DATA MANAGEMENT PLAN

OF

THE CZECH CENTRE OF PHENOGENOMICS AND ASSOCIATED DEPARTMENT 18 OF THE INSTITUTE OF MOLECULAR GENETICS

original setup: January 21, 2023
update: March 19, 2026
responsigble CCP data manager: Vendula Novosadova

Supervision:
CCP director: Radislav Sedlacek
CCP deputy manager: Jan Prochazka

1. DATA SUMMARY

INTRODUCTION

The large research infrastructure *The Czech Centre for Phenogenomics (CCP)* and Department 18 of the Institute of Molecular Genetics closely cooperate in research projects and technology development. The PI of both departments is Radislav Sedláček, ensuring strong interconnection and consistent data handling practices. Therefore, the Data Management Plan (DMP) for CCP and Dept18 has been designed as a joint plan to ensure unified standards of data usage and management across both departments.
The research work performed in CCP and Dept18 includes genome and gene sequence data, physiological, molecular biology and biochemistry data, and datasets arising from research on rare diseases, disease modelling, diagnostics, therapy, and infection biology. Consequently, this joint DMP covers a broad portfolio of biomedical data types and formats, including multi-omics datasets and computational outputs. Some multi-omics data will be reused from publicly available repositories. Researchers working with animal models investigating gene variants associated with rare diseases will also reuse phenotyping datasets deposited in the IMPC repository.
A consistent DMP is an essential foundation for standardized data usage, efficient management, and collaboration in a multi-beneficiary research environment where data are both generated and reused. This document describes the data management lifecycle for data collected, processed, generated, and preserved during and beyond the project period. It also ensures alignment with the principles of Findable, Accessible, Interoperable, and Reusable (FAIR) and provides a framework for responsible data handling.
To support the FAIR principles, the DMP describes how data will be collected, documented, curated, shared, and preserved, as well as the standards and methodologies applied
Solution for Individual Grant Projects
Since CCP and Dept18 maintain an extensive data management ecosystem—including internal repositories, databases, and analytical software—all grant and externally funded projects must be aligned with the overarching CCP & Dept18 DMP, including organisational rules and data flow processes. This applies to national grant schemes (e.g., MŠMT, GAČR, AZV, TAČR, Academy of Sciences) as well as European funding programmes (e.g., Horizon Europe)
This chapter provides an overview of the data generated, collected, and reused during the project. It outlines the

types of data, their formats, and estimated sizes, as well as how the data will be documented, shared, and made accessible for internal and external use. The purpose of this summary is to provide a clear understanding of the project's data landscape and support proper planning for storage, management, and reuse.

1.Data Categories

The project will generate four primary categories of data, which differ in their purpose, access rights, and long-term availability:

1. Research Data Scientific research data generated within internal or collaborative research projects.
These datasets support scientific publications and are intended to be made publicly available after the corresponding results are published, following institutional open access policies and FAIR principles. The detailed procedures for data curation, long-term storage, and access conditions will be described in later sections.

2. Phenotyping Data
Data produced through the standardized phenotyping pipeline.
These datasets follow IMPC/INFRAFRONTIER standards and are routinely deposited into international public repositories to ensure global accessibility and interoperability.
The specific data standards, metadata schemas, and repository submission workflows are addressed in subsequent chapters of this DMP.

3. Service generated data Data created as part of contract research or commercial projects.
These datasets are shared exclusively with the respective customers and remain confidential unless contractual agreements state otherwise.
Details on data ownership, confidentiality measures, and contract-specific access rights are discussed in later sections.

4. Metadata Metadata associated with Research, Phenotyping, and Commercial datasets.
They describe experimental design, workflows, protocols, instruments, genotype/phenotype parameters, and data provenance, ensuring full traceability and reproducibility. The structure, standards, and formats for metadata, including ontology usage, persistent identifiers, and interoperability, will be presented in the FAIR Data Principles chapter.

2.Types and formats of data generated

A wide spectrum of data types will be generated within the CCP–Dept18 research activities. To ensure **interoperability**, **long-term usability**, and **compliance with FAIR principles**, standardized and community-accepted file formats will be used wherever possible.
The project will generate both **raw data** and **processed/derived data**, depending on the experimental workflow and analysis pipeline. An overview of the expected file formats for each data category (e.g., genomic, phenotyping, imaging) is provided in **Table 1**.

Table 1: *Data types and formats*

Type of data	Raw data format	Processed data format(s)
Genomics	.fastq.gz	.bam, .vcf, .tsv
Proteomics	.raw	mzTab, .csv (peptide/protein lists from Proteome Discoverer)
Transcriptomics	.fastq.gz	.bam; additional file formats in the pre-processed data are: .html, .txt, .zip, .csv, .sf, .tsv, .pdf, .r and .gtf
MethSeq	.fastq	.bam; .bedGraph
Metabolomics	.raw	After raw data conversion to open format: .mzML After pre-processing: .tsv files (metabolite annotation file, MAF)
Phenotyping	.xml, xlsx	.html; .json; .sql
Images	.tiff, avi .bmp.dicom	.html, .nifty, .gif,
Patient-database	.xml .dicom tiff	.html; .json, DASTA, HL7-FHRI (communication formats)
Alphafold – protein modelling	.fasta	.pdb

Most datasets will be reused internally by project members as well as externally by the broader research community. All data related to model generation and characterization will be **searchable and accessible via the CCP web portal** following internal review and release procedures.

3.Data Size

Expected Data Volume

The expected sizes of the different data types generated or reused in the project are summarized in Table 2 based on previous experience

Table 2: *Expected size of the data generated*

Type of data	Expected size of raw data (per sample if not stated otherwise)	Expected size of processed data (per sample if not stated otherwise)
Proteomics	10 GB	tbd
DNA-seq.& genomics	2 GB – 2 TB	>5 GB
Transcriptomics	>5 GB	15 GB
Metabolomics	~1 GB per sample ~130 GB per analytical batch (~135 samples)	~800 MB total size for .tsv files
phenotyping	45 GB (cohort of RD model) 20MB (10 000 data files)	500 MB (cohort of RD model) 20MB (10 000 data files)
Images	1 GB (cohort of RD model)	100 MB (cohort of RD model)

Newly Generated Data

The majority of datasets will be newly generated (see Tables 1 and 2 for types and estimated sizes). These datasets include:

- Experimental data from animal and cellular models of human origin, including measurements derived through computational analyses.
- Human participant data collected under informed consent covering future research use.

Reused Data

Some datasets will be reused from existing studies, including:

- Patient or study participant data and case report forms, accessible through secure platforms such as ClinData.
- Pseudonymised data, associated with unique identifiers, with personal information accessible only to authorised users.

Data Utilisation Beyond the Project

Generated datasets, including metadata and validated disease models, will be valuable for:

- The broader biomedical research community
- Pharma and biotech users studying rare disease development, model comparisons, or treatment modalities

Additionally, newly collected and existing datasets will support computational approaches to explore disease mechanisms, optimize treatment design, and assess therapeutic effectiveness.

4. Re-using Existing Data

The CCP and Dept18 actively re-use existing datasets to support project objectives. Re-used data include:

1. Genomic DNA sequences for modeling and experimental design.
2. Phenotyping data from previously characterized genes, including datasets from IMPC.
3. Published rare disease literature accessible via PubMed.
4. Transcriptomics and other omics data from public and national repositories, including rare disease registries, the Czech genome/multiome project database, and the National Health Information Registry.

Use and review of these resources are an integral part of project workflows, ensuring efficiency and avoiding unnecessary duplication.

Purpose of Data Generation and Re-use

The primary purpose of data generation and re-use is to establish validated models of rare diseases that support:

- Designing and testing therapies.

- Advancing diagnostics, disease monitoring, and treatment evaluation.

Once the models are established and validated, additional data will be generated to characterize the models, including biomarker analysis and assessment of therapeutic interventions. These datasets, when shared and re-used, will provide significant value to the broader research community and facilitate further scientific discovery.

2. FAIR DATA / FAIR DATA

FINDABLE

All datasets generated or reused in the project will be accompanied by human- and machine-readable metadata to ensure they are easily findable. Each dataset will be assigned a stable and unique identifier, and where possible a **Digital Object Identifier (DOI)**, for example via Zenodo, to facilitate citation and long-term reference.
To enhance discoverability, datasets will follow **standardized naming conventions**, including a unique project code in the format **[Unit]-[Year]-[Project Number]** (e.g., VIS-2025-003). This naming scheme is applied consistently across folders, datasets, and metadata, ensuring traceability and simplifying integration with internal and international repositories.
Datasets will also use **controlled vocabularies and ontologies** appropriate for each data type, such as **IMPC EMPReSS** for mouse models and phenotyping, and community standards for omics data (FAIR genomes, GA4GH, B1MG, HUPO PSI). Metadata will include version numbers and timestamps to ensure reproducibility.
Anticipated storage in dedicated repositories and federated solutions, such as **Federated EGA** or **EOSC-CZ**, will enable indexing and discovery by both humans and machines. Metadata will be structured and enriched to support efficient search, harvesting, and reuse, ensuring that all data produced are fully **findable** within the research community.

MAKING DATA ACCESSIBLE / ZPŘÍSTUPNĚNÍ DAT (ACCESSIBILITY)

All datasets generated within CCP and Dept18 are stored in dedicated repositories according to their type, ensuring secure, long-term availability. Primary internal repositories include OneDrive, Primus, and Volta for general project data, and Vestlon and Pyrat for structured datasets. Internal users have full access to these systems
For external sharing, datasets intended for international consortia (e.g., IMPC, INFRAFRONTIER) receive persistent identifiers or DOIs for reliable location and citation. Where legal or contractual restrictions prevent public deposition, federated solutions such as Federated EGA or EOSC-CZ are used.
Most data generated from standardized procedures—including animal model generation, phenotyping, and preclinical studies—will be made openly accessible after curation, typically following an 18-month embargo to allow publication. Data forming part of intellectual property are released only under conditions that protect IP. Open-access data are available through standardized protocols; controlled datasets require approval by the Data Access Committee (DAC) and a signed Data Access Agreement (DAA).
Long-term preservation is ensured via a dual-branch strategy: one branch stores and displays data locally on CCP servers, while the second branch transfers datasets to the IMPC consortium, where they undergo quality control and integration into publicly accessible repositories. CCP servers serve as the primary site for storage and analysis, while IMPC repositories are maintained long-term with NIH support, ensuring both sustainability and global access.
Metadata are collected throughout all processes using standardized formats (e.g., ISA-tab) to support discoverability, integration, and reuse. Metadata are generally provided under public licenses unless IP or legal restrictions apply. Raw data requiring proprietary software will be provided in open formats whenever possible, with documentation and tools made available. Proprietary CCP software may be licensed for external use under Material Transfer or contract agreements.

MAKING DATA INTEROPERABLE / INTEROPERABILITA DAT

To ensure interoperability, all datasets and associated metadata will be generated and maintained using standardized formats and domain-specific metadata standards. Mouse model generation and phenotyping data will follow the IMPC EMPReSS standards, which include standardized statistics, metadata, and controlled vocabularies. Omics datasets will adhere to relevant guidelines, including MINSEQE for sequencing, MIAPE for proteomics, MIQE for quantitative PCR, and standards from domain-specific archives such as MetaboLights and the European Nucleotide Archive (ENA).

Table 3: *Controlled vocabularies and ontologies used.*

Data type	Identifiers, ontologies, controlled vocabularies
Metabolites	HMDB ID, ChEBI ID (Chemical Entities of Biological Interest)
Measurements and units identification	Units of Measurement Ontology (UO)
Phenotypes	Human Phenotype Ontology (HPO), National Cancer Institute Thesaurus (NCIT) EMPReSS/*: https://www.mousephenotype.org/impress/index. ORPHAcodes – www.orphadata.org and www.OMIM.org codes for rare disease classification.
Genes, transcripts	HUGO Gene Nomenclature (HGNC), Ensembl Gene ID, Ensembl Transcript ID
Gene annotation	Gene Ontology (GO)
Genomic coordinates	Unique identifiers based on chromosome (referred to via GenBank ID and version) and genomic coordinate (GRCh38) corresponding to, e.g., CpG site
Peptides and proteins	UniProtKB Sequence and UniProtKB Accession Number
microRNAs (sequencing and qRT-PCR)	miRbase ID (release 20)
Sample materials	Schema developed by FAIR genomes project
Analysis information	Schema developed by FAIR genomes project
Protocols, methods, experimental metadata	Ontology for Biomedical Investigations (OBI), Chemical Methods Ontology (CHMO), Experimental Factor Ontology (EFO), NCI Thesaurus OBO Edition, Metabolomics Standards Initiative Ontology (MSIO), PRIDE Controlled Vocabulary
Roles and contributions	CRO - Contributor Role Ontology

/ **IMPReSS** *(International Mouse Phenotyping Resource of Standardised Screens), contains standardized phenotyping protocols which are essential for the characterization of mouse phenotypes. IMPReSS contains definitions of the phenotyping Pipelines and mandatory and optional Procedures and Parameters carried out and data collected by international mouse clinics following the protocols defined, allowing comparability, shareability and ontological annotations.*

Wherever possible, standard vocabularies and ontologies will be applied to ensure consistent representation across datasets. In cases where project-specific or uncommon vocabularies are used, mappings to widely recognized ontologies will be provided to facilitate data integration and cross-disciplinary reuse.
All datasets will also include qualified references to related datasets or resources, ensuring clear provenance and traceability. This approach follows established practices in international consortia such as IMPC, INFRAFRONTIER, and EATRIS, and emerging standards in disease-specific initiatives like EOSC4Cancer, enabling seamless data exchange and integration across platforms and research domains.

INCREASE DATA RE-USE / ZVYŠENÍ OPAKOVANÉHO POUŽITÍ DAT

To facilitate data reuse, CCP has developed its own LIMS and data management system, which integrates data storage, analysis, and presentation into a single, user-friendly platform. Data collected from standardized pipelines, including animal model generation, phenotyping, and preclinical studies, are automatically analyzed and made accessible through the CCP portal, providing interactive visualization and statistical summaries for users.
All datasets generated following standardized procedures will be made available through the CCP portal, as well as international consortia portals such as IMPC and INFRAFRONTIER, enabling broad reuse by the research community. Preclinical datasets that are subject to intellectual property restrictions will be released only after internal review by the Data Access and Steering Committees and in compliance with institutional transfer office policies.
Quality assurance is embedded throughout all data generation pipelines. Standardized procedures ensure data integrity, reproducibility, and completeness, and quality control metrics will be provided alongside datasets and metadata. This ensures that all shared data are reliable, well-documented, and suitable for reuse in subsequent analyses.

3. DATA STORAGE

All datasets generated within CCP and Dept18 are stored in dedicated repositories according to their type, ensuring secure, long-term availability. Primary internal repositories include OneDrive, Primus, and Volta for general project data, and Vestlon and Pyrat for structured datasets. Internal users have full access to these systems.
To ensure consistency, traceability, and reproducibility, all data are organized following standardized folder structures based on project type, user group, and project naming conventions. Each project is assigned a unique code in the format [Unit]-[Year]-[Project Number] (e.g., VIS-2025-003), which is applied consistently to folder names, datasets, and metadata.

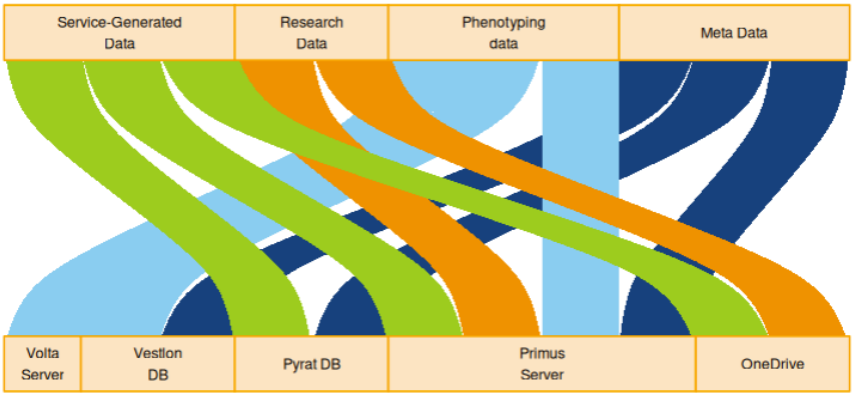

Figure 1:Data storage in CCP

Research Data are stored on for individual projects on \\primus.img.cas.cz\data\18_lab (e.g., Surname_Name),for collaborative projects involving multiple users on OneDrive - img.cas.cz\83_Projects and for projects requiring collaboration within a single unit on \\primus.img.cas.cz\data\83_Unit\UNIT\Experiments
Raw data are always stored on Primus, while preprocessed data, reports, and other outputs can be stored on OneDrive. Details about data processing, storage paths, and project information are recorded in the personal electronic lab notebook (ELN**)**.

Service-Generated Data are stored on Primus for individual requests ii \\primus.img.cas.cz\data\83_Request\UNIT or \\primus.img.cas.cz\data\64_tgu, For collaborative projects: \\OneDrive - img.cas.cz\83_Operations\Institute_ProjectPI\RequestNumber_ProjectName
These data include experimental plans, reports, data, statistics ouputs. Raw data for collaborative projects are always saved on primus.

Phenotyping Data
Stored similarly to service data: \\primus.img.cas.cz\data\83_IMPC\UNIT, Vestlon database, Volta server. Metadata and project information are maintained according to IMPC standards to ensure interoperability and reusability.

Structured animal/mouse data are meant as part of metadata and they stored in PyRat, They includes information about the animals and instructions on how to handle the mice. PyRat ensures traceability of each animal and compliance with internal and regulatory standards.

4. DATA WOKFLOW

The CCP data workflow comprises four interconnected stages, beginning with data acquisition and ending with the generation of fully curated data products. The process begins with data origination, which includes both external datasets downloaded from publicly available resources—such as genomics, transcriptomics and phenotyping datasets—and data generated internally within the institute. Internally produced data include primary and

secondary phenotyping outputs, preclinical datasets, animal model development data and rare disease model characterisation. All incoming datasets are accompanied by operational data that describe the experimental context, metadata, animal records and request-based information submitted through systems such as Pyrat or internal request forms.

Raw and preprocessed data are stored on a combination of internal and authorised external servers, and integrated into databases such as Vestlon, Pyrat and PDX. Data ingestion is facilitated by specialised tools, including VestlonApp for IMPC, and PDX Manager and GraviBase for the integration of preclinical study data. Scientific data are integrated through LabCollector, which contains personal Electronic laboratory books for every person in CCP and integrates information about running experiments.Prior to ingestion, all datasets undergo validation procedures to ensure integrity and consistency. Data are stored in open, machine-readable formats to ensure interoperability and support downstream analysis.

Preprocessed data enter a multi-step processing pipeline where they are cleaned, quality-controlled, validated and analysed using a variety of software environments, including R, Python, Matlab, Java and domain-specific tools. For more complex analytical needs, CCP employs advanced computational methods such as machine learning, multivariate modelling, neural networks and specialised image-analysis workflows. These processes produce enhanced data that are ready for publication or further dissemination.

The final stage focuses on data publishing and distribution. Depending on their nature, sensitivity and intended use, datasets are stored in internal repositories with restricted or institutional access, or made available through interactive Shiny applications such as Phenolyzer2, CohortsAPP, CTimgAPP, JuvAPP, LacZApp, YolkSackAPP and other specialised tools. Service-generated datasets are processed through automated reporting systems, including PhenoGenReport, which generate standardized HTML, Excel or Word reports for internal stakeholders or external customers. Data designated for public dissemination are deposited into external repositories such as Impress and FaseBase, ensuring long-term accessibility and compliance with FAIR principles.

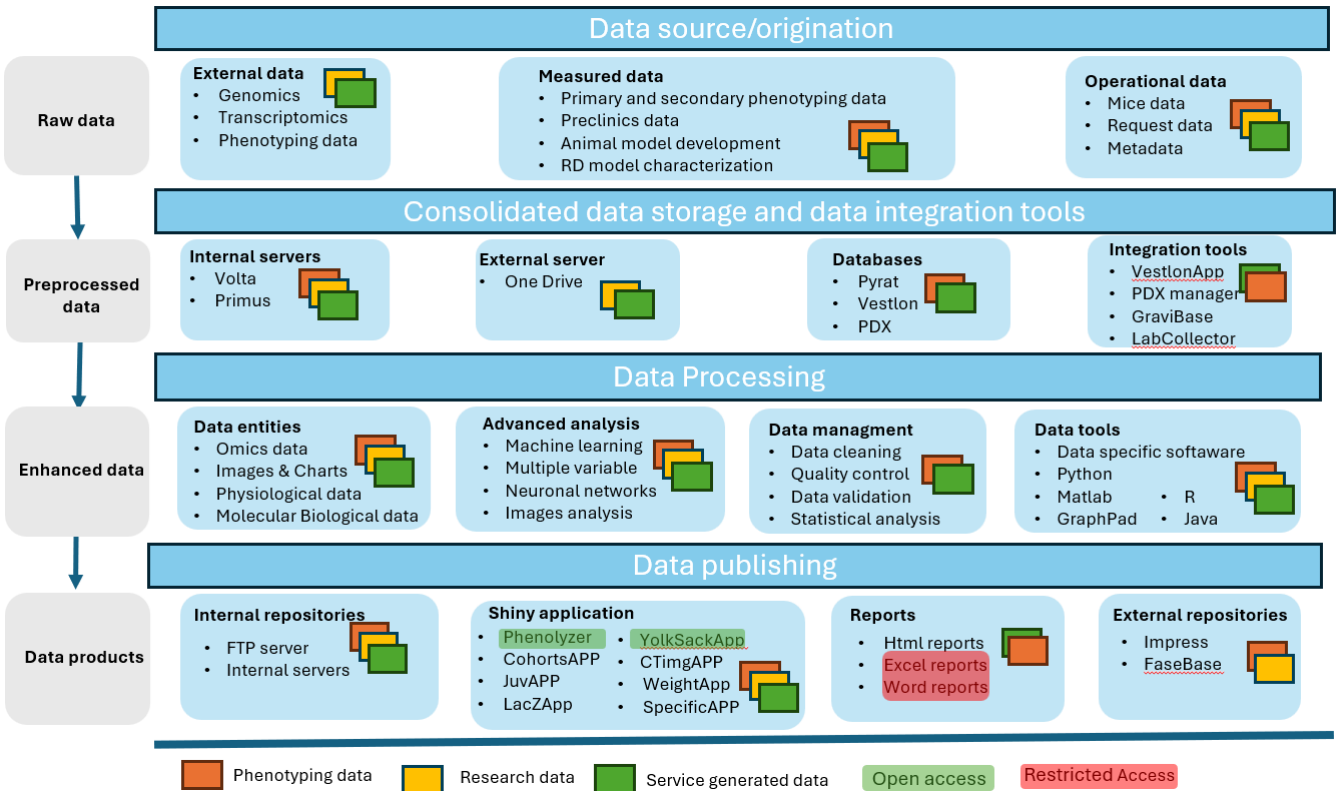

Figure 2: Overview of the CCP data lifecycle, illustrating the end-to-end workflow from data origination to the generation of final data products.

5. DATA SECURITY & PRIVACY

CCP and Dept18 implement comprehensive measures to protect all research data, particularly sensitive personal

or human subject data, throughout their lifecycle. Access to internal repositories is controlled through user authentication and role-based permissions, ensuring that only authorized personnel can view or modify data. Sensitive data, such as patient-derived samples or personally identifiable information, are pseudonymised and associated with unique identifiers. Direct identifiers are stored separately and securely, in compliance with GDPR and institutional policies. Data transmitted between systems or to external repositories is encrypted, and secure protocols are used for all data transfer.

The Data Access Committee (DAC) oversees requests for controlled data, reviewing applications from internal or external researchers and ensuring that data access agreements (DAAs) are signed when required. This process guarantees that sensitive data are only shared in accordance with legal, ethical, and institutional requirements.

For long-term storage, internal and federated repositories are regularly backed up, and data integrity is monitored to prevent corruption or accidental loss. Proprietary or preclinical data that may be subject to intellectual property restrictions are stored in controlled-access environments, with access limited to authorized personnel under material transfer or contractual agreements.

By combining technical, organizational, and administrative measures, CCP and Dept18 ensure that all data are securely stored, legally compliant, and ethically managed, while remaining accessible to authorized researchers for legitimate purposes.

6. ETHICS / ETIKA

CCP and Dept18 ensure that all research data, particularly those derived from animal models and preclinical studies, are collected, processed, and shared in accordance with institutional, national, and international ethical standards. Ethical compliance is an integral part of project planning, data management, and dissemination.

1. Ethical Handling of Animal Data

All work with animal models follows established protocols for animal care and experimentation, complying with relevant legislation and ethical guidelines. Data generated from animal studies are documented and stored responsibly, ensuring reproducibility, traceability, and accountability.

2. Data Sharing and Governance

The Data Management Committee (DMC) oversees controlled access to datasets. Open-access data are shared according to FAIR principles, while restricted data—such as preclinical datasets or proprietary research data—are shared only under Data Access Agreements (DAAs). This ensures proper use and prevents misuse of sensitive or potentially confidential information.

3. Intellectual Property and Contractual Considerations

Data generated in preclinical or contract research projects may be subject to intellectual property (IP) protection. Release of such data follows institutional transfer office rules and occurs only under appropriate agreements (e.g., Material Transfer Agreements or DAAs). This approach balances the ethical duty to share knowledge with the protection of IP rights.

4. Compliance with Legal and Institutional Policies

All data management activities comply with institutional policies, national legislation, and applicable international standards. CCP ensures that data storage, transfer, and sharing follow secure protocols and that users accessing restricted data are properly authorized.

7. ROLES AND RESPONSIBILITIES

Effective data management requires clear assignment of roles and responsibilities. At CCP and Dept18, responsibilities are distributed among governance, project leadership, and technical staff to ensure compliance with FAIR principles, security standards, and ethical guidelines.

1. Principal Investigators (PIs)

The PIs of CCP and Dept18 are ultimately responsible for the integrity, quality, and ethical compliance of all data generated within the departments. Their duties include:

- Oversight of data management practices.
- Approval of data sharing policies, embargo periods, and access agreements.

- Coordination with funding agencies and international consortia.

2. Data Management Committee (DMC)

The DMC supervises controlled data access, compliance, and preservation. Its members include senior representatives from CCP governance:

- **V. Novosadová** – Data Manager
- **J. Procházka** – CCP Deputy Director
- **R. Sedlacek** – CCP Director

The DMC's responsibilities include reviewing external requests for restricted data access, Approving Data Access Agreements (DAAs) and monitoring adherence to internal policies and international guidelines.

3. IT Staff

Technical staff manage the infrastructure required for data storage, preservation, and secure access. Responsibilities include:

- Maintaining CCP repositories (OneDrive, Primus, Volta, Vestlon, Pyrat) and ensuring backups and long-term preservation.
- Implementing secure access protocols, authentication, and role-based permissions.

4. Data Managers

- Monitoring data quality, integrity, and metadata compliance.
- Supporting FAIR implementation, including metadata curation and DOI assignment

5. Researchers and Collaborators

Researchers generating or processing data are responsible for:

- Ensuring proper documentation, metadata collection, and adherence to standardized formats.
- Following internal procedures for data security, embargoes, and ethical compliance.
- Depositing data in appropriate repositories according to project policies.

6. Institutional Oversight and Support

The CCP governance, including IT and legal teams, provide oversight and support for:

- Compliance with institutional, national, and international policies.
- Intellectual property protection and contract research agreements.
- Training and guidance on data management best practices.