



DEPARTMENT OF BIOTECHNOLOGY
MINISTRY OF SCIENCE AND TECHNOLOGY
GOVERNMENT OF INDIA



(FRAMEWORK FOR EXCHANGE OF DATA)

PROTOCOLS

FOR IMPLEMENTATION

OF BIOTECH-PRIDE

GUIDELINES

(PROMOTION OF RESEARCH AND INNOVATION THROUGH DATA EXCHANGE)

JANUARY 2025





DEPARTMENT OF BIOTECHNOLOGY
MINISTRY OF SCIENCE AND TECHNOLOGY
GOVERNMENT OF INDIA



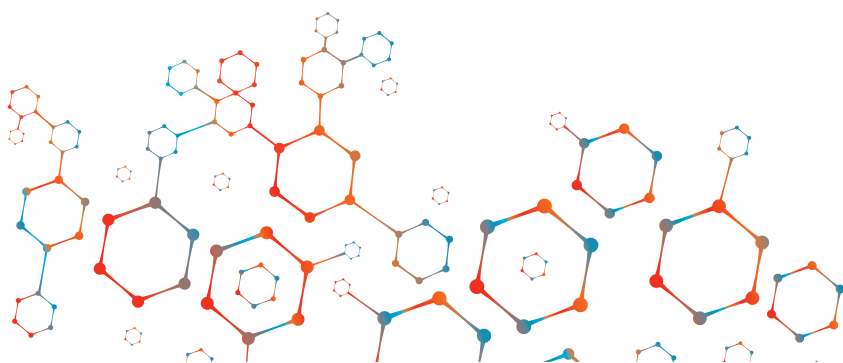
(FRAMEWORK FOR EXCHANGE OF DATA)

PROTOCOLS

FOR IMPLEMENTATION OF BIOTECH-PRIDE GUIDELINES

(PROMOTION OF RESEARCH AND INNOVATION THROUGH DATA EXCHANGE)

JANUARY 2025



डॉ० जितेन्द्र सिंह

राज्य मंत्री (स्वतंत्र प्रभार),
विज्ञान एवं प्रौद्योगिकी 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)

Ministry of Science and Technology;

Minister of State (Independent Charge)

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 high-quality biological data generated through various initiatives is vital for addressing pressing global challenges such as various health concerns including rise of infectious diseases, biodiversity loss, and climate change. Prompt sharing of this data empowers researchers with swift access to analyze critical information, accelerating discoveries and advancements in different fields of biological sciences. However, it is equally important to establish comprehensive protocols for data sharing to ensure judicious use, integrity, security, and reproducibility.

In this context, I commend the Department of Biotechnology for the timely publication of the procedural document, **“Framework for Exchange of Data (FeED) Protocols,”** developed under the realm of Biotech-PRIDE (Promotion of Research and Innovation through Data Exchange) Guidelines, 2021. This document provides a robust framework that will enable the research community to fully harness the potential of large-scale data initiatives like GenomeIndia and other projects supported by the Government of India. By fostering responsible data sharing, the FeED Protocols represent a significant step forward in advancing research and innovation, paving the way for a future driven by collaboration and impactful discoveries.

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

(Dr. Jitendra Singh)

MBBS (Stanley, Chennai)

MD Medicine, Fellowship (AIIMS, New Delhi)

MNAMS Diabetes & Endocrinology

FICP (Fellow, Indian College of Physicians)

Anusandhan Bhawan, 2, Rafi Marg,
New Delhi-110 001
Tel. : 011-23321681, 23714230,

Prithvi Bhawan, Lodhi Road,
Opp. India Habitat Centre,
New Delhi-110003
Tel. : 011-24629788, 24629789

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



डॉ. राजेश सु. गोखले
Dr. RAJESH S. GOKHALE



सचिव
भारत सरकार
विज्ञान और प्रौद्योगिकी मंत्रालय
जैव प्रौद्योगिकी विभाग
ब्लॉक-2, 7वां तल, सी.जी.ओ कॉम्प्लेक्स
लोधी रोड, नई दिल्ली-110003
SECRETARY
GOVERNMENT OF INDIA
MINISTRY OF SCIENCE & TECHNOLOGY
DEPARTMENT OF BIOTECHNOLOGY
Block-2, 7th Floor, CGO Complex
Lodhi Road, New Delhi-110003



FOREWORD

In today's rapidly advancing scientific landscape, the sharing of biological data has become an essential aspect of research and discovery. By pooling resources and collaborating on data sharing initiatives, researchers can accelerate the pace of scientific progress and drive innovation in the interdisciplinary fields of biology. By promoting transparency and collaborations among researchers, biological data sharing enables validation of findings, leading to more robust, reliable and reproducible scientific conclusions.

This cross-pollination of ideas and expertise is essential for tackling the multidimensional challenges of modern biology. However, at the same time proper data management and governance are critical for maintaining the integrity and reliability of shared datasets.

Earlier, this Department has issued Biotech-PRIDE Guidelines to assist in the dissemination of biological information and data being generated by research groups across the nation. Now the 'Framework for Exchange of Data (FeED) Protocols' for implementation of Biotech-PRIDE Guidelines has been shaped which sets the stage for fair and transparent data sharing in biological sciences. These protocols highlight the procedures laid down for responsible data sharing thereby facilitating interdisciplinary collaboration, and maximizing the utility of valuable datasets. The document also emphasizes the critical role of data submitter, data user, data keeper and funding agency in diligent data exchange; to ensure submission of quality datasets, compliance with regulatory requirements, and best practices for data sharing, practices on ethical standards, protection of participants privacy, and promote trust in scientific research.

I congratulate all the domain experts, the representatives of concerned government agencies and all other stakeholders for their active participatory role in formulating the 'FeED Protocols'.

(Dr. Rajesh S. Gokhale)





PROLOGUE:

Dr. Suchita Ninawe, Scientist - 'G' DBT



In the field of life sciences, the sharing of biological data plays a crucial role in advancing scientific knowledge and accelerating the pace of discovery. Unless data is exchanged and shared within a reasonable period after its generation, the utility of the knowledge derived from such data will be constrained. Consequently, the benefits of public investment in knowledge generation may be compromised. To address this challenge, this department released the Biotech-PRIDE Guidelines, underscoring the necessity of datasharing and exchange to maximize the benefits derived from public investments in knowledge and data generation; and stated a need to formulate a 'Framework for Data Sharing and Exchange'.

In the age of big data, protecting the confidentiality of sensitive information is paramount to maintaining public trust and compliance with data protection regulations. However, responsible data sharing of biological data is also essential for driving scientific research. Hence, as outlined in the Biotech-PRIDE Guidelines, the 'Framework for Exchange of Data (FeED) Protocols' has been formulated to establish a robust mechanism for data sharing and access.

The FeED Protocols define the imperative for researchers, institutions, and policymakers to prioritize responsible data sharing practices. They emphasize striking a balance between openness and data protection to ensure the integrity and sustainability of data-driven research endeavours. The protocols have been developed after extensive consultation within the 'Data Management Group (DMG)', followed by critical assessment by the 'Expert Advisory Committee (EAC)', and finalized through the rigorous discussions and recommendations by the 'Inter-ministerial National Steering Committee (IMC)'.

All the clauses stated in the Biotech PRIDE Guidelines and in this document - FeED Protocols are in compliance with other acts, rules, regulations, and national guidelines issued by the Government of India regarding data exchange, and are binding on all stakeholders. In cases of misalignment between national and international policies, national acts, rules, regulations, policies, and guidelines shall take precedence.

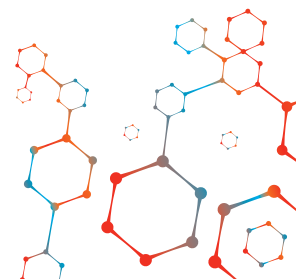
The FeED Protocols outline the roles and responsibilities of various stakeholders in the data-sharing ecosystem. These stakeholders include funders who support projects generating primary data, data producers who collect or generate primary/secondary data, individuals (in the case of human data) whose samples are used for data generation, users of the data, and (The document finalized by IMC) custodian i.e. Indian Biological Data Centre (IBDC). The role of DMG working under the guidance of EAC has been clearly stated to implement a responsible data sharing and decision-making system.

Moreover, the FeED Protocols provide details on data types, standardized data formats and metadata, presenting a significant impetus for effective data sharing. Standardized structures and protocols for organized datasets enable researchers to find, access, and utilize shared data efficiently, thereby enhancing the impact and utility of data sharing initiatives.

As envisioned in the Biotech-PRIDE Guidelines, this document on FeED Protocols will pave the way for responsible data sharing while addressing ethical considerations. By adhering to FeED Protocols, researchers can harness the power of shared data to unlock new discoveries and innovations in the field of biology. Furthermore, the integration of cutting-edge data science approaches into biological research will ensure a future-ready workforce, capable of leveraging digital tools to drive transformative advancements in science and healthcare.

The contributions by all the members of the Data DMG, the EAC, and the IMC are commendable. The FeED Protocols will be updated, as and when required, to bring more reforms in data exchange ecosystem.

New Delhi, January 2025







ACKNOWLEDGEMENT:

Data sharing is crucial for advancing scientific research and accelerating discoveries in various fields of biological sciences. Through data exchange, researchers can collaborate more effectively, avoid duplication of efforts, and gain access to a wider range of information to draw more robust conclusions. This promotes transparency, reproducibility, and accountability in research and enhances the overall quality and reliability of scientific findings. Realizing the significance of data-sharing, the Biotech-PRIDE (Promotion of Research and Innovation through Data Exchange) Guidelines were released by the Government in July 2021.

For implementation of Biotech-PRIDE Guidelines, 'Framework for Exchange of Data (FeED) Protocols' has been drafted through extensive Expert and Inter-Ministerial consultations. These protocols outline the procedures for deposition and sharing biological data, keeping in view data privacy, security, and proper attribution. By following these standardized Protocols, researchers can facilitate seamless sharing of biological data, promote scientific discovery, and accelerate progress in various fields of biosciences.

We extend our heartfelt gratitude to the Data Management Group (DMG) constituted by this Department comprising of 35 domain experts, co-chaired by Dr. B. Jayaram, IIT Delhi and Dr. Arvind Sahu, RCB Faridabad. The DMG formed a three member drafting committee involving Dr. Shandar Ahamd, JNU, Dr. Shantanu Sengupta, CSIR-IGIB, and Dr. Dinesh Gupta, ICGB, and tasked with preparing the initial draft. This initial draft was revised and refined by DMG during a series of meetings.

The Expert Advisory Committee (EAC) comprising of academicians, industries and private partners, and chaired by Dr. P Balaram, IISc Bangalore critically assessed the draft document prepared by the DMG and incorporated required revisions. We appreciate the valuable feedback and contribution by the EAC which was instrumental in making the draft ready for inter-ministerial consultations.

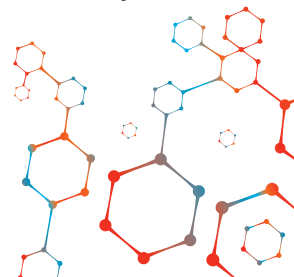
Under the ambit of Biotech PRIDE Guidelines, this department has constituted the Inter-ministerial National Steering Committee (IMC) chaired by Secretary DBT, co-chaired by Secretary DHR & DG ICMR and Chairperson National Biodiversity Authority (NBA) and comprised of representatives from DSIR & CSIR, DST, DARE & ICAR, MeitY, MoHFW, MoES, MoEFCC, Niti Aayog, the Office of the Principal Scientific Adviser, and MHA, and (The document finalized by IMC) domain experts. The draft submitted by EAC was subsequently presented for discussion by the IMC. We sincerely recognize the constructive critiques and suggestions of the EMC and its final recommendations on the document for its adoption. We place on records our gratitude to Dr. Rajesh S. Gokhale, Secretary DBT, Shri. C. Achalender Reddy, Chairperson NBA and all the members of the IMC for their guidance and contributions.

The invaluable contributions of Dr. Suchita Ninawe, Adviser at DBT, are deeply appreciated for her exceptional guidance and support throughout the drafting process. Her astute feedback has been instrumental in refining the document, and we are sincerely thankful for the time she devoted to reviewing and discussing the draft.

Furthermore, we wish to acknowledge the unwavering dedication and enthusiasm of Dr. Deepak Nair, RCB Faridabad and the IBDC team, whose efforts warrant special recognition. We also extend our gratitude to Dr. Deeksha, BTIC Apex Center and Mr. Deepak S. Verma, ProMentor Digital Solutions for their assistance in the preparation and design of the document. Additionally, we recognize the diligent work of the support staff at DBT. We express our heartfelt thanks to everyone who have contributed, directly and indirectly, to the development of this document.

New Delhi, January 2025

Dr. Richi V Mahajan





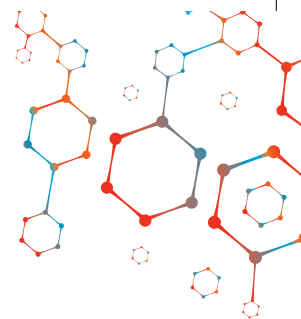
INDEX

Abbreviations

1. Introduction	3
2. Stakeholders	4
3. Data Types	4
4. Levels of Data	4
5. Data Access Levels	5
6. Timelines for Access to Data	5
7. Responsibilities of Data Producer/Submitter, Custodian (IBDC), User, and Funder	6
8. Role of Data Management Group (DMG)	8
9. Data quality check	9
10. Withdrawal of data	9
11. Intellectual Property and Legal Issues	9
12. SOPs and formats for data submission	10

Annexures:

i. Registration Portal format (Annexure-I)	12
ii. Pre-submission form (Annexure-II)	13
iii. Data User Agreement formats (Annexure-III)	14
iv. Nucleotide data (Annexure-IV)	17
v. Macromolecular Structure data (Annexure-V)	19
vi. Proteome data (Annexure-VI)	21
vii. Metabolome data (Annexure-VII)	22
viii. Biological Imaging data (Annexure-VIII)	24
ix. Phenome data (Annexure-IX)	26
x. Data Management Group, 2024 (Annexure-X)	30
xi. Expert Advisory Committee, 2024 (Annexure-XI)	32
xii. Inter-ministerial National Steering Committee, 2024 (Annexure-XII)	33
xiii. Flowchart for Data submission and Access (Appendix I to III)	34-36

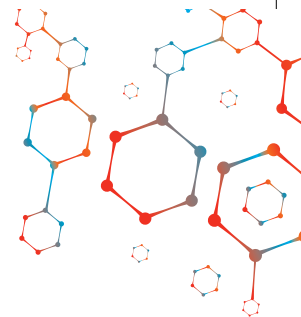


ABBREVIATIONS

BMRB	Biological Magnetic Resonance Data Bank
CryoEM	CryoElectron Microscopy
ddRADseq	Double digest restriction-site associated DNA followed by sequencing
DMG	Data Management Group
EMDB	Electron Microscopy Data Bank
FAIR	Findable, Accessible, Interoperable, and Re-usable
FeED	Framework for Exchange of Data
FSC	Fourier Shell Correlation
GBS	Genotyping by sequencing
IBDC	Indian Biological Data Center
ICPD	Indian Crop Phenome Database
IPD	Indian Proteome Databank
ISDA	Indian Structural Data Archive
MA	Managed Access
MIAME	Minimum Information About a Microarray Experiment
MIGS	Minimum Information About a Genome Sequence
MIMAG	Minimum Information on Metagenome Assembled Genomes
MIMARKS	Minimum Information on MARKerGene Sequence
MISAG	Minimum Information for Shotgun Assembled Genomes
MX	Macromolecular Crystallography
NMR	Nuclear Magnetic Resonance
PRIDE	Promotion of Research and Innovation for Data Exchange
SSRs	Simple Sequence Repeats
STRs	Short Tandem Repeats
wwPDB	Worldwide PDB







1 INTRODUCTION

The Government of India, through the Department of Biotechnology (DBT), has released Biotech-PRIDE (Promotion of Research and Innovation through Data Exchange) Guidelines in July 2021, (https://dbtindia.gov.in/sites/default/files/Biotech%20Pride%20Guidelines%20July%202021_0.pdf). The guidelines provide a well-defined framework and guiding principle to facilitate and enable sharing and exchange of biological knowledge, information and data and are specifically applicable to high-throughput, high-volume data generated by research groups across the country. These guidelines ensure data sharing benefits, viz. maximizing use, avoiding duplication, maximizing integration, ownership information, better decision-making and equity of access. These guidelines enable the sharing of data publicly and within a reasonable time after data generation to promote the maximal utility of the generated data. Resultantly, accrual of the benefit of public investment for data generation will not be compromised.

The Biotech-PRIDE Guidelines comprehensively identify the data types, the nature of the provision of access and sharing, and the resources required to manage data submission and sharing. The guidelines envisage a National Repository for biological knowledge, information, and data, which will be responsible for enabling its exchange, developing measures for safety, standards, and quality for datasets, and establishing detailed modalities for accessing data. Currently, these guidelines are being implemented through the National Repository named Indian Biological Data Centre (IBDC), established by DBT. The Biotech-PRIDE Guidelines in clause 5(b) under the section on 'FRAMEWORK FOR DATA SHARING AND ACCESS', envisaged the need for a 'Data Management Group (DMG)' to work under the guidance of an Expert Advisory Committee (EAC), which will be responsible for putting in place a responsive data sharing and management guidelines and decisionmaking system. This document aims to fulfill these objectives with inputs from various constituent data management subgroups, which recommended the specialized requirements and issues about different data types.

This document has been prepared, conforming to the Biotech-PRIDE Guidelines, 2021 and in harmonization with the International Agreements in place. This document aims to address the modalities and decision-making processes involved during the data submission and sharing to harmonize, synergize and encourage the data sharing for research and analysis in the country as per the FAIR principles. This will promote scientific work and foster progress by building on previous work.

All the clauses stated in the Biotech-PRIDE Guidelines and this document, as well as other acts, rules, regulations and national guidelines and policies issued by Govt. of India on data exchange, shall be binding on all stakeholders and in a circumstance where there is a misalignment of national and international policies, National acts, rules, regulations policies and guidelines shall prevail.

This document, named 'FeED Protocols for Implementation of Biotech-PRIDE Guidelines' is endorsed by the Inter-ministerial National Steering Committee in its meeting held on November 12th 2024.

This document may be revised from time to time as per the requirements of existing datasets, and national and international norms.





2 STAKEHOLDERS

(REFERENCE TO BIOTECH-PRIDE GUIDELINES, PAGE 2-3, SUBSECTION 1.2)

The Biotech-PRIDE Guidelines have identified primarily four main stakeholders of biological databases. Considering the management perspective of the data, the fifth stakeholder, the Data Custodian, i.e., IBDC, has been added. Thus, the five stakeholders are as follows:

- I. Funders, who fund projects generating the primary data**
- II. Producers, who collect or generate primary/secondary data**
- III. Individuals (in the case of humans) whose samples are used for data generation**
- IV. Users of the data, and**
- V. Custodian (IBDC)**

3 DATA TYPES

(REFERENCE TO BIOTECH-PRIDE GUIDELINES, PAGE. 6-7, SUBSECTION 3.2)

From the perspective of the biological domain, the Biotech-PRIDE Guidelines have identified 10 different types of data, e.g., sequencing data, epigenetic data, etc. After careful deliberation, biological data has been regrouped into six data types in this document as follows:

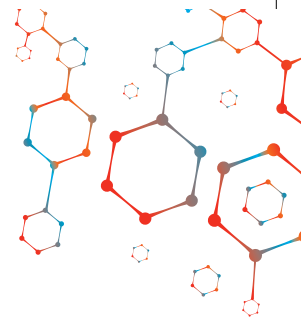
- a) Nucleotide(including sequencing and other gene expression/all DNA and RNAbased) data**
- b) Macromolecular Structure data**
- c) Proteome data**
- d) Metabolome data**
- e) Biological Imaging data**
- f) Phenome data**

4 LEVELS OF DATA

(REFERENCE TO BIOTECH-PRIDE GUIDELINES, PAGE 7, OF SECTION 3)

The Biotech-PRIDE Guidelines identify three levels/ components of data as follows:

- a) Raw data (level 1)**
- b) Processed data (level 2)**
- c) Meta-data**



5 DATA ACCESS LEVELS

(REFERENCE TO BIOTECH-PRIDE GUIDELINES, PAGE 10, OF SECTION 5(C))

Each of these data may be characterized as having three access levels:

- a) Open access (available to all)**
- b) Managed access (permissions needed)**
- c) No access (sensitive* data to be stored by IBDC but not to be made accessible)**

**Sensitive data will be all such data as defined in Biotech PRIDE Guidelines, 2021*

Data will be proposed to fall into one of the three access levels by the submitter, and the DMG will review these access levels as per the Biotech-PRIDE (The document finalized by IMC) Guidelines. The decisions of DMG will be reviewed by EAC on a periodic basis. Submission of any data to IBDC under 'No access level' will be reported by DMG/EAC to the Inter-ministerial National Steering Committee.

6 TIMELINES FOR ACCESS TO DATA

Timely data sharing is important to scientific progress. The Data under the 'Managed Access' level will be shared no later than 12 months of data submission or as soon as the main findings from the submitted study dataset are published, whichever is earlier. If findings are yet unpublished within the 12 months of data submission, the submitter may submit a request for extension of embargo time for data under 'Managed Access' along with the revised time-plan for data sharing, to IBDC for consideration by DMG.





7 RESPONSIBILITIES OF DATA PRODUCER/SUBMITTER, CUSTODIAN (IBDC), USER, AND FUNDER

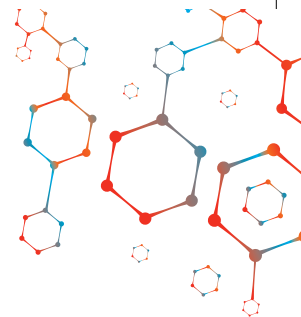
I. Data Producer/Submitter will ensure that:

- a) The relevant extant laws, rules, regulations and guidelines of GoI for data generation using 'Human' and 'Non-human' samples are followed and all the required clearances for collecting and sharing the data have been obtained.
- b) The data generated from humans have been sufficiently anonymized/deidentified. However, associated metadata, that must not enable identification of the individual to whom the data pertain, are also made available.
- c) For non-human data, the passport data of the sample is also provided to facilitate appropriate benefit sharing, if a situation arises.
- d) Sufficient quality controls have been implemented, and where applicable, quality metrics have been provided for each data object or the entire batch of data.
- e) The data is submitted to the relevant IBDC portal, following the login authentication and provisioning necessary steps, including the data submission forms as provided by IBDC from time to time. The type of data and level of data are chosen appropriately.
- f) The required permissions from their parent organization for submission of data are obtained.

The Data submitters are encouraged to submit the data under the open access portal for timely sharing of data except wherein the data is deemed to be characterized as 'Managed Access' or "No Access".

II. The Custodian (IBDC) will:

- a) Provide services for smooth data submission and sharing. Facilitate the data upload and maintenance.
- b) Address any issues of the data submitter arising in the various checks being implemented.
- c) Verify the pre-submission forms for Data types, volume, access level, and channel according to respective portals. IBDC will also encourage interoperability of the data submitted with national / international repositories. In cases where IBDC does not agree to the proposed data type or data access levels, the case may be referred to DMG.



- d) Ensure safety and security of the data to the extent possible. Ensure that no individual identifying information is received, as defined under the sensitive data in Biotech PRIDE Guidelines. Security audits should be carried out as per the MeitY guidelines
- e) Facilitate sharing of the data with users as per the Biotech-PRIDE Guidelines and FeED protocols only and in accordance with the appropriate ethical guidelines as applicable. And applicable rules and regulations existing at the time of data sharing.
- f) Responsible for acquiring any administrative and legal approvals, as and when required.
- g) Conduct the meetings of DMG, Expert Advisory Committee (EAC) and Interministerial Committee as listed in Biotech-PRIDE Guidelines from time to time, as and when required.
- h) Execute/communicate the decisions of DMG and/or EAC in a timely manner.
- i) Maintain all the datasets submitted to the IBDC, permanently accessible as independent and/or as a part of the scientific publication.
- j) Allow updating or corrections of errors in the datasets submitted to IBDC, with the approval of a DMG. The updated version of the data will replace the original dataset, while retaining the original submission.
- k) Maintain a time-stamped record of the changes in the dataset associated with a unique Accession ID.

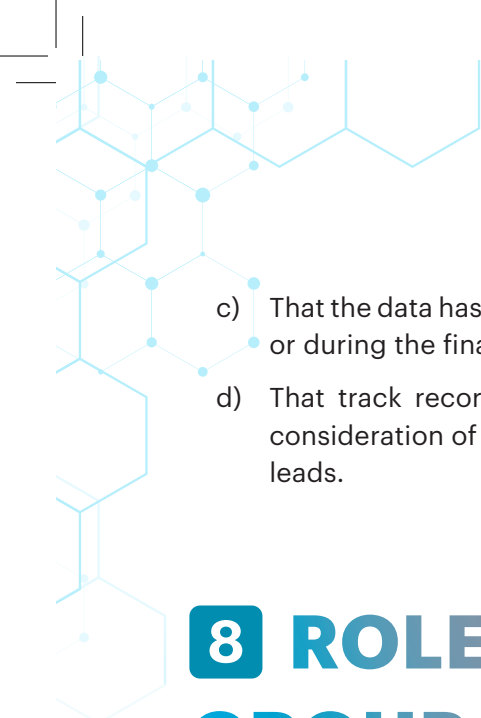
III. The Users/Data requester will ensure

- a) Safe custody of the data accessed from IBDC and no further transmission (private or public) of the data. Secondary data generated from the accessed data should follow the legislations/ laws/ acts/ policies/ guidelines laid down by Govt. of India from time to time.
- b) That for human data, re-identifying the individual whose data has been accessed should not be attempted.
- c) That original data generators are cited in derivative publications and other products.
- d) That if any request is received by the approved Data User from any third party seeking access to the data in his/her custody, such request must be directed to the IBDC by the User in order to ensure compliance.
- e) Use of appropriate agreed safeguards to prevent the disclosure of the data information and shall report to the IBDC about any violation of the agreement of which they become aware.
- f) To inform any addendum to the approved application or a new request from the same Requester/User for a fresh review by the DMG.

IV. The Funders will ensure

- a) That a data sharing plan is submitted along with the original proposal by the Project Investigator, before funding.
- b) That the required clearances for any proposed data collection, generation and sharing have been obtained from the concerned authorities.

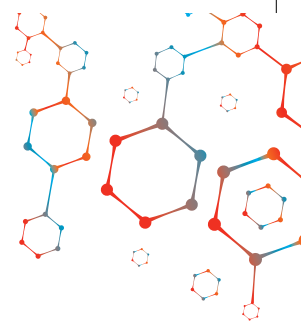


- 
- c) That the data has been submitted to IBDC, before the release of subsequent grants in the same proposal or during the final settlement of accounts.
 - d) That track record of previous data submission is made as one of the key parameters during the consideration of the new proposal submitted by the Project Investigator based on the earlier research leads.

8 **ROLE OF DATA MANAGEMENT GROUP (DMG)**

- a) The DMG will respond to all data access requests for “Managed Access Data” preferably within 45 days.
- b) The DMG will ascertain that the requester is ‘appropriately qualified/bona fide’ for use of the data responsibly, based on his/her request/proposal and initiated/forwarded by the authorized personal as per the institutional norms (to be verified by IBDC).
- c) The requester’s acceptance of the terms and conditions as per FeED protocols for data access is followed by a decision on the data sharing by the DMG. This should be done on a case-by-case basis and shall rely on the merits of the data access request.
- d) Preferably, the DMG should allow the release of managed data only to those with an institutional email address. This reassures the IBDC, research participants, and the general public that an appropriate individual is accessing and using the data.
- e) The DMG will also review the access levels for Data Submission from time to time as per the Biotech-PRIDE Guidelines.
- f) The DMG will consider the request of data submitters for an extension of embargo time for Managed Access.
- g) The DMG will mentor the Data Portals implemented at IBDC.
- h) The DMG will apprise the EAC and/or Inter-ministerial Committee on the activities mentioned in points above.
- i) The DMG will evolve FeED protocols for new data sets if not covered under the Data Types mentioned at Point 3 in this document, as and when received in IBDC.

The decisions made by DMG will be revisited annually by Expert Advisory Committee wherein, two independent experts may be called as nominated by Department of Biotechnology.



9 DATA QUALITY CHECK

Each upload dataset is validated during the submission process. Automated scripts verify the uploaded files for correct metadata format, unique accession list (duplicate entries will be rejected), and missing or blank data. Upon successful upload, valid IBDC accessions will be assigned to each data file. Submission of duplicate files and these submissions may be suppressed without warning. To update an existing record, do not resubmit the data files, instead, contact the helpdesk for assistance.

10 WITHDRAWAL OF DATA

Once deposited, data can be withdrawn after release per the provisions mentioned in Biotech-PRIDE Guidelines, 2021. The cases of data withdrawal requests will be taken up by the DMG. If any time after submission, the correctness, integrity, ownership, or provenance of a dataset is called into question, the Custodian - IBDC possesses the right to make it obsolete or remove the controversial dataset altogether or its controversial parts from the archive. However, this action is subject to the approval of the DMG. For example, if a publication describing the IBDC dataset is retracted, the retractor can request IBDC to remove corresponding data/entries from the IBDC records. However, DMG will make the final decision regarding the retraction of the submitted data.

11 INTELLECTUAL PROPERTY AND LEGAL ISSUES

Issues of intellectual property, re-identification of individuals, legal issues and other issues are elaborated in the Biotech-PRIDE Guidelines. The servers and software of IBDC shall be audited periodically by the MeitY empanelled vendor.





12 SOPs AND FORMATS FOR DATA SUBMISSION:

Data can be submitted by any researcher or organization supported in part or fully by public funds, private entities and other National/ International organizations as per the provisions of data submission mentioned in this document.

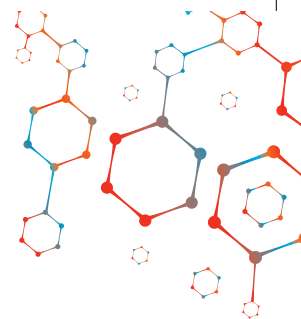
I. Common data agreement forms

- a) As a first step, the Data Submitter/ or Data User are required to register with IBDC for any kind of operation (Annexure I).
- b) Data Submitters will fill out a pre-submission form as per the template given in the Annexure II.
- c) A data-sharing agreement as per the template given in the Annexure III will be signed by the Submitter before data submission, and by the User before downloading/ accessing the data.

Annexures I, II and III are to be used across all the data types.

II. Use of separate templates depending on the Type of Data

Different data will be uploaded to the IBDC portals. Annexures IV, V, VI, VII, VIII and IX provide the data submission templates for, Nucleotide data, Macromolecular Structure data, Proteome data, Metabolome data, Biological Imaging data and Phenome Data, respectively.



A SUMMARY OF THESE TEMPLATES FOR QUICK REFERENCE IS PROVIDED BELOW:

Data type	Submission templates	Data formats*	Remarks
Nucleotide data (DNA Sequencing, RNA sequencing, Genotype, Epigenetic)	Annexure IV	FASTA, FASTQ, SAM/BAM/CRAM, BED/GTF, and bedgraph	Open and managed access; Raw and processed; Level 1 and Level 2 data; Human, pathogen, and plant data
Structure data (Macromolecular structure data and associated experimental data.)	Annexure V	PDB, CIF, MTZ, NEF, NMR-STAR, EM Map	Open access
Proteome data	Annexure VI	RAW, .D, .WIFF, .LCD, .JMS, .mzML, .mzXML, .mzData, T2D	Open and managed access
Metabolome data	Annexure VII	abf, cdf, cdf.cmp, cmp, d, dat, hr, ibd, jpf, lcd, mgf, qgd, raw, scan.wiff, xp, T2D	Open and managed access
Biological Imaging data (Radiological, Whole slide, MRI, Microscopy images, Videos)	Annexure VIII	Dicom (images), Text (metadata), JPEG, TIFF, PNG.	Open and managed access
Phenome data	Annexure IX	CSV, TSV, TIFF, JPEG, TXT, or SDF, xlsx, etc	Open and managed access



REGISTRATION PORTAL

First name:

Middle name (optional):

Last name:

Email (preferably Institutional):

Mobile (required for OTP):

Organization name/Affiliation:

Designation:

ORCID:

Address:

City:

State:

Country:

Username:

Password:

Password confirmation:

PRE-SUBMISSION FORM

(To be filled by Data Submitter subsequent to completion of the registration process)

GENERAL INFORMATION

1. Principal Investigator (PI)

Name:

Email:

Phone:

Primary Contact:

2. Source of funding:

3. Type of Data:(Imaging/Nucleotide/Proteome/Metabolome/Structure/Phenome/Others):

4. File Format:

5. Volume of Data (File Size):

6. Proposed Access Level: Open Access/ Managed Access/No Access

7. Ethical clearance obtained (if required):

8. Data Sharing Plan:

9. Justification for Proposed Data Access Level:

10. Approximate time period for making data available at Open Access level, in case it is proposed at Managed Access:

☐

I undertake that I have read the Biotech-PRIDE Guidelines and the FeED Protocols. I assure that I will abide by the clauses therein.

(PDF documents will be available here and the Submitter will have to open the documents and only then he will be able to check the box and only then this agreement submission will be completed.)

DATA USER AGREEMENT

(To be filled by Data User subsequent to completion of the registration process)

1. Common guidelines for data access

The sharable data from IBDC shall be available on an “as-is” and “where-is” basis under two Access Levels as per the Biotech-PRIDE Guidelines. The users will sign separate agreements/ request forms for the two types of access.

I. For use of data at Open Access Level

Name:

Designation and affiliation with complete address:

Email ID:

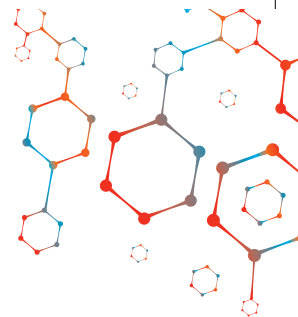
Mobile number:

1. I am aware that the shareable data from IBDC is available on an “as-is” and “where-is” basis. I agree to maintain safe custody of the data accessed from IBDC
2. That further transmission of the data shall not be done.
3. That secondary data generated from the accessed data shall be reported back to IBDC.
4. That re-identification of the individual whose data has been collected shall not be attempted.
5. Those original data generators shall be cited in derivative publications and other products.

☐

I undertake that I have read the Biotech-PRIDE Guidelines and the FeED Protocols. I assure that I will abide by the clauses therein.

(PDF documents will be available here and the data requester will have to open the documents and only then he/she will be able to tick the checkbox and only then this agreement submission should be completed.)



II. For use of data at Managed Access Level

A. DETAILS OF THE APPLICANT

Name:

Designation and affiliation with complete address:

Email ID:

Mobile number:

B. DETAILS OF THE REQUESTED DATA

1. The IBDC Project ID/Study ID/File IDs for the dataset to be downloaded.
2. Data files required:
3. How will other researchers benefit from sharing this dataset with you?

C. PROPOSED PROJECT DETAILS

1. Project Title, Project ID and Funding agency:
2. Keywords that best summarize your proposed research project(upto 6):
3. Description of the research question(250words):
4. Background and current status of the research(500 words):
5. Description of the methodology to be adopted(500 words):
6. Expected outcomes(250 words):
7. Do the data contain information collected from human research subjects?
8. Social impact of the proposed work (250 words):





DATA USER AGREEMENT

(To be filled by Data User subsequent to completion of the registration process)

9. Will the research project generate any new data fields derived from existing datasets? (300 words)
10. Estimated duration of your project, in months(maximum being 3 years):
11. Expected quantitative outcomes (publications, patents, technology transfer, etc.):
12. Status of IEC approval of the Project (attach a copy):
13. How will the research conducted with the data be funded?
14. Will the data be used in conjunction with other research? If yes, what research?
15. Do you anticipate receiving any Confidential Information as part of the data transfer?
16. Declaration Certificate as follows, forwarded by the executive Authority of the organization (to be uploaded)
17. I am aware that the shareable data from IBOC is available on an" as-is" and where-is" basis.
18. I agree to maintain safe custody of the data accessed from IBDC
19. That further transmission of the data shall not be done.
20. Secondary data generated from the accessed data should follow the norms laid down by Govt. of India from time to time
21. That re-identification of the individual whose data has been collected shall not be attempted.
22. That original data generators and IBDC accession numbers shall be cited in derivative publications and other products.

☐

**I undertake that I have read the Biotech-PRIDE Guidelines and the FeED Protocols.
I assure that I will abide by the clauses therein.**

(PDF documents will be available here and the data requester will have to open the documents and only then he will be able to check the box and only then this agreement submission should be completed.)



TEMPLATE/ CHECKLIST FOR UPLOADING NUCLEOTIDE DATA

PRE-SUBMISSION REQUIREMENTS:

- i. Up to 1,000 samples can be submitted in a single submission. For more than 1,000 samples, split your data into multiple submissions referencing the same Project/Study.
- ii. File compression with gzip or bzip2 is recommended, and tarballs are accepted. Avoid using zip format.
- iii. Studies exceeding 2 TB of data require splitting into submissions under 2 TB each. Upload them one after another, linking them to the same Project/Study for unified search ability.
- iv. For submissions exceeding 2 TB of data, please contact IBDC support before uploading. IBDC support can advise you on the best approach to handling large data submissions.
- v. Duplicate file submissions are not accepted; these submissions might be deleted accordingly.

A. SUB-TYPES OF DATA

The nucleotide data at this point pertains to all raw and processed DNA and RNA data obtained by various sequencing methods. This includes but is not necessarily restricted to the following categories: [Adapted from Biotech-PRIDE Guidelines]

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).
2. **RNA sequence transcriptomic data:** The nature of the data in this category is similar to those generated by a massively parallel DNA sequencer since usually cDNA synthesis is performed before sequencing. However, recent technological developments allow singlemolecule direct RNA sequencing without cDNA synthesis.
3. **Genotype data:** Modern methods use high-density microarrays to genotype individuals at a large number of loci spread across the entire genome. Whole genome resequencing, Genotyping by sequencing (GBS), and double digest restriction-site associated DNA followed by sequencing (ddRADseq) are increasingly used for genotyping, especially in plants. However, for various specific purposes, small-scale genotyping using microsatellites/ short tandem repeats (STRs)/ simple sequence repeats (SSRs), AFLP, PCR-RFLP, and other similar technologies continue to be used.
4. **Epigenomic data:** These data are also primarily generated using high-throughput methods analogous to a DNA microarray or DNA sequencing after suitable preprocessing.
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 microorganisms 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 the 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 purposes like 16S rRNA, gyrase, and many other genes.

B. LEVELS OF DATA

The submitted data can fall in to one of the following categories:

1. **Raw (Level 1) Data:** This will include the first-level data obtained from raw images/signals. Files and file formats under this category will mainly include, but are not necessarily restricted to FASTA, FASTQ, FAST5, SAM/BAM/CRAM, etc.
2. **Processed (Level2) Data:** Raw (Level1) data are curated, processed, and analyzed to provide value-addition and ease inferences. File and file formats under this category can include assembly data (Including Fasta, GFFs, AGP, and flat file formats), normalized read count table, alignment files, BED, GTF, VCF/BCF, MAFF, WIG, bigWig, bedGraph, etc. As this is a rapidly evolving field, other types of processed sequencing data may also evolve, and will be accepted. Ideally, the file formats are standard in the field and non-proprietary.
3. **Metadata:** Submitters must submit metadata along with the raw or processed data files. This metadata describes how the associated data have been obtained. This helps in enriching the scientific value of the data and also ensures its reproducibility. In reality, the “standards” for metadata requirements differ depending upon the type of data being submitted, therefore, the minimum information required for each type has to be documented. The metadata mainly includes submitter details, project description, sample description, experimental details, file types submitted, and analysis done.

Further details under each attribute are as follows: Submitter details: Name, Affiliation, Contact details
Project details: Description of project
Sample description: The common details under minimum information required for the submission of different kinds of samples will include -Title, Description, Scientific Name, Tax Id, Centre Name, Geographic location (country/sea/state).

Further information required depends upon the sample type as listed below:

- Animal-strain, isolate, breed, cultivar, ecotype, age, dev_stage, sex, tissue
- Human-age, gender, biomaterial_provider, ethnic/ linguistic background, geographical region, tissue
- Plant-isolate, cultivar, ecotype, age, dev_stage, tissue
- Metagenome-host, isolation_source, lat_lon
- Microbial-strain, isolate, host, isolation_source, sample_type
- Pathogen-Strain, Isolate, Collected_by, Isolation_source, Lat_lon, Host, Host_disease, Collection_date, Country, Region, Host_age, Host_sex, Sample_type, Laboratory, Sequencing_platform
- Viral-Strain, Isolate, Collected_by, Isolation_source, Lat_lon, Host, Host_disease, Collection_date, Country, Region, Host_age, Host_sex, Host_health_status, Host_treatment, Sample_type, Laboratory, Sequencing_platform, Viral_load, Passage_history, Symptoms, Treatment_history

The definition of the sample fields is as per the controlled vocabulary of International Nucleotide Sequence Database Collaboration (INSDC) guidelines (<https://www.ncbi.nlm.nih.gov/biosample/docs/attributes/>).
Experimental details: Alias, Samples used, Labeled phenotypes (if any), Library preparation details, Technology used, Instrument used, Library Layout (Single/paired), Runs, Analysis, etc.
Files submitted: Data file and their formats
Analysis: Description of analysis carried out, e.g., denovo assembly, sequence annotation, abundance measurement, variant calling, etc. Some examples of standards for metadata format that can be followed include Minimum Information on MARKer gene Sequence (MIMARKS), Minimum Information for Shotgun Assembled Genomes (MISAG), Minimum information about a microarray experiment (MIAME), Minimum Information on Metagenome Assembled Genomes (MIMAG), Minimum Information About a Genome Sequence (MIGS).

TEMPLATE/CHECKLIST FOR UPLOADING MACROMOLECULAR STRUCTURE DATA

A. SUB-TYPES OF 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 to understand various aspects of protein synthesis under different conditions. [Adapted from Biotech-PRIDE Guidelines]

The three methods utilized to obtain atomic-level three-dimensional structures of biological molecules are **Macromolecular crystallography (MX)**, **Nuclear Magnetic Resonance (NMR)** and **Cryo-Electron Microscopy (CryoEM)**.

B. FILE FORMATS:

1. **Macromolecular crystallography:** Structural coordinates in pdb or mmCIF format and Structure factors in mtz or mmCIF format.
2. **NMR:** Structural coordinates in pdb or mmCIF format. Restraint file along with chemical shift data in a single unified NMR-STAR format.
3. **Cryo-Electron Microscopy:** Structural coordinates in pdb or mmCIF format, primary map, a raw map (unmasked, unfiltered, unsharpened), unmasked half-maps, and Fourier Shell Correlation (FSC) data.

C. LEVELS OF DATA:

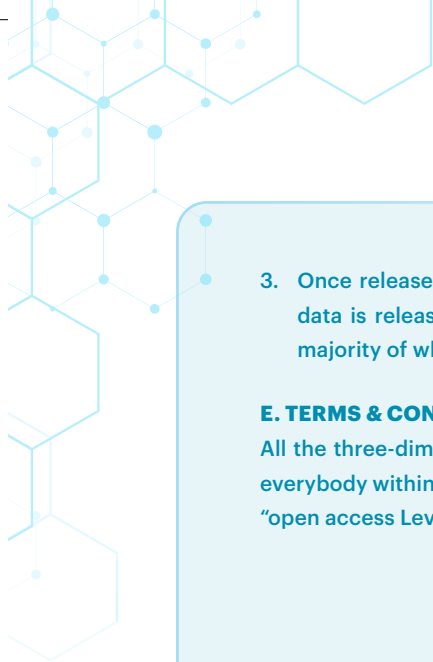
The submitted data can fall into one of the following categories:

1. **Raw (Level 1) Data:** As per Biotech-PRIDE Guidelines, the raw or Level 1 data for structural data is defined as "coordinates for 3D structures for biological macromolecules". The associated experimental data may also be deposited.
2. **Metadata:** The metadata, along with the structural data files, must be submitted. Metadata represents how the associated data have been obtained, which helps improve the data's scientific value and ensures its reproducibility. As the metadata requirements differ depending on the data type, the required information regarding a structure submission has to be documented. The metadata mainly includes submitter details, information about associated publications, project description, sample description, experimental details, and data quality descriptors.

D. DEPOSITION AND SHARING OF STRUCTURAL DATA

In the Biotech-PRIDE Guidelines released by DBT, level 1 structural data is defined as "coordinates for 3D structures for biological macromolecules". The co-ordinates, along with associated experimental data, should be deposited at IBDC. An additional database may be prepared for structural data generated by researchers working in India.

1. The structure portal will house structural coordinates and associated experimental data from these three methods: Macromolecular crystallography (MX), Nuclear Magnetic Resonance (NMR) and Cryo-Electron Microscopy (CryoEM).
2. The data will be under managed access for two years or until the corresponding manuscript is published, whichever is earlier. Before the two-year deadline, the depositor can choose to ask for an extension of the release date, withdraw/replace the data, or release the data.

- 
3. Once released, the structural data will be available under open access category without registration. The data is released in a form that can be viewed or analyzed using standard visualization/analysis tools, the majority of which are open-access and freely available.

E. TERMS & CONDITIONS

All the three-dimensional structures determined by Indian researchers are archived in India and accessible to everybody within the nation without registration. The structural data generated in India is, therefore, available as “open access Level 1 data” as per the Biotech-PRIDE guideline definitions.

TEMPLATE/CHECKLIST FOR UPLOADING PROTEOME DATA

A. PRE-SUBMISSION REQUIREMENTS

1. Two types of dataset submissions can be accommodated, namely “complete” and “partial,” tailored to various proteomics data workflows. The data formats must be organized accordingly for submission, depending on the submission type.
2. The dataset should be arranged based on the data workflow (e.g., DDA, DIA, MRM/PRM/HR-MRM) utilized. If multiple workflows (e.g., DDA, DIA, MRM/PRM/HR-MRM) are employed, it is recommended to split the data into different datasets to facilitate easier interpretation for future users.
3. It is encouraged that the submission of “complete” type data submissions for DDA. All requirements for this type of submission are mentioned in the section below.
4. Files can be compressed using gzip or bzip2, and may be packaged in a tarball.

B. DATA TYPES WITH THE NECESSARY FORMAT FOR SUCCESSFUL SUBMISSION

1. **Mass Spectrometry output files:** Raw data (mandatory) and the derived peak lists (optionally). Raw data file formats: .RAW, .D, .WIFF, .LCD, .JMS, .mzML, .mzXML, .mzData, T2D.
2. **Experimental and Technical metadata:** Minimum sufficient information is required to define the experimental and technical details about the used workflow.
3. **Processed Results:** Two submission types are supported:

Complete submission: A complete submission ensures that the processed results (at least the identification data) and the corresponding mass spectra can be parsed, integrated, and visualized by the IPD resource, connecting the identification data to the corresponding mass spectra. To achieve that, processed identification results need to be provided in a Proteomics Standards Initiative (PSI) open standard format (mzIdentML, mzTab).

Partial submission: When processed identification results are submitted in other data formats than the ones defined for complete submissions. These types of results are not suitable for parsing, integrating, and visualizing the identification and/or connecting the processed results to the corresponding mass spectra. However, all the submitted files are made available to download.

Other files: Other optional components of submitted datasets are mentioned

- a. Output of additional analysis software used (e.g. quantification results).
- b. Protein Sequence database, as used in the search.
- c. Spectral library, if it was used during the analysis.
- d. Images, e.g. Gel images.
- e. Scripts
- f. Additional metadata files
- g. Other

Data type/submission type	IPD
Partial	Yes
Complete: mzIdentML	Yes
Complete mzTab	Yes
Targeted SRM/MRM/PRM/HR-MRM	Partial only
DIA MS/MS	Partial only
Top-down	Partial only

TEMPLATE/CHECKLIST FOR UPLOADING METABOLOME DATA

A. TYPES OF DATA

Types of data to be submitted will include the data generated through analytical techniques such as

- Mass Spectrometry (MS)
- Nuclear Magnetic Resonance (NMR) Spectroscopy

B. LEVELS OF DATA

1. **Raw data:** The spectrometric, spectrographic, and chromatographic data created by the instrument software with the description of the platform and vendor's software version. Data submitters are encouraged to submit data in open-source text-based formats (mzML, nmrML) that others can open and reuse.
2. **Analytical or processed data:** Analysed or processed data includes information on Gas or Liquid chromatography parameters (like MRM/PRM transition, gas pressure, collision energy, detector, etc), software, and algorithms used to analyze the raw data.
3. **Final result matrix:** The result matrix must contain the measured known and unknown metabolites with their associated parameters (chemical shifts, coupling constants, m/z value, retention index, MS peak height, MS peak area, etc.) identified in the study.
4. **The metadata file:** The metadata file may contain the description of the project, its BioRRAPID, study, sample preparation, reference material, storage, and extraction, including the source, organism, and cell line information. It may also contain the details of the instrument and software used to generate, process, and analyze the data.

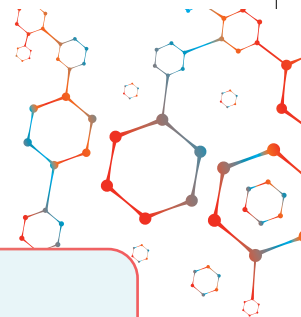
C. DATA SUBTYPES DETAILS

1. Microbial and cell culture (human and animal) specific metadata

- Type of cell type/organoids (like chemotaxonomic)
- Media, buffer compositions, and environmental exposure
- Methodology of obtaining metabolites (like harvesting, lysis, extraction, etc.)
- Storage and transportation conditions

2. Bio fluids and tissue (human and animal) specific metadata

- Sample type (tissues, biofluids such as plasma, serum, urine, fecal, saliva, CSF, BALf, etc.)
- Anonymized patient/animal data (all approved by IEC/IAEC/CTRI and obtained, such as disease type, medication, etc.)
- How it was obtained (Date, time, invasive/non-invasive, surgical procedure, etc.)
- Storage and transportation conditions



3. Plant-specific metadata

- Species/Variety/genotype, tissue type used for analysis, developmental stage,
- Genetic manipulation information (breeding, generation, directed manipulation, etc., and IBSC, GEAC approvals)
- Growth conditions (Like light, temperature, humidity, fertilizer/pesticide/herbicide/insecticide treatments used, time of treatment, stress conditions used, replicate numbers, physio-chemical & microbial data of soil and water, field GIS, irrigation profile, and any other environmental conditions that can impact the data.)

4. Environmental metadata

- Type of sample (soil, water, air etc)
- GIS data
- Sample obtaining procedure (like surface, digging, date, time, season, special geological event, etc.)
- Approvals (Authorities like forests, wetlands, etc.)

5. Food and consumer products metadata

- Type of sample (honey, milk, wine, cosmetics, etc.)
- Sample obtaining procedure (commercial or from the field, etc.)

If obtained commercially, then complete product details like brand, batch number, lot number, manufacturing date, expiry date, date of purchase, date of analysis, study approval number (if any), etc.

If obtained from the field, then: Approvals (Authorities like IAEC, forests, wetlands, etc.), GIS data, Animal/Plant/Microbial Species details, Growth condition details.

D. FILE FORMATS FOR DATA SUBMISSION

Mass Spectrometer

Raw file formats: abf, cdf, cdf.cmp, cmp, d, dat, hr, ibd, jpf, lcd, mgf, qgd, raw, scan.wiff, xp,T2D

Derived file formats: mzML, mzTab, imzML

NMR Spectrometer

Raw file formats: Free Induction Decay (FID) files like fid, ser, jdf. Pulse program and acquisition parameter files.

Derived file formats: Fourier transform data files like 1r, 2rr, jdf, etc.

Deposition in nmrML, mzML, mzTab format is encouraged.

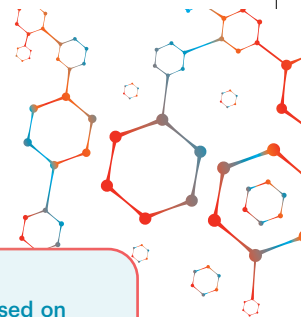


TEMPLATE/CHECKLIST FOR UPLOADING BIOLOGICAL IMAGING DATASET

1. SUBTYPES OF DATASETS: As given in Table XX, the image data can be of multiple types based on the categories given there. Each category may further comprise multiple imaging technologies and various file formats associated with them. In future, the list of subtypes of datasets will be updated with the availability of more data subtypes.

Biological image data sub-types (imaging technologies) and their file formats (this list will be updated whenever required).

S.No	Category	Imaging Technology	File formats
1	Biological Images	Microscopy	SVS, DICOM, PNG, GIF, TIFF, JPG, BMP, PDF, NDPI, VMS, VMU, SCN, MRXS, TIFF, SVSLIDE, BIF
2	Medical Images	Magnetic Resonance (MR)	NiftI, MINC, NRRD, ITK, MGH, MGZ, FSL, DICOM, Analyze, DICONDE, PAR/REC
		CT	DICOM, Analyze, MINC, DICONDE, PAR/REC
		X-ray	DICOM, TIFF, JPEG2000, PNG, RAW, BMP, JPEG
		Ultrasound	DICOM, JPEG, PNG, AVI, MP4
		PET	DICOM, ECAT 7
3	Agricultural Images	Plant Photography	HDF5, ENVI, RAW, NDVI, TIFF, GeoTIFF, BMP, JPEG, PNG
		Thermal Imaging	Radiometric JPEG, BMP, TIFF, JPEG, ENVI, RAW, GeoTIFF
		UAV	TIFF, JPG, PNG, ENVI, GeoTIFF
		Hyperspectral Imaging	TIFF, ENVI
		Multispectral imaging	BIL, ENVI
		LiDAR	LAS, LAZ
		NDVI	TIFF, PNG
		Satellite Remote Sensing	GeoTIFF, NetCDF
4	Marine Images	Hyperspectral Camera	NetCDF, GRIB, NcML
		Azor Drift-Cam,	MP4, MOV, AVI, DNG, WAV, AAC, H.264, H.265
		Photogrammetry	OBJ, STL, PLY, LAS, XYZ, RAW, GeoTIFF, JPEG, TIFF, JSON
5	Livestock Images	Infrared thermography	IDN, IMG, JPG, SIT, IS2, IRI, ANA, TIF, FTS
		Animal photography	TIFF, JPG, PNG, BMP
		Magnetic Resonance (MR)	NiftI, MINC, NRRD, ITK, MGH, MGZ, FSL
		CT	DICOM, Analyze, MINC, DICONDE, PAR/REC
		X-ray	DICOM, TIFF, JPEG2000, PNG, RAW, BMP, JPEG
		Ultrasound	DICOM, JPEG, PNG, AVI, MP4



2. FILE FORMATS: Table 1 provides a comprehensive list of the most common image file formats. Thus, based on the imaging technology used to acquire the images, there can be hundreds of image file formats data available at the data submitter's end. Moreover, for some of the imaging technologies, a common file format may be available. For example, DICOM is a standard file format for the exchange of most medical images. Furthermore, there can be file formats that may be specific to a particular imaging technology image.

3. LEVEL OF DATA: The data deposited by the submitters can have any or all the three levels as given below:

(i) **Raw data (Level 1):** The image data directly received from the biological imaging machines (such as MRI machines, PET scans, Microscopes, CT scan, Drones, Cameras, etc.) with no further processing prior to deposition in the imaging resource. This type of data shall be considered original image data (with no human or software intervention involved) at the time of submission to the image repository.

(ii) **Processed data (Level 2):** This will include any kind of images that are retrieved after the image pre-processing, such as image file format conversion, cropped images, resized images, rescaled images, images with varying resolution (generated through computational tools/algorithms), segmented images, images with a region of interest information, labeled images, augmented images, etc.

(iii) **Metadata:** The images may have various kinds of data associated with them, such as Project, Study, Sample, and Experimental information. These data types may further provide more information such as project and study title, funding agency, collaborators details, type of imaging technology used to acquire biological images, study participants information (number of participants/specimens, demographic, clinical, therapeutic, diagnostic, prognostic, etc.), license type, keywords, organism name, organ, tissue, cells, experimental design summary, images annotation protocols used, experimental instrument information, experimental parameters used, etc. Thus, metadata will enhance the utility of image data for the users (biologists or clinicians, new imaging technology developers, computer scientists/algorithm developers involved in image analysis) interested in this data.

Data submitters should describe the permitted purposes for subsequent research projects, including associated limits and conditions, for all resources hosted in a repository using a common ontology, for example, such as the **GA4GH Data Use Ontology (DUO)**.



TEMPLATE/CHECKLIST FOR UPLOADING PHENOME DATA

A. PHENOME DATA TYPES

Phenome data encompasses the observable characteristics of an organism that result from the complex interplay between its genotype and the environment. These phenotypic characteristics can be systematically categorized based on different life forms, such as humans, animals, plants, and microorganisms. Below is a detailed list of sub-types of phenome data categorized by life forms:

A.1 Plant Phenome Data

- i. **Morphological Traits:** Observable physical characteristics like plant height, leaf size, flower colour, etc.
- ii. **Physiological or biochemical Traits:** Measurements of plant physiological functions, such as photosynthesis rate, water use efficiency, Nutrient uptake and utilization, transpiration rate, etc.
- iii. **Yield-related Traits:** Characteristics related to crop yield and productivity, including traits such as grain yield, biomass accumulation, seed size and weight, fruit yield, etc.
- iv. **Quality Traits:** Attributes pertaining to the nutritional quality, taste, and texture of harvested crops, including nutritional quality (e.g., protein content, vitamin and mineral content), taste and flavour attributes, texture, cooking properties, etc.
- v. **Developmental Traits:** Descriptors of plant developmental stages and processes like germination rate and flowering time as germination rate, flowering time, maturity time, etc.
- vi. **Disease and Pest Resistance/Tolerance:** Measure of plant's resistance or tolerance to diseases, pests, and environmental stresses, including traits such as disease severity, pest infestation levels, resistance to abiotic stresses (e.g., drought, salinity), etc.
- vii. **Root Traits:** These traits describe the characteristics of the plant's root system, including root length, root architecture, root density, root depth, root branching pattern, etc.
- viii. **Environmental Response Traits:** These traits describe the plant's response to environmental factors such as temperature, humidity, light intensity, drought, etc.
- ix. **Molecular and Genetic Traits:** These traits are based on molecular and genetic markers and include gene expression levels, marker-trait associations, QTL (quantitative trait loci) analysis results, etc.

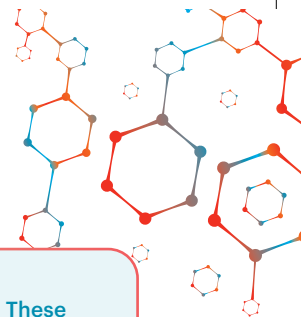
Metadata Standard Reference:

MIAPPE (Minimum Information About a Plant Phenotyping Experiment)

PATO (Phenotypic Quality Ontology) Crop Ontology

A.2 Human Phenome Data

- i. **Physical Traits:** Characteristics like height, weight, body mass index (BMI), skin color, eye color, and hair type. These traits are influenced by genetic makeup and environmental factors.
- ii. **Physiological Data:** Includes measurements like blood pressure, heart rate, respiratory rate, and metabolic rate. This data helps understand bodily functions and responses.



- iii. **Biochemical Measurements:** Levels of substances like blood glucose, cholesterol, and hormones. These measurements are critical for understanding metabolic and hormonal health.
- iv. **Behavioral Traits:** Patterns in sleep, diet, exercise, and psychological traits like mood and cognition. These behaviors influence and are influenced by health and well-being.
- v. **Clinical Data:** Information on disease history, medication responses, and surgical outcomes. This data is essential for personalized medicine.
- vi. **Imaging Data:** Data from MRI, CT scans, X-rays, and ultrasounds. Imaging data provides a detailed view of the internal structure and potential anomalies.
- vii. **Genetic Data:** Includes single nucleotide polymorphisms (SNPs) and gene expression profiles. This data helps link specific genes to physical and physiological traits.

Metadata Standard Reference:

MIABIS (Minimum Information About Biobank Data Sharing): Guidelines for describing biobank data.

HL7 FHIR (Fast Healthcare Interoperability Resources): Standard for electronic health records.

LOINC (Logical Observation Identifiers Names and Codes): Universal standard for identifying health measurements, observations, and documents.

A.3 Animal Phenome Data

- i. **Morphological Traits:** Physical characteristics such as body size, fur color, beak shape, and wing span. These traits are often linked to species identification and adaptation.
- ii. **Behavioral Data:** Includes feeding habits, mating behaviors, and social interactions. Behavioral studies help understand species ecology and social structures.
- iii. **Physiological Data:** Measurements like heart rate, temperature regulation, and metabolic rates. Physiological data provide insights into animal health and adaptation.
- iv. **Health and Disease Data:** Susceptibility to diseases, parasite load, and immune responses. Health data is critical for conservation and veterinary care.
- v. **Genetic Data:** Genotype-phenotype correlations and gene expression profiles. Genetic studies help in understanding inheritance and evolutionary biology.
- vi. **Reproductive Data:** Includes litter size, gestation periods, and fertility rates. Reproductive data is vital for breeding programs and population studies.

Metadata Standard Reference:

ARRIVE Guidelines (Animal Research: Reporting of In Vivo Experiments): Guidelines for reporting animal research to improve reproducibility.

OBO Foundry: Provides a suite of interoperable reference ontologies for various biological domains.

Animal Trait Ontology: Standardized vocabulary for describing phenotypic traits in animals.



A.4 Microorganism Phenome Data

- i. **Morphological Traits:** Cell shape, size, and colony morphology. Morphological data helps in microorganism identification and classification.
- ii. **Growth Parameters:** Growth rate, colony forming units, and biomass production. Growth data is essential for studying microorganism behavior and potential applications.
- iii. **Metabolic Data:** Metabolite profiles, enzyme activities, and respiration rates. Metabolic studies provide insights into microorganism functions and applications.
- iv. **Physiological Data:** pH tolerance, temperature tolerance, and oxygen requirements. Physiological data help in understanding the environmental preferences and survival strategies of microorganisms.
- v. **Genetic Data:** Plasmid content, gene expression profiles, and mutation rates. Genetic data is crucial for studying microorganism evolution and genetic engineering.
- vi. **Pathogenicity:** Virulence factors, toxin production, and host interactions. Pathogenicity data is important for understanding microorganism impacts on health and disease.

Metadata Standard Reference:

MIMARKS (Minimum Information About a Marker Gene Sequence): Standards for reporting marker gene sequences.

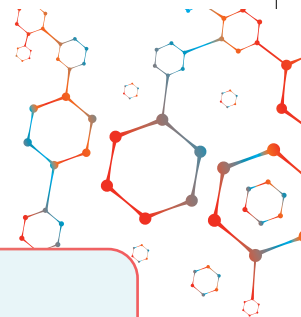
MIxS (Minimum Information about any (x) Sequence): Extension of MIMARKS for reporting various sequence data.

GSC (Genomic Standards Consortium): Standards for genome projects and associated metadata.

BRC (Biological Resource Centers) Standards: Guidelines for the collection, storage, and distribution of microorganism data.

A.5 Insect Phenome Data

- i. **Morphological Traits:** Wing patterns, body size, and antennae length. Morphological traits are essential for species identification and understanding insect diversity.
- ii. **Behavioral Data:** Foraging behavior, mating rituals, and nesting habits. Behavioral studies help in understanding insect ecology and social structures.
- iii. **Physiological Data:** Metabolic rate, thermal tolerance, and developmental stages. Physiological data provide insights into insect health and adaptation.
- iv. **Health and Disease Data:** Parasite load, immune responses, and disease susceptibility. Health data is critical for conservation and pest control.
- v. **Genetic Data:** Gene expression profiles, genetic mutations, and epigenetic changes. Genetic studies in insects help in understanding inheritance and evolutionary biology.
- vi. **Ecological Data:** Habitat preferences, interaction with plants, and predation rates. Ecological data help in understanding the role of insects in ecosystems.



Metadata Standard Reference:

ECOCORE (Environmental Conditions Ontology): Ontology for environmental conditions, useful for describing habitat data.

OBO Foundry: Provides ontologies relevant to insect phenotyping, such as the Insecta Ontology.

B. LEVELS OF DATA

The submitted Phenome data can be 'raw' or 'processed' and essentially associated with metadata as described below:

a. Data: The different levels of the data are as follows:

- **Raw Data (Level 1):** Original, unprocessed measurements or observations collected directly from experiments or phenotyping platforms, requiring pre-processing and quality control. File formats may include CSV, TSV, TIFF, JPEG, TXT, SDF, xlsx, JSON, XML etc.
- **Processed Data:** Raw data subjected to pre-processing to extract meaningful information and relevant features, commonly stored in CSV, XLS, TSV, etc.

b. MetaData: The metadata provides essential descriptors summarizing contextual information about datasets stored within the database, including submitter details, project and study descriptions, biological material, species/strain, age/developmental stage, traits, sex, location, environment details, treatment information, accessions, and data collection, instrumentation used and analysis methods. Minimum information requirements should be based on standards specific for different phenotype data types. Different levels of meta-data are as follows:

Project: Title, Grant number, funding agency, description, associated publication, data access type, project type (individual or consortium), organism and taxonomic information, ethical approval (if required).

Study: Data type, Meta trait, title, description, trial start date and end date, location, experimental design, growth facility, growth & environmental conditions (temperature, light intensity, relative humidity), treatment details (type, agent, description, qualifier, duration, development stage), disease details, Traits, age, development stage, tissue, method details, instrumentation details, clinical data, and author list and any other information specific to data types. The metadata for phenotypic data will be archived using different ontology standards.



DATA MANAGEMENT GROUP 2024

CORE EXPERT GROUP

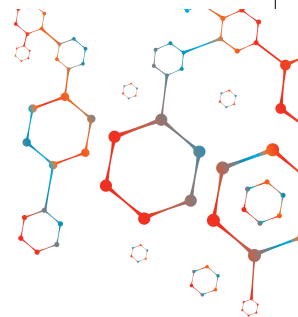
1.	Dr. B Jayaram, IIT Delhi	Co-Chair
2.	Dr. Arvind Sahu, Executive Director, RCB Faridabad	Co-Chair
3.	Dr. Nagasuma Chandra, IISc, Bangalore	Member
4.	Dr. Shandar Ahmad, JNU New Delhi	Member
5.	Conveners of respective Working Groups	Member
6.	Dr. Sunita Reddy, JNU New Delhi (Social Scientist)	Member
7.	Dr Deepak Nair, IBDC-RCB, Faridabad	Co-Member Secretary
8.	Dr Richi V Mahajan, Scientist – D, DBT	Co-Member Secretary

WORKING GROUP 1: NUCLEOTIDE DATA

9.	Dr. Shantanu Sengupta, IGIB New Delhi	Co-Convener
10.	Dr. K Thangaraj, CCMB Hyderabad	Co-Convener
11.	Dr. Vivek Singh, University of Delhi	Member
12.	Dr. Madulika Kabra, AIIMS New Delhi	Member
13.	Dr. D Sundar, IIT Delhi	Member
14.	Dr. Arindam Maitra, NIBMG, Kalyani	Member
15.	Dr. Jaspreet Kaur Dhanjal, IIIT Delhi	Member

WORKING GROUP 2: STRUCTURAL DATA

16.	Dr. Debasisa Mohanty, NII New Delhi	Convener
17.	Dr. Deepak Nair, RCB Faridabad	Member
18.	Dr. Lipi Thukral, IGIB New Delhi	Member
19.	Dr. Pravindra Kumar, IIT Roorkee	Member



WORKING GROUP 3: METABOLOMICS

20.	Dr. Neel Sarovar Bhavesh, ICGEB, Delhi	Convener
21.	Dr. Nirpendra Singh, inSTEM Bengaluru	Member
22.	Dr. Jyothilakshmi Vadassery, NIPGR New Delhi	Member
23.	Dr. Amit Yadav THSTI Faridabad	Member

WORKING GROUP 4: AGRICULTURE/ PLANT DATA

24.	Dr. Shubhra Chakraborty, NIPGR New Delhi	Convener
25.	Dr. Saurabh Raghuvanshi, UDSC New Delhi	Member
26.	Dr. Jitendra Thakur, ICGEB New Delhi	Member
27.	Dr. Mukesh Jain, JNU New Delhi	Member

WORKING GROUP 5: IMAGING DATA

28.	Dr. Dinesh Gupta, ICGEB New Delhi	Convener
29.	Dr. Debajani Paul, IIT Bombay	Member
30.	Dr. Swapnil Rane, TMC Mumbai	Member
31.	Dr. Himanshu Sinha, IIT Madras	Member

WORKING GROUP 6: PROTEOMICS

32.	Dr. Suman Kundu, BITS Pilani, Goa	Convener
33.	Dr. Tushar K Maiti, RCB Faridabad	Member
34.	Dr. Srikanth Rapole, NCCS Pune	Member
35.	Dr. Inderjeet Kaur, Central University of Haryana	Member



EXPERT ADVISORY COMMITTEE, 2024

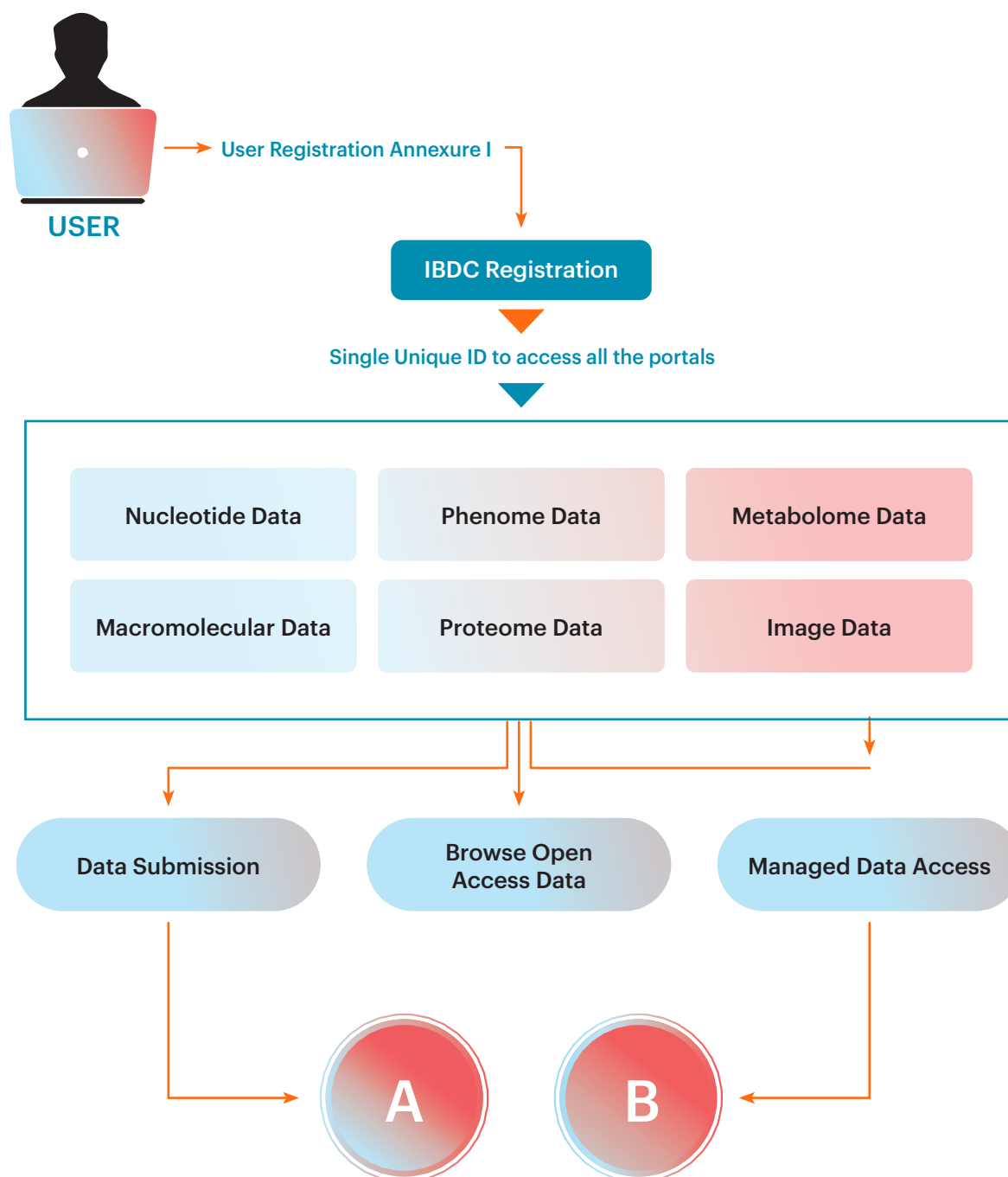
S.No	Committee Members	Designation
1.	Dr. P Balaram, IISc Bangalore	Chair
2.	Dr. Alok Bhattacharya, Ashoka Univeristy, Sonapat	Co-Chair
3.	Dr. B Jayaram, IIT Delhi	Co-Chair
4.	Dr. Suchita Ninawe, Advisor, DBT	Member
5.	Dr. Anand Deshpande, Persistent Systems, Pune	Member
6.	Dr. Arvind Sahu, Executive Director, RCB Faridabad	Member
7.	Dr. Debasisa Mohanty, NII, New Delhi	Member
8.	Dr. Rajendra Joshi, C-DAC, Pune	Member
9.	Dr. B Gopal, IISc Bangalore	Member
10.	Dr. Shandar Ahmad, JNU New Delhi	Member
11.	Dr. Apurva Sarin, India Alliance, Hyderabad	Member
12.	Dr. Dinesh Gupta, ICGEB	Member
13.	Dr. Tilak Raj Sharma, DDG, ICAR	Member
14.	Dr. JBV Reddy, Scientist F, DST	Member
15.	Dr. Harpreet Singh, Scientist F, ICMR	Member
16.	Dr. Sunita Reddy, JNU New Delhi (Social Scientist)	Member
17.	Dr Deepak Nair, IBDC-RCB, Faridabad	Co-Member Secretary
18.	Dr Richi V Mahajan, Scientist – D, DBT	Co-Member Secretary

INTER-MINISTERIAL NATIONAL STEERING COMMITTEE, 2024

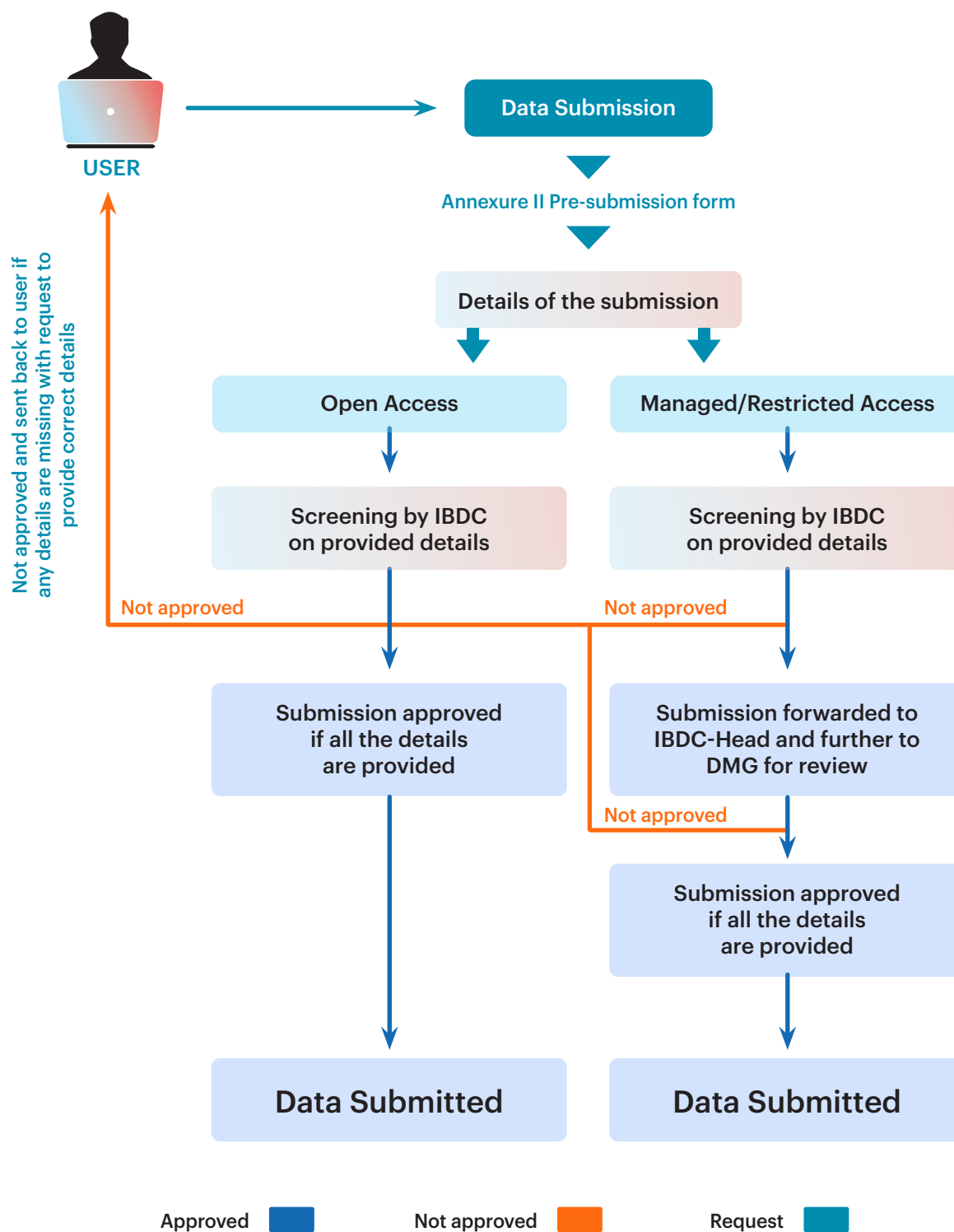
S.No	Committee Members	Designation
1.	Dr. Rajesh S Gokhale, Secretary, DBT	Chair
2.	Dr. Rajiv Bahl, Secretary DHR & DG ICMR	Co-Chair
3.	Sh. C. Achalender Reddy, Chairman, NBA	Co-Chair
4.	Dr. P Balaram, IISc, Bangaluru	Member
5.	Dr. Suchita Ninawe, Adviser, DBT, New Delhi	Member
6.	Dr. Alok Bhattacharya, Ashoka University, Sonapat	Member
7.	Dr. B Jayaram, IIT, Delhi	Member
8.	Dr. Anand Deshpande, Persistent Systems, Pune	Member
9.	Dr. Debasisa Mohanty, NII, New Delhi	Member
10.	Dr. Arvind Sahu, Executive Director, RCB Faridabad	Member
11.	*Representative of DST	Member
12.	*Representative of DSIR & CSIR	Member
13.	*Representative of DARE & ICAR	Member
14.	*Representative of MeitY	Member
15.	*Representative of M/o Health & FW	Member
16.	*Representative of MoES	Member
17.	*Representative of MoEFCC	Member
18.	*Representative of Niti Aayog	Member
19.	*Representative of Office of PSA	Member
20.	*Representative of Office of MHA	Member
21.	Dr. Deepak Nair, IBDC-RCB, Faridabad	Co-Member Secretary
22.	Dr. Richi V Mahajan, Scientist – D, DBT	Co-Member Secretary

*Not below the rank of Joint Secretary/Adviser

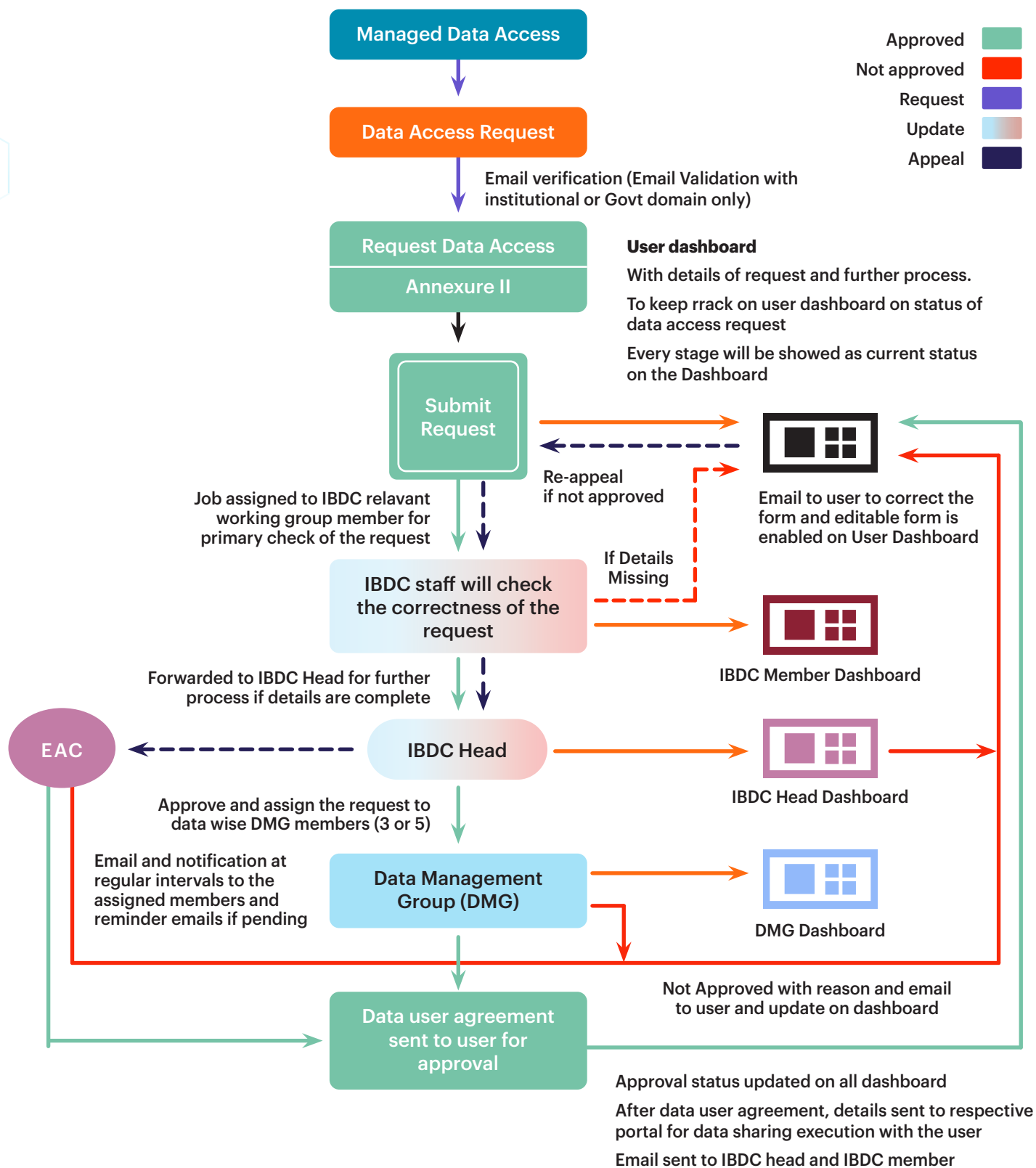
FLOWCHARTS FOR DATA SUBMISSION/DATA ACCESS

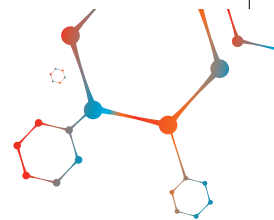


FLOWCHART FOR DATA SUBMISSION



FLOWCHART FOR DATA ACCESS





These protocols emphasize the critical role of data submitter, data user, data custodian and funding agency in conscientious data sharing, and ensuring data quality. By sharing biological data, researchers can collaborate more effectively, avoid duplication of efforts, and gain access to a wider range of information to draw more robust conclusions. This will promote transparency, reproducibility, and accountability in research, enhances the overall quality and reliability of scientific findings.



सत्यमेव जयते

**DEPARTMENT OF BIOTECHNOLOGY
MINISTRY OF SCIENCE AND TECHNOLOGY
GOVERNMENT OF INDIA**

