सत्यमेव जयते

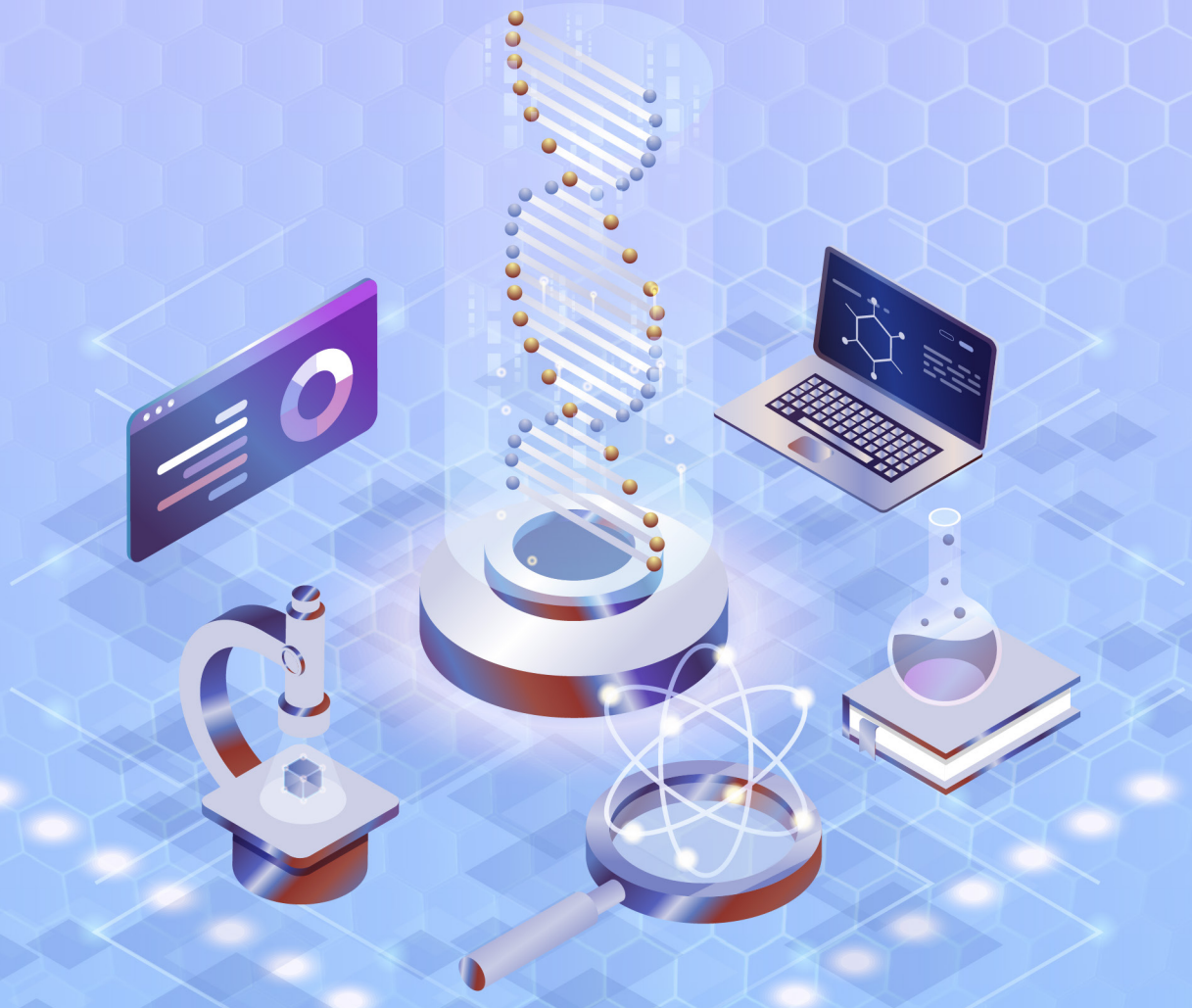
DEPARTMENT OF BIOTECHNOLOGY
Government of India

BIOTECH-PRIDE

(Promotion of Research and Innovation through Data Exchange)

GUIDELINES

July 2021



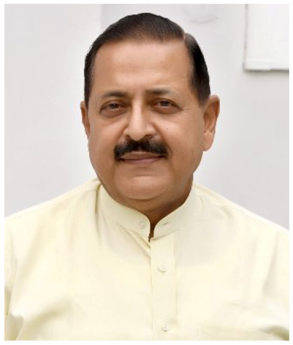
Published by:

Department of Biotechnology, New Delhi - 110003
July 2021

डॉ0 जितेन्द्र सिंह
राज्य मंत्री (स्वतंत्र प्रभार),
विज्ञान एवं प्रौधोगिकी मंत्रालय;
राज्य मंत्री (स्वतंत्र प्रभार) पृथ्वी विज्ञान मंत्रालय;
राज्य मंत्री, प्रधान मंत्री कार्यालय;
राज्य मंत्री कार्मिक, लोक शिकायत एवं पेंशन मंत्रालय;
राज्य मंत्री परमाणु ऊर्जा विभाग तथा
राज्य मंत्री अंतरिक्ष विभाग
भारत सरकार

सत्यमेव जयते

Dr. JITENDRA SINGH
Minister of State (Independent Charge)
of the Ministry of Science and Technology;
Minister of State (Independent Charge)
of the Ministry of Earth Sciences;
Minister of State in the Prime Minister's Office;
Minister of State in the Ministry of Personnel,
Public Grievances and Pensions;
Minister of State in the Department of Atomic Energy and
Minister of State in the Department of Space
Government of India



Message

The Government of India invests a large amount of public funds to generate data in various sectors, including in the biosciences for knowledge generation, to gain deep insights into intricate biological mechanism and other processes and for translation. The data generated from all sources within the country should be shared publicly and within a reasonable period of time of data-generation for effective utility and to accrue maximal benefit from investments in generation of data. Hence, it is also imperative that India puts the right framework to store, manage, archive and distribute all biological data generated in the country.

With this in view, and in conformity to the principles of NDSAP 2012, the Department of Biotechnology, Ministry of Science and Technology in consultation with all the relevant stakeholders have brought out the Biotech Promotion of Research and Innovation through Data Exchange (PRIDE) Guidelines. These Biotech- PRIDE GUIDELINES will enable storage, access and sharing of biological data in general and is specifically applicable to high-throughput, high-volume data generated in the country.

I congratulate all the members of the team who have brought out this document.

(DR. JITENDRA SINGH)

Anusandhan Bhawan, 2, Rafi Marg
New Delhi-110001
Tel. : 011-23316766, 23714230,
Fax. : 011-23316745

South Block, New Delhi-110011
Tel. : 011-23010191 Fax : 011-23017931
North Block, New Delhi-110001
Tel. : 011-23092475 Fax : 011-23092716

सचिव
भारत सरकार
विज्ञान और प्रौद्योगिकी मंत्रालय
जैव प्रौद्योगिकी विभाग
ब्लॉक-2, 7वां तल, सी0 जी0 ओ0 काम्पलेक्स
लोधी रोड, नई दिल्ली-110003
SECRETARY
GOVERNMENT OF INDIA
MINISTRY OF SCIENCE & TECHNOLOGY
DEPARTMENT OF BIOTECHNOLOGY
Block-2, 7th Floor, C.G.O. Complex
Lodhi Road, New Delhi-110003

सत्यमेव जयते

डॉ. रेणु स्वरूप
DR. RENU SWARUP

Foreword

A wealth of information, representing scientific disciplines in the healthcare, genomics, proteomics, metabolomics, microbiomes, protein structures, natural compounds, agriculture and population genetics, is being generated in India. Modern sciences have become data intensive and data dependent, often requiring integrative analysis. However, in the absence of any central data framework for biological data generated in the country, both data sharing and data dependent research are restricted. Accordingly, the 'Biotech-PRIDE Guidelines' have been formulated for sharing a wide range of large scale data so as to understand the molecular and biological processes that will contribute to human health on agriculture, animal husbandry, fundamental research and thus will extend to societal benefits.

I congratulate all the domain experts, representatives from the Industries, concerned ministries and Govt. organizations for their participatory role which has led to formulation of this document on Biotech Promotion of Research and Innovation through Data Exchange (Biotech-PRIDE) Guidelines. I am confident that Biotech-PRIDE guidelines will facilitate and enable exchange of information to promote research and innovation in different research groups across the country. Also this guideline document mentions about the Indian Biological Data Centre (IBDC) which is an initiative of Department of Biotechnology. This Centre will serve as the national repository for deposition, storage, quality control and annotation of biological data and will also help in consolidation of past, current, and future biotechnology data for variety of applications.

(Dr. Renu Swarup)

Tele : 24362950 / 24362881 Fax : 011-24360747 Email : secy.dbt@nic.in

TABLE OF CONTENTS

PREFACE

Keeping in view the emphasis of the Government on engaging citizens in Governance Reforms, placing of non-strategic data in public domain and the provisions of RTI Act 2005 for empowering the citizens to secure access to information under the control of public authority leading to the transparency and accountability in the working of every public authority, the National Data Sharing and Accessibility Policy (NDSAP) has been published in March 2012. Sharing of data generated is important not just to allow access to other research groups but also to develop a strong research response ecosystem by allowing access to data to researchers and students from universities and laboratories across the country. Currently the biological data which is generated is deposited in International Repositories; moreover there are no guidelines which mandate this. The value of data and knowledge will be enhanced multifold if this is shared. Globally all major funding agencies and governments have well defined data sharing policies/guidelines for biological data deposit and sharing.

Biotech - PRIDE (Biotech - Promotion of Research and Innovation through Data Exchange) Guidelines of India are to facilitate and enable sharing and exchange of biological knowledge, information and data generated through research conducted within the country and is specifically applicable to high-throughput, high-volume data like nucleic acid and protein sequences generated by instruments like next-generation sequencers, microarrays and mass spectrometry; biomolecular structures as determined by X-ray crystallography, Nuclear Magnetic Resonance (NMR), CryoEM etc.; images of whole body (like CT scans, PET scans, X-rays and MRI images), organs and cells; and flow cytometry data. Sharing of 'sensitive data' as defined in the document is not allowed under these guidelines.

These guidelines do not deal with generation of biological data per se. These guidelines create an enabling mechanism to share and exchange information and knowledge that is produced/generated/submitted by the Data Producer/ Generator/ Submitter following relevant extant laws, rules, regulations and guidelines of Government of India (GoI). These guidelines have been harmonized with relevant extant norms of GoI.
Since the technology platforms for acquiring data along with the nature of data are changing rapidly and the related policies by other agencies are under consideration for approval by the Government of India, it is anticipated that this guidelines document will be reviewed periodically and modified as appropriate.

ACKNOWLEDGEMENT

Data generated from one study can be utilized to explore other research questions and thus may result in amplifying the scientific value of data. With this in view, and in conformity to the principles of NDSAP 2012, the Biotech -PRIDE guidelines have been formulated through extensive Stakeholders and Inter-Ministerial Consultation for enabling the sharing, access and storage of biological data.

In the current scenario the biological data being generated is deposited in international repositories and there are no existing guidelines which mandate this. Hence, the formulation of Biotech Promotion of Research and Innovation through Data Exchange (Biotech-PRIDE) Guidelines has been an uphill task. The Department initiated steps towards formulation of the document on sharing and exchange of biological knowledge, information and data. A 'Zero Draft' was prepared by 'Working Group' comprising of Dr. Partha Majumder, NIBMG, Kalyani, Dr. Alok Bhattacharya, Ashoka University, Sonipat, Dr. Binay Panda, JNU, New Delhi, Dr. Ramesh Sonti, CCMB, Hyderabad and Dr. Yogesh Shouche, NCCS, Pune. We profusely thank the Working Group for taking first strides towards drafting of this document. The Advisory Committee chaired by Dr. G. Padmanaban, IISc., Bengaluru considered and discussed the draft prepared by Working Group and modified to the 'First Draft'. We are grateful to the Chair and members of the Advisory Committee for their kind suggestions and efforts throughout the journey for preparation of document.

The Draft was placed in the public domain for the comments from various stakeholders. Comments of concerned government agencies – DST; DSIR & CSIR; DARE & ICAR; DHR & ICMR; MoHFW; NBA; MoEFCC; MoES; Meity, MoH, Office of PSA and NITI Aayog were also sought. We sincerely acknowledge the constructive criticism and suggestions by each respondent which helped us to shape the document further.

The draft was modified, taking into cognizance the comments of public consultation and from concerned government agencies and all other stakeholders by the Inter-Ministerial Committee chaired by Dr. Renu Swarup, Secretary, DBT. We copiously thank the Chair and members of the Inter-Ministerial Committee for finalizing the document which has now been promulgated as Biotech-PRIDE Guidelines.

Last but not the least, untiring efforts and enthusiasm of the DBT team are to be put on record with high appreciations. The efforts of Dr. Shahaj Uddin Ahmed, Scientist 'E', Dr. A. Vamsi Krishna, Scientist 'E' and Dr. Onkar Nath Tiwari, Scientist 'E' in preparation of this document are commended. The efforts and zeal of Dr. Richi V Mahajan, Scientist 'C', are also appreciated. We further compliment the support extended by Mr. Ankit Agrawal and Ms Prachi Grover for designing of the document. Hard work of support staff in DBT is also acknowledged.

We express our sincere thanks and gratitude to everyone who has directly or indirectly contributed to this document.

Dr. Suchita Ninawe
Adviser, DBT

ABBREVIATIONS

BAM	Binary Alignment Map
BCF	Bit Clear File
BED	Browser Extensible Data
CEL	CIMFast Event Language file
CRAM	Challenge Response Authentication Mechanism
DBT	Department of Biotechnology
DHR	Department of Health Research
DNA	Deoxyribonucleic acid
dta	Defence Technology Security Administration
EXP	Protected mode executable program
GA4GH	Global Alliance for Genomics and Health
GBS	Genotyping by sequencing
GFF	General Feature Format
GoI	Government of India
IBDC	Indian Biological Data Centre
ICMR	Indian Council of Medical Research
MIMAG	Minimum Information on Metagenome Assembled Genomes
MIMARKS	Minimum Information on MARKer gene Sequence
MISAG	Minimum Information for Shotgun Assembled Genomes
NBA	National Board of Accreditation
NDSAP	National Data Sharing and Accessibility Policy
PCR-RFLP	Polymerase chain reaction-Restriction fragment length polymorphism
PRIDE	Promotion of Research and Innovation through Data Exchange
RNA	Ribo-Nucleic Acid
SGA	System Global Area
TPM	Trusted Platform Module
TXT	Text
VCF	Visual Component Framework

1. INTRODUCTION

1.1 The necessity of data exchange and sharing and related issues

Sharing of data is encouraged as emphasized through an Indian National Data Sharing and Accessibility Policy (NDSAP) promulgated in 2012. The Government of India invests a large amount of public funds to generate data in various sectors, including in the biosciences for knowledge generation, to gain deep insights into intricate biological mechanisms and other processes and for translation. Unless the data is exchanged and shared with other research groups within a reasonable period of time after data-generation, the utility of the knowledge generated data will be constrained. Resultantly, accrual of benefit of public investment for the knowledge generation will be compromised. The necessity of data-sharing and exchange, is, therefore, to accrue maximal benefit from public investment in generation of knowledge and data. In India there is no specific guidelines for storage access and sharing of Biological data. The Biotech PRIDE Guidelines will facilitate this and enable exchange of information to promote research and innovation in different research groups across the country. The Biotech PRIDE Guidelines are applicable for all biological data generated through research conducted within the country. Major data types are given in section 3 of this document.

Maximal accrual of benefit from data is possible if there is a national repository (hub) of data that stores all biological data in its servers linked to various other centres/ datasets (spokes). The national repository will act as a facilitator responsible for access and sharing of Biological Knowledge, Information and Data. This document provides a framework and principles for exchange and sharing of data generated from all biological resources while protecting the rights of individuals and populations and without causing any harm to them. It also ensures benefit sharing arising due to the sharing of data generated from accessing biodiversity of the country. The data provider/ generator/ submitter and users shall be bound by the relevant extant laws, rules, regulations and guidelines of GoI.

1.2 Imperatives of data exchange and sharing

There are primarily four stake-holders of this resource – funders who help generate the resource, producers, individuals whose samples are used for data generation, and users of the resource. All four stake-holders must assume responsibility on how the data are exchanged and shared. There are many issues that must be taken into account while sharing public resource data.

Data-sharing 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, and no harm must be done to individuals as a result of human- data sharing. It should also honour relevant National and International agreements accepted by GoI on protection of rights on biodiversity and benefit sharing. The guidelines do not allow sharing of 'sensitive data'.

It is mandatory that the Data Generator/Provider/Submitter shall follow the relevant extant laws, rules, regulations and guidelines of GoI for data generation using 'Human' and 'Non-human' samples.

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

- ***Data quality, storage and security:***

The quality of the data must be of a high standard, unbiased and verifiable. For the non-human data, it must be accompanied by minimum essential metadata as mentioned above. Metadata of the population must be available in case of mapping populations (such as recombinant inbred lines, F2 lines and others) along with phenotyping data. For genome-wide association studies (GWAS), the phenotypic data of the population under study must also be made available along with their metadata. The data submitter is responsible for ensuring high quality and authenticity of submitted data as per the defined standards for data ollection which are prescribed and regularly updated by the national repository. 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 guidelines 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.

- ***Protection of privacy and confidentiality:***
  Shared data generated from humans 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, unless and until there is any legal order to do so. In addition, care must be taken to ensure that the data resource is not used to discriminate communities - ethnic, religious, geographical or any other. Appropriate ethical approval(s) need to be obtained by the data-submitter prior to data generation using human samples.
- ***Transparency of guidelines:***
  Data-sharing guidelines must be transparent and must state in a publicly-accessible manner the guidelines of data transfer within and across national boundaries, with public and private organizations, for knowledge and commercial use.
- ***Public engagement and complaints:***
  Citizens should be engaged in the development of data-sharing policies and modalities. The engagement should result in improvement of future guidelines. There should also be a formal mechanism to register complaints of data misuse and to handle such complaints.

1.3 Harmonization with international policies, ensuring that national policies supersede

Many international consultations have been held to establish and evolve norms, rules and regulations for sharing of biological information and data. These include the Bermuda Agreement, 1996; Fort Lauderdale Agreement, 2003; Nagoya Protocol on Access and Benefit Sharing, 2010, Guidelines of the Global Alliance for Genomics and Health (GA4GH), 2013 and the European Union's General Data Protection Regulation, 2016. This present framework on Biotech Resource Information and Data sharing respects and upholds the principles and tenets of the international discourses and agreements emphasizing the principle that data should be rapidly released and shared after generation. This framework is also in harmonization with the International Agreements in place, however, in a circumstance in which there is a misalignment of national and international policies, national policies and guidelines shall supersede.

2. DEFINITIONS

'**Access**' means retrieval of information/ data by a user from a repository for research/ patient care/ commercial purpose.

'**Anonymization**' means the process of encrypting or removing all information that can identify an individual from the data.

'**Biological data**' means all information related to living organisms, including their nucleic acids, protein sequence, metabolites, and other molecular and functional characteristics.

'**Consent**' means expressed informed consent granted with full knowledge and understanding the nature, purpose and consequences of the collection, use, storage or disclosure of their data in any written/electronic/video form.

'**Custodian**' means the Data Centre that is responsible for storing the biological data, developing measures for safety, standards and quality for datasets and establishing detailed modalities for data access.

'**Data**' means a representation of information, numerical compilations and observations, documents, facts, maps, images, charts, tables and figures, concepts in digital and/or analog form.

'**Data repository**' or '**National Biological Knowledge, Information and Data Centre**' is a place for storage of all biological data and responsible sharing of the data with the user.

'**Data storage**' means retaining biological data in digital form on storage devices.

'**De-identification**' means the process of removing, obscuring, redacting or delinking all personally identifiable information from an individual's data in a manner that eliminates the risk of unintended disclosure of the identity of the individual.

'**Metadata**' means the information that describes the data source and the time, place, and conditions under which the data were created. Metadata informs the user of who, when, what, where, why, and how data were generated. Metadata allows the data to be traced to a known origin and know quality.

'**Processed data**' means the raw data that has been manipulated or modified to obtain certain information.

'**Producer or Generator or Submitter**' means those who are involved in generating and capturing data through automated and manual means.

'**Public funds**' means funds from any agency or autonomous body of the Government of India.

'**Public resource data**' means the data that has been generated to support development of research or any other public good.

'**Raw data**' means primary data collected from a source and not been modified or changed. For example, data coming out from a sequencing machine will be considered as raw data.

'**Reference data**' means data that are used for comparison and can be used as standards in further analysis. For example, a reference human genome is highly useful to find variants in analysis of patient genomic sequences.

'**Research data**' means data that is collected or created for purpose of analysis leading to understanding a scientific question.

'**Security**' refers directly to protection of data, and specifically to the means used to protect the privacy and use of data.

'**Sensitive data**' means data that are sensitive from a personal standpoint (e.g. racial or ethnic origin, some health or behaviour related data) or from a conservation viewpoint (metadata that reveals location of a rare and endangered species) or from a national security standpoint that is not publicly accessible and as defined in various Acts and rules of the Government of India.

'**Sharing**' means facilitating access and use by users.

'**Users**' means individuals or organizations who make use of the data by accessing data from a repository after obtaining necessary permissions, if required.

3. SOURCE AND TYPES OF DATA

3.1 Public Resource Data

Individuals engaged in research on scientific or social problems generate biological data that are of interest to them. These data comprise a resource, but may not be of immediate use to others. Even so, the biological data must be shared in a timely manner, especially if the data were generated using public funds. On the other hand, many government agencies generate biological data that are not meant to answer an individual researcher's questions but to become a public resource. Such resource data include disease 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 Major Data Types

It is almost impossible to define all biological data-types that are generated by the biotechnological methods. Data-types, particularly high-throughput data, also change with changing technologies. However, currently data can be classified into some broad and major types. 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, exomes, certain coding regions, DNA fragments 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 high-through put methods analogous to a DNA microarray or DNA sequencing after suitable pre-processing.

3.2.5 Microbiome data – These data are also nucleic acid sequence data and currently are of three major subtypes (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 and (c) genome sequences of individual isolates In addition, there is also data in the form of individual gene sequences used for Multi Locus Sequence Analysis and Multi Locus Sequence Typing, or for taxonomic purpose like 16S rRNA, gyrase and many other genes.

3.2.6 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.

3.2.7 Mass Spectrometry data – Mass spectrometry 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.

3.2.8 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 cells in a heterogeneous fluid mixture.

3.2.9 Imaging data – IImages of individual cells, organs or body parts, for example, chest X-rays or images of human eyes or mouth cavity.

3.2.10 Metabolome data – Metabolomics is increasingly used to understand metabolite levels either independently or in relation to gene expression. It is also used in conjunction with microbiome data to better understand host- microbiome interaction. Small molecule metabolite patterns are generated using either LC MS or GC MS or CE MS.

Data on phenotypes (e.g. plant height, insulin level, morphology, biochemical and chemotaxonomic characters for microbes) on individual samples also comprise biological data. Further, for the purpose of interpretation of results of analyses of biological data, environmental exposure data (e.g., dietary pattern, level of air-pollution) may also be collected in research. These data may also be considered essential to be shared, if collected. Examples of some common high-throughput data levels are provided below.

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.

Metadata: Examples of certain metadata include but not limited to gender, ethnic background, phenotype, demographic information etc. for human data; for non-human samples - Minimum Information on MARKer gene Sequence (MIMARKS), Minimum Information for Shot gun Assembled Genomes (MISAG) and Minimum Information on Metagenome Assembled Genomes (MIMAG) for metagenomics data. For non-human samples, the biological metadata must include passport data including family, genus, species, variety, accession number, geographic location from which isolated, etc.

4. DATA DEPOSIT STRATEGY AND TIMING

4.1 Data Deposit 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 deposited almost immediately after it is generated, but data-processing may take time and hence processed data may not become immediately amenable to deposit in the repository. Further, often there is no unique method of data-processing; methodological development may also be a part of data-processing.

It is the responsibility of the data-generator/producer/submitter to deposit data in an appropriate database in the notified Data Repository. However, depositing data in the national repository does not always mean that data will be available for sharing.

It is recommended that – when funds provided by any agency of the Government of India to generate data, either wholly or partially – such data after appropriate processing be shared in accordance with the following:

4.1.1 Raw (Level-1) data must be deposited, by placement on a database in the identified repository, within one year of generation of the data. Experimental conditions and specifications of the equipment used to generate these data (experimental metadata) shall also be deposited along with the raw data, 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 deposited within six months of data-generation.

4.1.2 Processed (Level-2) data based on data generated wholly or partially with funding from Government of India must also be deposited. 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 must be deposited with others within two years of data-generation.

4.2 Deposit 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, demographic information etc. Metadata must be deposited concurrently with other types of data (e.g., DNA sequence data) in order that the value of the released data is not diminished. Released metadata, alone or in combination with other data, must not enable identification of the individual to whom the data pertain. For non-human data, the passport data of the sample including place of collection must be provided so that appropriate communities can receive a share in the benefits that might accrue from any commercial applications that are derived by use of the data. It is extremely important that this information is made available at the time of data submission.

4.3 Exemptions to Data Deposition

Release of 'sensitive data' shall be exempted. Therefore, if an exemption to deposit of 'sensitive data' is requested, the request may be considered and granted

4.4 Withdrawal of Data

Data withdrawal may be granted if the individual or the organization, whose data have been placed on a publicly accessible database, make a justified request either directly or through the submitter, with valid claims to the data. Such requests may be considered and granted provided that the data are identifiable in the database.

5. FRAMEWORK FOR DATA SHARING AND ACCESS

(a) Initially, these Guidelines will be implemented through Indian Biological Data Centre (IBDC) at Regional Center for Biotechnology supported by Department of Biotechnology. Other datasets/ data centres will be bridged to the IBDC which will be called Bio-Grid. The Bio-Grid will be a National Repository for all biological knowledge, information and data generated through research within the country and will be responsible for enabling its exchange to facilitate the Research and Innovation, developing measures for safety, standards and quality for datasets and establishing detailed modalities for accessing data.

The expansion of IBDC will be considered after assessing the volume of biological data in the country.

(b) A Data Management Group under the guidance of Expert Advisory Committee will be responsible to put in place a responsive data sharing guidelines management and decision-making system in place. The implementation of the guidelines shall be overseen and administered by an Inter-ministerial National Steering Committee with Secretary DBT – Chair, Secretary DHR & Director General ICMR and Chairman NBA – Co-chairs, Representatives of all concerned government agencies (not below the rank of Joint Secretary) – Members, 4-5 Domain Experts. Ministry of Home Affairs shall be observer in the Inter-ministerial National Steering Committee to analyze datasets from the national security standpoint.

(c) The modalities for data sharing shall be managed by the IBDC under three categories as follows:

- **Open access:**
  Open access data are those which are intended to be shared openly by the data provider. All data, under 'open access' category, generated from public-funded research will be available to everyone (larger scientific community and public) under FAIR (findable, accessible, interoperable and reusable) principles. However, the user shall be bound by relevant extant national regulations, which he/she will have to agree to comply with at the time of downloading from the IBDC by electronically checking a box on the screen. Most data stripped of all personal identifiers and data that are not subjected to any intellectual property or patent restrictions should be covered under 'open access', especially if the data are generated using public funds. The data under open access category will also be available at www.data.gov.in

- **Managed access:**
  Managed access data are those which are shared with specific restrictions imposed by the data producer/generator/submitter. In case of data generated using public funds, restrictions to access and use of such data are to be established by the funding agency before its deposition. If there are restrictions placed on the use of data, these restrictions must be made known publicly and adhered to. In 'managed access,' prior to any data download attempt, a written proposal shall be provided by the data- requester intending to download such data to the IBDC regarding use and downstream sharing of the data. The proposal will be considered for approval by the IBDC. The data downloader, prior to actual download, will be required to sign a data use and sharing agreement form.

- **No access:**
  Access to 'sensitive data' shall not be permitted, even if generated using public funds.

The sharable data from IBDC shall be available on an "as-is" "where-is" basis. The access to a data set may be periodically reviewed by the Data Management Group and the nature of access to the data set may be altered.

(a) 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.

(b) 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.

(c) Data generator may require privileged use of the data. Therefore, the data generator may request exclusive access to the data and a reasonable period of moratorium to the IBDC before public release of the data. The period of moratorium may vary with the nature of the data and it is expected that data generated by public funds will be released without any significant time lag.

6. DATA USER AGREEMENT

The National Repository – IBDC will be responsible for framing detailed Data User Agreement as appropriate for the category and type of data, and will operationalize data sharing.

6.1 Open Access Data

The following terms shall be included in the Data User Agreement:

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

6.1.2 Intellectual property created from the shared data: The onus of arriving at a decision on who has the right to hold intellectual property created from shared data will be on the national regulatory authorities and to a large extent will depend on prior intellectual property rights granted by regulatory authorities to others, notably the data-generator/ producer.

6.1.3 Re-identification of individuals using shared human-data: Re-identification is prohibited and appropriate reprimands or legal provisions will be imposed by appropriate statutory or legal bodies.

6.2 Managed Access Data

In addition to the terms mentioned under 6.1, the following terms shall be included in the Data User Agreement:

6.2.1 Purpose of access: The data-requester must apply for access and use of 'managed access' data.

6.2.2 Competence of data-user requesting data access: The data provider shall assess the competence of the data-requester 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 individuals 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 managed 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 managed data-access is requested before approval.

6.2.6 Renewal of data access: 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. No access would be granted for unlimited time without renewals.

6.2.7 Confidentiality and security of shared data: 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.

6.3 Duration of Data Access

There is no upper limit to the duration of access to open-access or managed-access data and The duration may be variable depending on intent of use of the data. However, technologies (e.g., data storage space) may be the determining factors to limit the duration of access. It may also be noted that data categories are not static; managed-data at a point of time may be declared, by relevant statutory authorities, as open data and no-access data may be declared as managed-access 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.

7. AUDIT AND LEGAL ISSUES

Data will remain the property of agency/ Department/ Ministry/ entity which is responsible for its collection and reside in their IT enabled facility for sharing and providing access. Access to sharing of data under these guidelines will not be in violation of any act and rules of GOI in force. Report of any breach in data-access or data-usage shall be appropriately dealt within the Legal framework of NDSAP published by GOI vide Gazette notification No. DL(N)-04//0007/2003-05 dated 23rd March 2012

ANNEXURE-I

The Advisory Committee constituted by the Department of Biotechnology vide OM No.BT/Data Policy/2019 dated March 19, 2019 and reconstituted on July 9, 2019

1.	Dr. G. Padmanaban, IISc., Bengaluru	Chairman
2.	Dr. Partha Majumder, NIBMG, Kalyani	Co-chair
3.	Dr. Alok Bhattacharya, JNU, New Delhi	Co-chair
4.	Dr. Vijay Chandru, Strand Life Sciences, Bengaluru	Member
5.	Dr. Binay Panda, Ganit Labs, Bengaluru	Member
6.	Dr. P. Anandan, Wadwani Institute of Artificial Intelligence, Mumbai	Member
7.	Dr. Sanghamitra Bandopadhyay, ISI, Kolkata	Member
8.	Dr. Dinakar Salunke, ICGEB, New Delhi	Member
9.	Dr. N. K. Arora, The INCLEN Trust International, New Delhi	Member
10.	Dr. Ramesh Sonti, NIPGR, New Delhi	Member
11.	Dr. Gagandeep Kang, THSTI, New Delhi	Member
12.	Dr. Alok Srivastava, CMC, Vellore	Member
13.	Dr. Saurabh Raghuvanshi, DU, New Delhi	Member
14.	Dr. Yogesh Shouche, NCCS, Pune	Member
15.	Dr. Subeer Majumdar, NIAB, Hyderabad	Member
16.	Dr. Dinesh Gupta, ICGEB, New Delhi	Member
17.	Dr. Debasisa Mohanty, NII, New Delhi	Member
18.	Shri R. S. Mani, NIC, New Delhi	Member
19.	Dr. Rajender Joshi, C-DAC, Pune	Member
20.	Representative of DST • Dr. Neeraj Sharma, Scientist 'G' • Dr. Anita Aggarwal, Scientist 'F'	Member
21.	Representative of CSIR • Dr. Anurag Agarwal, IGIB, Delhi	Member

22. Representative of ICAR	Member
• Dr. Anil Rai, ADG, ICAR	
• Dr. N. K. Singh, NRCPB, New Delhi	
• Dr. Dinesh Kumar, IASRI-ICAR	
23. Representative of ICMR	Member
• Dr. Prashant Mathur, NCDIR-ICMR, Bengaluru	
• Dr. Sukanya R, NCDIR-ICMR, Bengaluru	
24. Representative of National Biodiversity Authority	Member
• Dr. Purvaja Ramachandran, Secretary, NBA	
• Dr. K. P. Raghuram, NBA	
• Dr. Prabha Nair, NBA	
25. Representative of MeitY	Member
• Shri S. Gopalakrishnan, MeitY, New Delhi	
• Shri Vikas Chourasia, MeitY, New Delhi	
26. Representative of BIRAC	Member
• Dr. Shirshendu Mukherjee, Mission Director	
• Dr. Madhavi Chandra, BIRAC	
27. Dr. Suchita Ninawe, Adviser/Scientist G, DBT	DBT Coordinator
28. Dr. Vamsi K. Addanki, Scientist 'E', DBT	Co-Member Secretary
29. Dr. Shahaj Uddin Ahmed, Scientist 'E', DBT	Co-Member Secretary
30. Dr. Onkar Nath Tiwari, Scientist 'E', DBT	Co-Member Secretary

ANNEXURE-II

The Inter-Ministerial Committee constituted by the Department of Biotechnology vide OM No.BT/Data Policy/2019 dated July 9, 2019

1.	Dr. Renu Swarup, Secretary, DBT	Chairperson
2.	Dr. G. Padmanaban, IISc., Bengaluru	Co-chair
3.	The Secretary DST/Representative	Member
	• Dr. Neeraj Sharma, Scientist 'G'	
	• Dr. Anita Aggarwal, Scientist 'F'	
4.	The Secretary, DSIR & DG, CSIR/Representative	Member
	• Dr. Anurag Agarwal, IGIB, Delhi	
5.	The Secretary, DARE & DG, ICAR/Representative	Member
	• Dr. Anil Rai, ADG, ICAR	
6.	The Secretary, DHR & DG, ICMR/Representative	Member
	• Dr. Prashant Mathur, NCDIR-ICMR, Bengaluru	
	• Dr. Sukanya R, NCDIR-ICMR, Bengaluru	
7.	The Secretary, M/o Health & FW/Representative	Member
	• Dr. Mandeep Bhandari, Joint Secretary	
	• Dr. Sanjeev Kumar, M/o Health & FW, New Delhi	
8.	The Chairman, National Biodiversity Authority/ Representative	Member
	• Dr. V.B Mathur, Chairman, NBA	
	• Dr. Purvaja Ramachandran, Secretary, NBA	
9.	The Secretary, M/o Environment, Forest & Climate Change/ Representative	Member
	• Dr. Sujata Arora, Scientist 'G'	

10. The Secretary, M/o Earth Sciences/Representative	Member
• Dr. Gopal Iyengar, Scientist 'G'	
• Dr. R.S. Mahesh Kumar, Scientist 'F'	
11. The Secretary, MeitY / Representative	Member
• Shri Gopalakrishnan S, MeitY , New Delhi	
12. Representative of CEO, NITI Aayog	Member
• Dr. Neeraj Sinha, Adviser (S&T), New Delhi	
13. Dr. Partha Majumder, NIBMG, Kalyani	Member
14. Dr. Alok Bhattacharya, Ashoka University, Sonepat	Member
15. Dr. Binay Panda, Ganit Labs, Bangaluru	Member
16. Dr. Ramesh Sonti, NIPGR, New Delhi	Member
17. Dr. Yogesh Shouche, NCCS, Pune	Member
18. Dr. Suchita Ninawe, Adviser/Scientist G, DBT	DBT Coordinator
19. Dr. Vamsi K. Addanki, Scientist 'E', DBT	Co-Member Secretary
20. Dr. Shahaj Uddin Ahmed, Scientist 'E', DBT	Co-Member Secretary
21. Dr. Onkar Nath Tiwari, Scientist 'E', DBT	Co-Member Secretary

DEPARTMENT OF BIOTECHNOLOGY
Ministry of Science & Technology
Government of India

dbtindia.gov.in **/dbtIndia** **@dbtIndia** **@dbtIndia**

Contact Information:

Dr. Shahaj Uddin Ahmed
Scientist 'E'
Department of Biotechnology
Ministry of Science and Technology
Government of India
E-mail: shahaj.ahmed@nic.in

Dr. Richi V Mahajan
Scientist 'C'
Department of Biotechnology
Ministry of Science and Technology
Government of India
E-mail: rv.mahajan@dbt.nic.in