Data: MIABIS

Loreana Norlin, Martin N. Fransson, Mikael Eriksson, Roxana Merino-Martinez, Maria Anderberg, Sanela Kurtovic, and Jan-Eric Litton

Numerous successful scientific results have emerged from projects using shared biobanked samples and data. In order to facilitate the discovery of underutilized biobank samples, it would be helpful if a global biobank register containing descriptive information about the samples existed. But first, for shared data to be comparable, it needs to be harmonized. In compliance with the aim of BBMRI (Biobanking and Biomolecular Resources Research Infrastructure), to harmonize biobanking across Europe, and the conclusion that the move towards a universal information infrastructure for biobanking is directly connected to the issues of semantic interoperability through standardized message formats and controlled terminologies, we have developed an updated version of the minimum data set for biobanks and studies using human biospecimens. The data set called MIABIS (Minimum Information About BIobank data Sharing) consists of 52 attributes describing a biobank's content. The aim is to facilitate data discovery through harmonization of data elements describing a biobank at the aggregate level. As many biobanks across Europe possess a tremendous amount of samples that are underutilized, this would help pave the way for biobank networking on a national and international level, resulting in time and cost savings and faster emergence of new scientific results.

Introduction

The term "biobank" has been defined by Organisation for Economic Co-operation and Development (OECD) as a collection of biological material and the associated data and information stored in an organized system for a population or a large subset of a population.[1,2] With that in mind, the scientific value of biobanks can be made much more significant for many research purposes if multiple biobanks are connected to enable sharing of samples and information.[3] Connecting detailed and up-to-date bioinformatics databases for the purpose of open access to knowledge has long been identified within the EU as two of the three pillars for the harvesting of the potential of life sciences and biotechnology.[4] In order for shared data to be comparable, it needs to have the same definitions. However, this is seldom the case. Although minimum information checklists are successfully being developed in different projects for diverse biological and technological areas,[5,6] biomedicine is in fact lacking guidance when it comes to sharing and pooling data from samples, clinical trials, electronic health records, and questionnaires. The primary reasons are the difficulty to harmonize data and personal data protection.

Numerous successful research results have emerged from projects using shared biobanked samples and data.[7] In fact, many biobanks possess a tremendous amount of samples that are underutilized but make up a potential gold mine for research projects[3,8] once the samples are found, and ethical and legal aspects straightened out. To overcome the hurdles in data sharing that personal data protection issues brings about, we suggest standardizing data on a descriptive level as a first step.

In this article, we present the minimum biobank and sample collection attributes that will help researchers initiate collaborations on biobanked samples. The data set called Minimum Information About BIobank data Sharing (MIABIS), is an elaboration of the minimum data set proposed by 8 countries in 2009 within the preparatory phase of the European project BBMRI (Biobanking and Biomolecular Resources Research Infrastructure). BBMRI was one of the first projects entering the European Research Infrastructure preparatory phase of the ESFRI roadmap funded by the European Commission (EC). The preparatory phase of BBMRI came to its end in January 2011. Over the past 3 years, BBMRI has grown into a 53-member consortium with more than 280 associated organizations (largely biobanks) from over 30 countries, making it one of the largest research infrastructure projects in Europe. One of the outcomes was a minimal dataset for biobanking, which is the background for this article. The MIABIS presented in this study aims at standardizing data elements used to describe a biobank's content on a meta-data and aggregate level to facilitate data

Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden.

discovery and is not intended to standardize data on an individual sample or subject level.

Materials and Methods

The MIABIS, like the initial BBMRI minimum data set, aims to provide an easy way to present which data elements are considered common for all biobanks; hence the name *minimum*. The initial data set consists of 21 attributes describing three levels: a biobank level, a study level and an object level (*i.e.,* individual subjects/cases/samples).[9] When looking at implementing the BBMRI minimum data set within the Swedish node BBMRI.se, the data set was further evaluated and led to the updated version presented in this article.

The main methods for creating the updated version (*i.e.,* MIABIS) of the BBMRI minimum data set included the following:

i) The concepts of "biobank" and "study" were separated into two distinct levels, as one study can be conducted with samples from multiple biobanks and, additionally, one biobank can host biological samples collected within multiple (unrelated) studies.
ii) Since a sample collection can exist without being for a particular study, the "study" level was renamed to "sample collection" level. Renaming this attribute will have a broader application and extend the use of the data set.
iii) Additional attributes that were considered required for describing biobanks and sample collection were included in the data set. This was done after consulting experienced researchers, biobank experts, and The National Board of Health and Welfare in Sweden.

Additional methods included:

iv) Implementing the data set in a pilot project where descriptive data about research projects using biobank samples was gathered in a paper-based form. The data set was structured in a survey with free text answers limited to the widest extent possible. Thirty Principal Investigators and biobank experts answered the questionnaire, which led to further improvements and validation of the data set.
v) Defining the attribute descriptions was made in accordance with epidemiological literature and terminology such as P^3G, The Public Population Project in Genomics, [http://www.p3g.org] and ISBER International Society for Biological and Environmental Repositories [http://www.isber.org/].
vi) Discussions with biobank IT experts regarding seeing the benefits of implementing the data set as part of their Laboratory Information Management Systems (LIMS) with positive response.

Results

One of the objectives for the BBMRI consortium during the preparatory phase was to develop a plan to integrate existing quality controlled biobanks, biomolecular resources, and enabling technologies into a novel pan-European biomedical research infrastructure. The MIABIS describes the minimum information facilitating the sharing of samples and data among biobanks on a global scale. The data set consists of a total of 52 attributes presented in Table 1. Multiple values are allowed for many of the attributes and their definitions are in some cases overlapping. The attributes for biobanks and sample collections/studies and their inter-relationship are visualized in Figure 1.

Current uses of the data set

National register of studies and other sample collections in Sweden. In conformity with the Catalog of European Biobanks [http://www.bbmri.eu/index.php?option=com_content&view=article&id=24&Itemid=26], developed within the pan-European BBMRI project, BBMRI.se is working on creating a National register with descriptive data about studies and other sample collections using human biological samples. The aim of the register is to raise awareness about sample accessibility for potential scientific collaborations.

The questionnaire used to gather information about the biobanks is based on the data set presented in this article. The questionnaire is currently only available in Swedish on the bbmri.se website [http://www.bbmri.se/sv/Forskningsprovsamlingar/Enkaten/]. The answers are also presented on the website, grouped by the studied diagnosis (ICD-10 code) for each sample collection [http://www.bbmri.se/sv/Forskningsprovsamlingar/Register/]. The register will be made searchable on any attribute.

BBMRI Wiki. The minimum data set is currently being implemented on a Wiki-platform by BBMRI.se, [http://bbmri-wiki.wikidot.com/]. The BBMRI Wiki is intended to help establish a standard vocabulary within the European BBMRI project and facilitate the updating and adding of terms. In excess of the data set, the Wiki also contains a detailed information model with guidelines on how to implement the data set in a relational database such as a LIMS. Within the Wiki, the present version of the BBMRI Biobank Lexicon has also been incorporated. To overcome the language barrier that comes with crossing borders when conducting research, the Biobank Lexicon has been translated into multiple languages.

Discussion

Limitations of the data set

The data set is limited in that it does not include data elements relevant to all medical areas or data on individual subjects and samples. Certain medical areas will require additional information on both a descriptive and individual sample level in order to be fully useful. As the planned research field determines the minimum data set recorded within the project or by the biobanks,[10] we encourage domain experts in each medical area to create "add ons" to the data set with additional attributes for their particular area of science or medicine. The data set does also not cover sample quality information more than on a collection level that for some analyses may be insufficient. Furthermore, the current data set is only intended to direct researchers to a source of samples and lacks information concerning the ethical standards under which the samples are collected, any restrictions on research use, and access requirements to the samples. As crucial as it is to know if samples are available and under what circumstances, the data on these aspects is difficult to harmonize as the legislation differs between countries and as the extradition of samples is always decided upon by each sample owner, communication between a researcher and the sample owner will be inevitable.

Table 1. Attribute Definitions for Version 2.0 of the BBMRI Minimum Data Set

	Attribute	Description
	Data Describing Biobanks	
1	Biobank ID	Textual string of letters starting with the country code (according to standard ISO1366 alpha2) followed by the underscore "_" and post-fixed by a biobank ID or name specified by its juristic person (nationally specific).
2	Name of biobank	Textual string of letters denoting the name of the biobank in the local language.
3	Juristic person	Textual string of letters denoting the juristic person (e.g., a university, concern, county council etc. for the biobank).
4	URL	Textual string of letters with the complete http-address for the biobank URL.
5	Country code	Textual string of letters of the two letter code for the country of the biobank according to ISO-standard 3166 alpha2.
6	Biobank type	Textual string of letters indicating the underlying (primary) medical area within which the samples are collected. Can be one or several of the following values: Pathology, Cytology, Gynecology, Obstetrics, Transfusion, Transplant, Clinical chemistry, IVF and similar, Bacteriology, Virology, Other. Definitions for the values are as follows: Pathology = Study of disease [1]; Cytology = Medical and scientific study of cells. Cytology refers to a branch of pathology [1], Gynecology = Branch of medicine particularly concerned with the health of the female organs of reproduction and diseases thereof [1]; Obstetrics = Art and science of managing pregnancy, labor, and the puerperium (the time after delivery) [1]; Transfusion = Transfer of blood or blood products from one person (the donor) into another person (the recipient's) bloodstream [1]; Clinical chemistry = Specialty of analytical chemistry applied to assays of physiologically important substances found in blood, urine, tissues, and other biological fluids for the purpose of aiding the physician in making a diagnosis or following therapy [2], IVF and similar = *In vitro* fertilization, a laboratory procedure in which sperm are placed with an unfertilized egg in a Petri dish to achieve fertilization. The embryo is then transferred into the uterus to begin a pregnancy or cryopreserved (frozen) for future use [1]; Bacteriology = Science and study of bacteria and their relation to medicine and to other areas such as agriculture (e.g., farm animals) and industry [1]; and Virology = Study of viruses[1]; Other.
7	Contact person	Textual string of letters denoting the name of the contact person for the biobank.
8	Contact phone	Textual string of letters denoting the telephone number to the contact person, including international call prefix.
9	Contact email	Textual string of letters denoting the email address of the contact person.
10	Contact department	Textual string of letters denoting the department, or corresponding (e.g., division), of affiliation of the contact person.
11	Contact address	Textual string of letters denoting the street name and street number or PO Box of the Contact person.
12	Contact ZIP	Textual string denoting the ZIP of the contact person.
13	Contact city	Textual string of letters denoting the city of the contact person.
14	Contact country	Textual string of letters of the two letter code in following the ISO-standard (3166 alpha2) format for the country of the contact person.
15	Hosted studies	Textual string of letters identifying the studies/sample collections that the biobank is physically hosting. Can be multiple values.
16	Date of entry	Date in ISO-standard (8601) time format when data about the biobank was reported into a database.
17	Last Updated	Date in ISO-standard (8601) time format when data about the biobank was last updated in a database.
	Data Describing Sample Collections	
	Attribute	*Definition*
18	Sample Collection/ Study ID*	Textual string depicting the unique ID or acronym for the sample collection or study.
19	Study name*	Textual string of letters denoting the name of the study in English.
20	Description	Textual string of letters describing the sample collection or study aim (max 200 characters).
21	Sample Collection Responsible/ Principal Investigator*	Textual string of letters denoting the name of the sample collection responsible or principal investigator.
22	Contact person	Textual string of letters denoting the name of the contact person for the sample collection or study.
23	Contact phone	Textual string for telephone number to the contact person, including international call prefix.
24	Contact email	Textual string for the email address of the contact person.

(*continued*)

Table 1. (Continued)

	Attribute	*Description*
25	Contact department	Textual string of letters denoting the department, or corresponding (e.g., division), of affiliation of the contact person.
26	Contact address	Textual string of letters denoting the street name and street number or PO Box of the contact person.
27	Contact ZIP	ZIP of the Contact person.
28	Contact city	Textual string of letters denoting the City of the contact person.
29	Contact country	The two letter code in format following ISO-standard (3166 alpha2) for the country of the contact person.
30	Type of Collection	Textual string of letters denoting the type of sample collection or study design. Can be one or several of the following values: Case-control, Cohort, Cross-sectional, Longitudinal, Twin-study, Quality control, Population-based, Other. Definitions for the values are as follows: Case-control = A case-control study design compares two groups of subjects: those with the disease or condition under study (cases) and a very similar group of subjects who do not have the disease or condition (controls) [3]; Cohort = A group of individuals identified by a common characteristic (e.g. demographic, exposures, illness etc.) [4]; Cross-sectional = A study in which participants are examined at only a single time for characteristics of a disease [3]; Longitudinal = Research studies involving repeated observations of the same entity over time. In the biobank context, longitudinal studies sample a group of people in a given time period, and study them at intervals by the acquisition and analyses of data and/or samples over time [4]; Twin-study = A twin study design is a study design in behavior genetics which aid the study of individual differences between genetically identical twins by highlighting the role of environmental and genetic causes on behavior [3]; Quality Control = A quality control testing study design type is where some aspect of the experiment is quality controlled for the purposes of quality assurance [3]; and Population-based = Multidisciplinary study done at the population level or among the population groups, generally to find the cause, incidence or spread of the disease or to see the response to the treatment, nutrition or environment [3]; Disease specific = A study or biobank for which material and information is collected from subjects that have already developed a particular disease [6]; Other
31	Collection start	Date in ISO-standard (8601) time format specifying when the sample collection starts
32	Collection end	Date in ISO-standard (8601) time format specifying when the sample collection ends, if applicable
33	Planned sampled individuals*	Number of individuals with biological samples planned for the study (also see Current sampled individuals)
34	Planned total individuals*	Number of individuals planned for the study (also see Current total individuals)
35	Sex	Textual string of letters denoting the sex of the sample donors. Can be one or both of the following values: Female, Male
36	Age interval	Age interval of youngest to oldest participant in sample collection
37	Average age	Average age of all sample donors in the sample collection
38	Main diagnosis	Textual string of letters for the ICD-10 codes for the studied diagnoses. Can be several values
39	Comorbidity	Textual string of letters indicating if information about co-morbidity is available. Can be Yes or No
40	Categories of data collected	Can be one or several of the following values: Biological samples, Register data, Survey data, Physiological measurements, Imaging data, Medical records, Other
41	Material type	Textual string of letters denoting the nature of the biological samples that make up the sample collection. Can be one or several of the following values: Whole blood, Plasma, Serum, Urine, Saliva, CSF, DNA, RNA, Tissue, Faeces, Other
42	Storage temperature	Textual string of letters with the temperature for the long-term storage of the biospecimens in the sample collection. Can be one or several of the following values: Room temperature, +4 °C, −18 °C to −35 °C, −60 °C to −85 °C, Liquid nitrogen, Other. The intervals are chosen according to SPREC [8].
43	Survey data	Textual string of letters covering additional information existing about the sample donors. Can be one or several of the following values: Individual Disease History, Individual History of Injuries, Medication, Perception of Health, Women's Health, Reproductive History, Familial Disease History, Life Habits/Behaviors, Sociodemographic Characteristics, Socioeconomic Characteristics, Physical Environment, Mental Health, Other [7].
44	Medical records	Free text specifying which medical record data is available in the sample collection/study
45	Registers	Free text specifying which registry data is available in the sample collection/study

(*continued*)

Table 1. (Continued)

	Attribute	*Description*
46	Omics experiments*	Textual string of letters denoting the -omics experiment(s) that have been performed on the samples in the sample collection . Can be one or several of the following values: Genomics, Transcriptomics, Proteomics, Metabolomics, Other. Definitions for the values are as follows: Genomics = The study of an organism's entire genome; Transcriptomics = The study of the transcription, i.e., the expression levels of mRNAs in a given organism, tissue, etc. (under a specific set of conditions). Proteomics = The study of proteins, their structures, and their functions, namely the study of the proteome; and Metabolomics = The identification, quantification, and characterization of the small molecule metabolites in the metabolome (i.e., the set of all small molecule metabolites found in a specific cell, organ, or organism) [5]; Other
47	Sample handling	Textual string of letters describing how the samples in the sample collection have been handled as an indication of sample quality. Can be one or several of the following values: Freeze chain, indicating if the samples in the collection have been kept cool from needle to freezer. Freeze time, time in hours from needle to freezer. SPREC compliant, if the samples are labeled according to SPREC, Other
48	Current sampled individuals	Number of individuals with biological samples in the study at the date of Last updated (also see Planned sampled individuals)
49	Current total individuals	Total number of individuals in the study at the date of Last updated (also see Planned total individuals)
50	Hosting biobank	Textual string of letters of the biobank/s storing the biological samples that are part of the sample collection. Can be several
51	Date of entry	Date in ISO-standard (8601) time format when data about the sample collection was reported into a database
52	Last updated	Date in ISO-standard (8601) time format when data about the sample collection was last updated in a database

[1] MedicineNet.com; [2] MedConditions.net; [3] EMBL (EFO); [4] P3G; [5] Methods in Molecular Biology 593; [6] George Davey Smith, Lyle Palmer, Paul R. Burton, An Introduction to Genetic Epidemiology; [7] http://www.datashaper.org/Datashaper.html#dataschemasTab$intro (Accessed Nov 2, 2011); [8] Betsou F. Standard Preanalytical Coding for Biospecimens: Defining the Sample PREanalytical Code. *Cancer Epidemiol Biomarkers Prev* 2010;19:1004–1011.

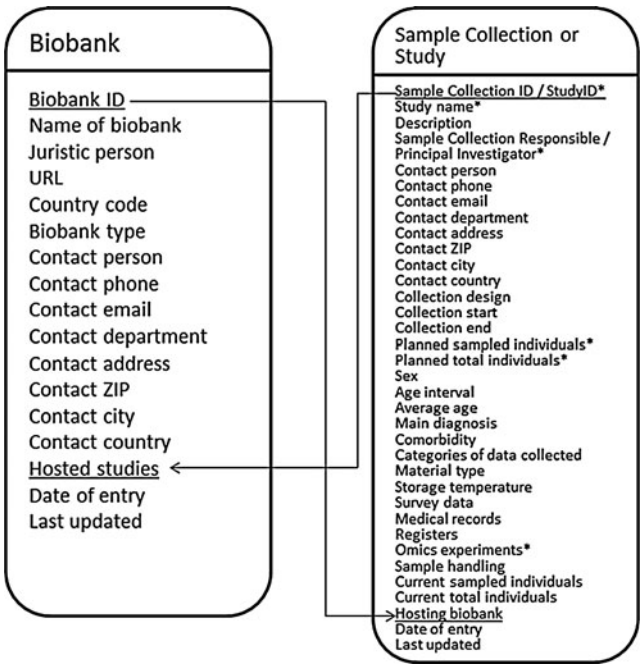

FIG. 1. The relationship between the types of attributes. *highlights attributes relevant for studies only.

Conclusions

A standard for sharing the key biobank attributes is the first logical step towards open and stimulating research collaborations. By its novelty, researchers will meet in new constellations enriching their research. In the 'omics' era, collaboration is not only the essence of good research, but also crucial for performing large scientific studies. Through collaboration, it is possible to get more value out of data that is already collected and, in many cases, underutilized. A meta-data model for biobanks and sample collections is a first step towards data sharing. A common set of attributes facilitates data sharing and helps researchers find attractive samples and potential collaborative partners. Furthermore, we believe the uses of a national register of studies and other sample collections as described in this article are many; containing enough information to help decide whether the samples are relevant for use without disclosing sensitive information about the sample donors. With this approach, underutilized sample collections can be made visible and fully utilized. Sharing samples and data also increases the transparency of science allowing quality and results to be tested and verified.

Using a standard for sharing aggregate data about biobanks and sample collections in a global and frequently revised register would pave the way for biobank networking by stimulating research collaborations and optimizing usage of biobank samples and correlated data. It would save time

and costs as samples are reutilized, shortening the time for the generation of new scientific results.

Acknowledgments

We would like to thank the people involved in the European BBMRI preparatory phase, financially supported by the European Commission (grant agreement 212111) and the Swedish Research Council for granting the BBMRI.se project (grant agreement 829-2009-6285). We would also like to thank Assistant Professor Mathias Brochhausen at the University of Arkansas for Medical Sciences (UAMS), Arkansas, U.S. for valuable feed-back.

Author Disclosure Statement

No competing financial interests exist.

References

1. Yuille M, van Ommen GJ, Brechot C, et al. Biobanking for Europe. Brief Bioinform 2008;9:14–24.
2. Sampogna C, Organisation for Economic Co-operation and Development., Organisation for Economic Co-operation and Development. Directorate for Science Technology and Industry. *Creation and governance of human genetic research databases*. (OECD, Paris; 2006).
3. Pukkala E, Andersen A, Berglund G, et al. Nordic biological specimen banks as basis for studies of cancer causes and control—More than 2 million sample donors, 25 million person years and 100,000 prospective cancers. Acta Oncol 2007;46: 286–307.
4. Communication from the Commission Official Journal (2002).
5. Taylor CF, Field D, Sansone SA, et al. Promoting coherent minimum reporting guidelines for biological and biomedical investigations: The MIBBI project. Nat Biotechnol 2008;26:889–896.
6. Fortier I, Burton PR, Robson PJ, et al. Quality, quantity and harmony: The DataSHaPER approach to integrating data across bioclinical studies. Int J Epidemiol 2010;39:1383–1393.
7. Thorisson GA, Muilu J, Brookes AJ Genotype-phenotype databases: Challenges and solutions for the post-genomic era. Nat Rev Genet 2009;10:9–18.
8. Zika E, Paci D, Braun A, et al. A European survey on biobanks: Trends and issues. Public Health Genom 2011;14:96–103.
9. Biobanking and Biomolecular Resources Research Infrastructure (BBMRI) Annex 16: Final Report from WP5 (WP5) (Medizinische Universitaet Graz, Graz; 2011).
10. Riegman PH, Morente MM, Betsou F, et al. Biobanking for better healthcare. Mol Oncol 2008;2:213–222.

Address correspondence to:
Loreana Norlin, M.Sc.
Department of Medical Epidemiology and Biostatistics
Karolinska Institutet
171 77 Stockholm
Sweden

E-mail: loreana.norlin@ki.se