BIOLOGICAL DATA STORAGE, ACCESS AND SHARING POLICY OF INDIA

सत्यमेव जयते

Department of Biotechnology
Ministry of Science & Technology
Government of India

.... 2019

BIOLOGICAL DATA STORAGE, ACCESS AND SHARING POLICY OF INDIA

INDEX

BIOLOGICAL DATA STORAGE, ACCESS AND SHARING POLICY OF INDIA

1. Purpose of this Policy Document

To define guidelines for sharing of data generated by scientists in India using modern biotechnological tools and methods.

Although, this document provides broad guidelines for biological data in general, it specifically pertains to modern high-throughput, high-volume data, for example, data generated by nucleic acid sequencing and microarrays, biomolecular structures and flow cytometry.

2. Introduction and Background

2.1 *The necessity of data-sharing and related issues*

The Government of India invests a large amount of money to generate data in various sectors, including the biotechnology sector. These data are generated in the context of furthering knowledge, gaining deeper insights into biological and other processes and for translation. Ultimately, all data are to be used for the benefit of humankind; that is the major reason for investment of public funds to generate data. Sharing of data maximizes the collective utility of data.

However, there are many issues that must be taken into account in the context of data-sharing, the most important of which is that data-sharing must be done in a responsible manner. Data may induce vulnerability to individuals and to populations. The rights to privacy and confidentiality of individuals and populations must be protected as emphasized in the U.N. Declaration of Human Rights, and no harm must be done to them as a result of data-sharing.

This document provides a framework and principles for sharing of data while protecting the rights of individuals and populations and without causing any harm to them.

2.2 *Ethical imperatives of data-sharing*

Data generated from public funds are for public good. Data are, therefore, a resource for human development. Unless the data are shared publicly and within a reasonable period of time after data-generation, the utility of the data will be constrained. Resultantly, accrual of benefit of public investment for data generation will be compromised. The necessity of data-sharing is, therefore, to accrue maximal benefit from public investment in generation of data.

Since data are a public resource, there are primarily three stakeholders of this resource – funders who help generate this resource, producers and users of this resource. All three stakeholders must assume responsibility on how the data may be shared. Even though data are a public resource, sharing of data have ethical implications. Data pertain to individuals and contain private information. When such data are shared, there may be breach of privacy. Therefore, sharing of data must be done in a responsible manner. Modalities in which data are shared must protect privacy, confidentiality, security and should be non-discriminatory and fair. These issues have been emphasized in the National Ethical Guidelines for Biomedical and Health Research involving Human Participants, 2017, established by the Indian Council of Medical Research (ICMR); it is expected that these guidelines are followed in research involving humans.

Responsible data-sharing implies that certain principles are to be followed. These include,

- *Protection of privacy and confidentiality*: Shared data must not include any personal identifiers and must have been collected with informed consent, including consent to share data after adequate anonymization/de-identification. Re-identification after anonymization/de-identification must not be attempted, without legal orders. In addition, care must be taken to ensure that the data resource is not used to ostracize communities; ethnic, religious, geographical or any other. Appropriate ethical approval(s) need to be obtained by the data-submitter prior to data submission.
- *Data quality, storage and security*: The quality of the data must be of a high standard, unbiased and verifiable. The data submitter is responsible for ensuring high quality and authenticity of submitted data. The storage must be done in a manner that protects privacy and confidentiality, and promotes ease of access, search and long-term maintenance. Appropriate security features must be embedded in the storage and access framework to avoid breach of data-trust. Features to enable tracing of chain of data access may be built-in. The storage and security policy must also address duration of data storage and accessibility. Mechanisms for obtaining feedback from resource users must be put in place in order to improve data quality, data access, data integrity and interoperability.
- *Transparency of policy*: Data-sharing policy must be transparent and must state in a publicly-accessible manner the policy of data transfer within and across national boundaries, with public and private organizations, for knowledge and commercial use, etc.
- *Public engagement and complaints*: Citizens must be engaged in the development of data-sharing policies and modalities. The engagement must result in improvement of future policy. There should also be a formal mechanism to register complaints of data misuse and to handle such complaints.

2.3 *Harmonization with international policies, ensuring that national policies supersede*

Many international consultations have been held to establish and evolve norms, rules and regulations for data sharing. These include the Bermuda Agreement, 1996; Fort Lauderdale Agreement, 2003; Nagoya Protocol, 2011, guidelines of the Global Alliance for Genomics and Health (GA4GH), 2013 and the European Union's General Data Protection Regulation, 2016. This current data-sharing policy respects and upholds the principles and tenets of the international discourses and agreements. These discourses have emphasized, in the main, the principle that data should be rapidly released after generation. The present data-sharing policy also strongly supports this principle. An Indian National Data-Sharing and

Accessibility Policy was promulgated in 2012. While the national data-sharing policies are generally harmonized with the international policies, it is emphasized that in a circumstance in which there is a misalignment of national and international policies, national policies shall supersede.

2.4 *Data-sharing: Open access, Managed access, No access*

Data resource sharing can be of three types.

- *Open access*: These data are placed in the public domain and are shared without any restriction imposed by the data provider.
- *Managed access*: The responsibility of protecting privacy and confidentiality of data providers is supreme. A data donor may provide limited consent for sharing her/his data with others. Under these circumstances, unrestricted access to data may not be provided. For such data, sharing must be managed. One common way to provide managed access is to ask the data user why she/he wishes to access the data and how the accessed data will be used. If there are restrictions placed on the use of data, these restrictions must be made known publicly and adhered to.
- *No access*: Access to some types of data may not be permitted. Data that are sensitive from a personal standpoint (e.g., data on illnesses), or from a national security standpoint may not be shared at all. Such data may not be made publicly accessible.

3. Types of Data

3.1 *Research Data and Public Resource Data*: Individuals engaged in research on scientific or social problems generate data that are of interest to them. These data comprise a resource, but may not be of immediate use to others. Even so, the data must be shared in a timely manner, especially if the data were generated using public funds. On the other hand, agencies such as Department of Biotechnology or Indian Council of Medical Research or Indian Council of Agriculture Research generate data that are not meant to answer an individual researcher’s questions but to become a public resource. Such resource data include cancer registry data, genome sequence data on members of various ethnic groups, crops/core collections of crops, animals, etc. Public resource data must be shared rapidly after generation and curation.

3.2 *Broad and Major Types of Data*: It is almost impossible to define all biological data-types that are generated by modern biotechnological methods. Data-types, particularly those that are high-throughput, also change with changing technologies. However, currently data types can be classified into some broad and major classes. These include, but are not to be considered as exhaustive,

3.2.1 *DNA sequence data* – Such data can be at the level of a whole genome, or single genes. Such data can be a single sequence (such as, sequence data generated by a Sanger sequencer) or multiple fragmented sequences from a genomic region with a high depth of coverage (such as those generated by a massively-parallel DNA sequencer).

3.2.2 *RNA sequence transcriptomic data* – The nature of the data are similar to those generated by a massively-parallel DNA sequencer, since usually cDNA synthesis is performed before sequencing. However, recent technological developments allow single-molecule direct RNA sequencing without cDNA synthesis.

3.2.3 *Genotype data* – Modern methods use high-density microarrays to genotype individuals at a large number of loci spread across the entire genome. Genotyping by sequencing (GBS) is being increasingly used for genome wide association studies especially in plants. However, for various specific purposes, small-scale genotyping using PCR-RFLP and other similar technologies continue to be used.

3.2.4 *Epigenomic data* – These data are also primarily generated using a BeadChip (that is similar to a DNA microarray). However, epigenomic data may also be generated using sequencing methods after bisulfite conversion.

3.2.5 *Microbiome data* – These data are also nucleic acid sequence data and currently are of two types (a)Amplicon sequencing data from which specific groups of microrganisms present in any sample (e.g., human stool, soil, sediment, etc.) can be identified, or (b) Shotgun metagenomic sequence data that allows comprehensive assessment of all microbial organisms present in a sample.

3.2.5 *Protein Structure data* – Atomic coordinates and other information that describes a protein and other important biological macromolecules comprise such data. These data provide 3D shapes of proteins, nucleic acids, and complex assemblies that help understand various aspects of protein synthesis under different conditions.

3.2.6 *Mass Spectrometry data* – Mass spectrometry (MS) is a key analytical technology in current proteomics and mass spectrometers are widely used to generate data that allow protein identification, annotation of secondary modifications, and determination of the absolute or relative abundance of individual proteins.

3.2.7 *Flow Cytometry data* – Flow cytometry is a technique used to detect and measure physical and chemical characteristics of a population of cells or particles. Flow cytometry data pertain to counts and multi-parameter profiles of different types of cells in a heterogeneous fluid mixture.

3.2.8 Imaging data – Images of individual cells, organs or body parts, for example, chest X-rays or images of human eyes or mouth cavity.
3.2.9 Metabolome data- Metabolomics is increasingly used in conjunction with microbiome data to better understand host-microbiome interaction. Small molecule metabolite patterns are generated using either LC MS, GC MS or CE MS.

Examples of common high-throughput data-types are provided in Appendix-1.

4. Framework for Data Sharing and Access

(a) Data generated from publicly-funded projects should be shared openly for public good, with few restrictions and in a timely manner, safeguarding the ethical issues that may arise out of shared data.

(b) High standards and best practices should be used in generation, management and access to data. Data that are valuable in the long-term should be stored in a manner that these remain accessible for a long time.

(c) Shared data will always be de-identified

(d) Under specific circumstances, even data generated using public funds may not be provided open access, and may be provided under a managed/controlled access protocol.

(e) To enhance use of data, metadata must also be released in a timely manner.

(f) Access to data that are of "sensitive" nature may be barred, even if generated using public funds.

(g) The conduct of research must not be jeopardized by release of data. The research organization must ensure that due consideration is given to protect the interest of the data generator.

(h) Data generator may require privileged use of the data. Therefore, there may be a period of moratorium before the data generator releases the data in the public domain. The period of moratorium may vary with the nature of the data; public resource data need to be released without any significant time lag.

5. Data Release Strategy and Timing of Data Release

5.1 *Data Release and Timing*

In current research and other scientific activities, large volumes of data are generated. These data comprise *raw data* that are produced by the various equipment that are used, e.g. DNA sequencer, Flow cytometer, etc. The raw data are then processed and analysed by researchers to draw scientific inferences. When public funds are used to generate data, these data must be made accessible to others in a form that is valid and user-friendly. It is recognized that raw data can be shared almost immediately after it is generated, but data-processing may take time and hence processed data may not become immediately amenable to sharing. Further, often there is no unique method of data-processing; methodological development may also be a part of data-processing.

It is recommended that – when funds provided by any agency of the Government of India to generate data, either wholly or partially – data, after appropriate clean-up and curation, be shared in accord with the following guidelines:

5.1.1 Raw (Level-1) data must be shared, by placement on a database identified and approved by the funding agency of the Government of India, *within one year* of generation of the data. If no such database is identified by the agency, then raw data must be made available to anyone working in any Indian institution, public or private, requesting for these data. The sharing of raw data must also include the experimental conditions and specifications of the equipment used to generate these data (experimental metadata), where relevant.

Sometimes an agency of the Government of India funds activities that are solely devoted to data generation, usually for the purpose of generating a "reference" data set. When data from such a project are generated, these data must be released *within six months* of data-generation.

5.1.2 Processed (Level-2) data based on data generated wholly or partially with funding from Government of India must also be shared. In recognition of the facts that (a) processing of data takes time, and (b) the research group that was funded to generate data and draw inferences from the data must be accorded the first right to publish the findings, it is recommended that *processed data may be shared with others within two years of data-generation*.

5.2 *Release of Metadata*

Use of certain types of data even if made publicly available may be of limited use unless some associated metadata are also made available. Such metadata include gender, ethnic background, phenotype, etc. Metadata should be released concurrently with other types of data (e.g., DNA sequence data) in order that the value of the released data is not diminished.

5.3 *Data Deposition*

In addition to releasing data, it is the responsibility of the data-generator to deposit data in an appropriate database in a National Biological Data Centre, as identified by the Department of Biotechnology. Until a National Biological Data Centre is established, raw and processed data must be stored on an institutional data storage facility. The data should be made available to anyone working in any Indian institution, public or private, who may request access to these data. Along with sharing of processed data, details of relevant methods used for processing raw data must also be shared.

5.4 *Exemptions to Data Release ("Sensitive" data)*

Release of data that compromises or impacts on national security shall be exempted. There may also be other circumstances when data-release exemption may be granted. There are some tribal populations in India that are numerically small. Whole genome sequence data released on individuals from such populations, with or without metadata, can result in individual identification. Therefore, if an exemption to release of such data is requested, the request may be considered and granted.

5.5 *Withdrawal of Data*

An individual donor whose data have been placed on a publicly accessible database may request for withdrawal of data, even if the donor had provided consent initially. Such requests may be considered and granted provided that the data are identifiable in the database.

6. Data User Agreement

Most data stripped of all personal identifiers and data that are not subjected to any intellectual property or patent restrictions should be made accessible openly (*open access data)*, especially if the data are generated using public funds. Sometimes, data generated in even in publicly-funded projects may not be allowed open access for a variety of reasons, prominent among them being that the data provider may need to be recontacted or may belong to a vulnerable subgroup or intellectual property issues may be under consideration. Such data should still be made available to others under *managed access*, that is, data accessibility should be provided only if the data-requester provides sufficient justification to request access the data and the purpose of data-use is valuable and ethically appropriate.

An individual requesting data that are accessible may do so using a Data Request Form (Appendix-2).

A Data Usage Agreement Form must be signed by the data recipient using the form provided in Appendix-3.

6.1 *Open Access Data*

6.1.1 Acknowledging the data provider: If the data-provider is identified in the database, then it is expected that the data user will adequately acknowledge the data-provider in publications and such documents in which results generated from the data are announced.

6.1.2 Who shall hold intellectual property arising from the shared data?: The onus of arriving at this decision is on national regulatory authorities and to a large extent depends on prior intellectual property rights granted by regulatory authorities to others, notably the data-generator.

6.1.3 Will the data be shared with others?: Open access data are public and are shared with anyone interested in accessing the data.

6.1.4 Will efforts be made for re-identification of individuals who may have provided data for inclusion in the database?: Re-identification is prohibited for open access data.

6.1.5 Legal Issues

6.1.5.1 Who will be held liable for data misuse?: If legal provisions are invoked for data misuse, such decision will be made by the court of law.

Otherwise, national funding agencies or peer groups (e.g., national science academies) may consider the nature of misuse and provide suitable reprimands.

6.1.5.2 If there is a dispute, how to resolve?: If the dispute is escalated to a court of law, then national legal modalities will apply for resolution. Otherwise, national funding agencies or peer groups (e.g., national science academies) may provide a platform for negotiation and resolution.

6.1.5.3 Duration of data access: There is no upper limit to the duration of access to open-access data. However, technologies (e.g., data storage space) may be the determining factors to limit the duration of access.

6.2 *Managed Access Data*

6.2.1 Purpose of access: Description, Ethics approval: As mentioned above, managed data usually have ethical, intellectual property and similar issues attached to the data. Therefore, unless the need for data-access outweighs the burden that may arise from these issues, access to such data may not be provided. Modalities of management of access to such data are usually established by the institution responsible for data-generation. Therefore, the data-requester must apply for data-access by providing a detailed description to the data-management group for reasons to request access to the data, the possible uses to be made of the data and the ethical precautions to be followed during and after data access.

6.2.2 Competence of researchers requesting data access: The data-management group shall assess the competence of the researchers to responsibly use the data for the purposes described by the data-requester before access to the data is provided.

6.2.3 Authority designated to sign on behalf of the data user: Unless otherwise stated, normally the head of the institution to which the data-user belongs or a designated nominee shall sign applications and other documents pertaining to data access and use on behalf of the data-user.

6.2.4 List of users authorized to access the data: Access to managed-data shall normally be given to a one or a small number of users of an institution. The list of persons who plan to access and use the data shall be provided on the application for data-access. The data-management group shall examine the list of possible users and their levels of competence before providing approval to data-access.

6.2.5 Duration of data access: The duration may be variable depending on intent of use of the data. The duration for which access is requested must be specified in the application and the data-management group may examine the appropriateness of the duration for which data-access is requested before approval.

6.2.6 Whether renewal of data access may be sought?: Yes. A fresh application must be submitted to the data-management group, in which the past use of the data and the future intended use shall be clearly described and justified.

6.2.6 How will confidentiality and security of shared data be ensured?: The application for data-access must clearly describe the plan to uphold the confidentiality of the data and the security of the data to prevent access by unauthorized users.

7. Audit

For open-access data, there may be a national committee established by a consortium of national funding agencies to monitor access and use. For managed-access data, the institution that manages data shall be responsible for data-audit. The data-management group shall regularly seek reports from users who have been provided data-access. Report of any breach in data-access or data-usage for open access or managed access data shall be appropriately dealt with by the national committee or the data-management group. Penalty for proven breach of access/use shall minimally comprise barring of access to the database for an extended period of time.

8. Other

For publicly-funded projects, if data-generation is outsourced to any GoI or private agency, then a Data Delivery and Non-Retention Agreement must be signed by the agency to which data-generation is outsourced. This Agreement must clearly describe the nature of samples being sent to the agency, the data to be generated, the date of delivery of data. Further, the Non-Retention portion of the Agreement must state that (a) data more than what has been requested and paid for shall not be generated by the Agency, (b) quantities of samples remaining after data-generation shall be returned by the Agency to the P.I., and (c) no data shall be retained by the Agency after despatch to the P.I.

A model Data Delivery and Non-Retention Agreement form is provided in Appendix-4.

References: To be added
Abbreviations

APPENDIX-1
Examples of High-throughput Data Types

Data on imaging of cells, molecules, whole genome, whole exome, gene-panel, transcriptome, gene expression, chip-seq, methylome and other epigenome or epitranscriptome, metagenome, proteome, metabolome and any data with information on more than a single gene/protein/metabolite from microbes, plant or animal cells, cell lines, animal and human organs using multiple of platforms like next-generation sequencing, microarrays, mass-spectrometry and microscopy and bimolecular structures. Examples of some of the high-throughput data generated are data on genome-wide association study (GWAS), genome and exome sequencing, gene expression studies using RNA-seq or miRNA-seq, disease-specific gene panels, DNA microarrays, proteomics studies using mass spectrometry, imaging different types of cells in human body, understanding microbial population in human gut using metagenome sequencing and disease metabolite profiling using gas chromatography coupled with mass spectrometry (GC-MS) or liquid chromatography coupled with mass spectrometry (LC-MS).

Levels of data

Raw (Level 1) Data:

First level data converted from raw images. Examples of Level 1 data are FASTQ/CSFASTA/HDF5/SSF files, intensity (idat) files along with the manifest (bpm/bgx/TXT) file for Illumina microarrays, EXP and CEL files for Affymetrix arrays, dta/pkl/ms2/mgf files for protein mass spec data, spectrum data for metabolites, TIFF/JPG/PNG files for images, higher level coordinates for 3D structures for biological macromolecules.

Processed (Level 2) Data:

Raw (Level 1) data are curated, processed and analyzed to provide value-addition and to ease inferences. Examples of such data are BAM/CRAM/FAST5/ProBAM files for sequencing, nmrML files for metabolite profiling experiments, CHP file for Affymetrix microarrays, gtc files for Illumina microarrays.

Processing of raw or semi-processed data are done in a variety of ways. Examples of such higher-level processed data are VCF/BCF files for variants, TXT file with genes with analyzed expression values (FPKM/RPKM/TPM normalized BED files), BED/ProBED files for genomics and proteomics data respectively, SGA/GFF files for chip-seq experiments.

APPENDIX-2
Data Request Form

1. Full Name of the Research Project for which Data are Requested

2. Applicant Information

Applicant (person in charge of the research project)
Name
Telephone
Work Address
E-mail

Contact person (if other than applicant)
Name
Telephone
Work Address
E-mail

3. Brief Description of the Research Project
Aims of the study, study design, scientific value and significance (max 300 words)
Timetable of the study
Agency funding the study

4. Description of the Requested Data

5. Plan for Publishing the Results
Describe, how the results are planned to be published (e.g. scientific article, PhD thesis)

6. Signature (Principal Investigator)
Place and date
Signature
Name in Capital Letters

APPENDIX-3

Data Use Agreement

General Information

Principal Investigator (PI)

Name

Email

Phone

Primary Contact

Secondary Contact

Do the data contain information collected from human research subjects? ☐ Yes ☐ No

Have the data been collected with IEC approval and with informed consent? ☐ Yes ☐ No

Do the Data contain any identifiers or individually identifiable health information ☐ Yes ☐ No

How will the research to be conducted with the Data be funded?

Do you anticipate that any inventions or intellectual property will be developed from the use of the data? ☐ Yes ☐ No

If yes, by whom?

Will the data be used in conjunction with other research? ☐ Yes ☐ No

If yes, what research?

Do you anticipate receiving any Confidential Information as part of the data transfer? ☐ Yes ☐ No

By signing this Agreement, you agree that the data that you shall receive will not be transmitted to anyone else and that standard ethical practices will be followed during the usage of these data.

Signature

Name in Capital letters

Place and Date

APPENDIX – 4

Data Delivery and Non-Retention Agreement

1. Name and Address of the P.I.:

2. Name of the Project

3. Name and Address of the Agency

4. Nature of Samples being sent to Agency

5. Number of Samples Sent

6. Quantity of Each Sample Sent

7. Date of Despatch of Samples

8. Date Within which the Data will be delivered by Agency to the P.I.

9. Agreement: The Agency to which samples are sent for data generation agrees to: (a) Generate only the data that have been requested and paid for, i.e., additional data may not be generated; (b) Remaining quantities of samples shall not be retained by the Agency and shall be returned to the P.I., and (c) No data shall be retained by the Agency, i.e., after despatch of data by the Agency and after receipt of the data by the P.I., the Agency shall destroy all data generated for the P.I.