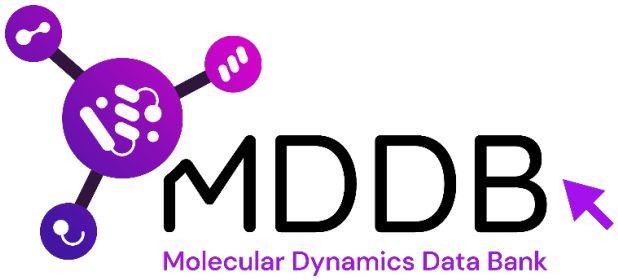

Project Acronym: MDDB

Project title: Molecular Dynamics Data Bank. The European Repository for Biosimulation Data

Call: HORIZON-INFRA-2022-DEV-01

Topic: HORIZON-INFRA-2022-DEV-01-01- Research Infrastructure Concept Development

Project Number: 101094651

Project Coordinator: Institute for Research in Biomedicine (IRB Barcelona)

Project start date: 01/03/2023

Project end date: 28/02/2026

Deliverable 4.1: Data management plan

Work Package: WP4- Implementation, Sustainability and community engagement

Lead beneficiary: EMBL-EBI

Dissemination level: PUBLIC

Due date: 31/08/2023

Actual submission date: 31/08/2023

Funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Document history

Version	Main Contributor(s)	Partner	Date	Comments
0.1	Adam Bellaiche	EMBL-EBI	July 2023	First draft
0.2	Adam Hospital Erik Lindahl Josep L. Gelpí	IRB-CERCA KTH BSC-CNS	August 2023	Revision and contributions to the document
1.0			31/08/2023	Version 1

Table of contents

Executive summary

This document contains the first version of the first deliverable of the MDDB project: the data management plan. It evokes the various FAIR principles and how we will respond to them, while laying the foundations for our data management.

1 INTRODUCTION

The MDDB project is part of the EU's Open Data Institute (ODI). To make the most of open data, it is necessary not only to store the data, but also to make it findable, accessible, interoperable and reusable (FAIR). We support open data and FAIR, but we also take into account the need to protect individual data sets.

The purpose of this document is to provide guidelines on the principles guiding data management in the MDDB project and on the data that will be stored using the responses to the EU Data Management Plan (DMP) questionnaire. We will answer these questions in one or more paragraphs.

The detailed DMP indicates how the data will be handled during and after the project. The MDDB DMP is modelled according to the Horizon 2020 and Horizon Europe online manual. It will be updated and checked for validity several times during the course of the MDDB project.

2 Data summary

2.1 Aims of the project

Will you re-use any existing data and what will you re-use it for? State the reasons if re-use of any existing data has been considered but discarded.

What is the purpose of the data generation or re-use and its relation to the objectives of the project?

What is the origin/provenance of the data, either generated or re-used?

To whom might your data be useful ('data utility'), outside your project?

The project will provide storage and sharing to Molecular Dynamics Simulations (MDS) data to different communities. This will comprise both existing data and newly generated data. MDS databases already exist but **(a)** those are more specific to a certain family or certain families of molecules/macromolecules/proteins or **(b)** these databases don't enforce FAIR principles in all aspects particularly when it comes to interoperability, reuse, and providing quality measure indicators to also *promote* and not merely enable reuse.

There is a growing interest and multiple reasons in using and reusing molecular dynamics data: **(1)** Not re-running similar simulations and thus reduce the computational load on supercomputers and the carbon footprint, **(2)** making it possible to critically assess previous studies and identify issues in algorithms, analysis or program code by having full access to the raw results, and **(3)** providing a fast and secure access to quality MDS data which can be reused and analysed in many ways in order to review experiments, to draw new interpretations, and perform meta-analyses e.g. about the accuracy of simulations in general or parametrizing machine-learning and artificial intelligence methods. This will provide a better vision of the dynamics of molecules and will allow advances in many fields such as medicine, biology, physics, chemistry or computer science.

Data production is not part of the project. It intends to collect, store and share existing or newly generated data from the community of experts in MDS. The objectives are multiples:

- provide the mechanisms and recommendations to make MDS data FAIR,
- promote and allow the full and useful integration of MDS data to the Life Sciences Data Ecosystem (LSDE),
- define and promote good practices for the MDS data generation but also for their analysis. It will allow tracking provenance and make MDS reproducible using metadata and documentation.

From a technical point of view, this project will provide appropriate structures, governance models and policies for such data.

MDS data is widely scattered in the community in large proportions. We want to bring this community to both share their data and establish community-accepted procedures for how to define and assess quality of data generation. It will provide significant information for many communities: biologists, physicists/chemists, physicians, pharmaceutical industries, agri-food industries or bioinformaticians. However, these data are difficult to interpret and to understand by inexperienced communities. Therefore, a real effort will be made to make them accessible and understandable. The project also aims to save resources and improve quality of published science by providing early warnings to users of common simulation codes when their chosen setup would result in generated simulation data that fails one or more criteria required for deposition in the database.

2.2 Data types and formats

What types and formats of data will the project generate or re-use?

What is the expected size of the data that you intend to generate or re-use?

MDS data is usually grouped into two main categories, that depending on the software used can correspond to single files or collections of them:

- **(1)** Static data that does not change during a simulation. This includes the molecular "topology" describing the atoms and connectivity of the system, labels on molecules and all the simulation parameters describing the physical algorithms used for the simulation (in the case of the GROMACS code, this is contained in the .tpr "run input file") For convenience this type of data may also contain the initial simulation coordinates to have a single file that can be used to start/reproduce a simulation.
- **(2)** The actual output data from the simulation in the form of trajectories that contain frames corresponding to the 3D positions (x,y,z) of each atom of the system at a specific time (for GROMACS the compressed-coordinate file is .xtc). In some cases, it will also contain velocities, forces and energies of the system at a specific time (for GROMACS the file is .trr). There might also be other types of data such as potential energies, free energies, or other statistical analyses performed on the fly, which are stored for convenience (for GROMACS, this is present in the so-called "energy file", .edr).

Historically, many simulation codes have not been explicit about storing metadata such as the commands used, history of files, software versions and the specific hardware the simulation was run on. It is however an explicit goal of the MDDB project to expand towards storing this type of information too.

There are a lot of data formats for the two categories of files depending on the software used. MDAnalysis, a famous python package for analysing MDS data, has a web page presenting all the formats it can read. There are no fewer than **37 different formats** for storing trajectories. Whereas MDTraj, another famous python package, can read **15 formats** (see here). This goes some way to explaining how difficult it can

be to analyse MDS. It's going to require some concessions, moving from one package to another, or even not having a package at all.

Those formats are not adapted for storage and sharing. MDDB aims to create new community-based standards and highly compressed formats to allow storage, sharing and re-use following FAIR principles, in particular by defining translators and libraries so that all codes can use highly efficient/compressed formats without writing an implementation from scratch.

The size of the data generated by molecular dynamics simulation is variable and depends on the system studied. A trajectory (the output of a simulation) can reach a large volume in the terabyte range. It is possible to reduce a trajectory with a slight loss of information. In this case, raw data approaches terabytes while pre-processed data can be reduced from 10 GB to 100GB.

For special cases, for instance when it is sufficient to only store coordinates of slowly moving parts such as the protein (but not water or lipids), modern techniques for multi-frame compression (similar to movies, whereas traditional compression methods work on one frame at a time) will enable us to reduce this by another order of magnitude.

3 FAIR data

3.1 Making data findable, including provisions for metadata

Will data be identified by a persistent identifier?

Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

Will search keywords be provided in the metadata to optimise the possibility for discovery and then potential re-use?

Will metadata be offered in such a way that it can be harvested and indexed?

As said above, the MDS data are mainly represented by two types of files: (1) the topology and (2) the trajectories (to which energy files belong). The topology can be linked to several trajectories. However, the trajectory is normally associated with a single topology. Thus, these topologies will be findable via a unique persistent identifier (DOI). We will also work with code implementations such that future file formats can contain provenance information about what other files/DOIs they were generated from, and add support for digital signatures to guarantee the integrity of data.

There is a large set of metadata that can be associated with a simulation and there are no standards yet. The table 1 shows, as a example, an initial set of MDS metadata to make them F.A.I.R. The metadata can be divided in six categories:

1. Hardware: the hardware's components used for the simulation,
2. Software: the software's components used for the simulation, both the version of the simulation code and build metadata such as the compiler, libraries and options used for the build,
3. The user commands executed to produce a particular output,
4. Molecule(s) of interest: a deep and organised description of the studied molecule(s), and the modifications introduced from the original structure if any.
5. Physical-Chemical (PC) parameters of the system,
6. Comments from the author and related articles.

Table 1: MDS metadata

HARDWARE	SOFTWARE	COMMAND	MOLECULE(S)	PC PARAMETERS	COMMENTS
CPU type GPU type CPU number GPU number Nodes numbers **If exist:** reference of the supercomputer	MD software Integrator Time step Duration Job scheduler OS Constraints	Full command line Environment variables affecting the run ID of the user producing the output	Type/family wwPDB ID N atoms N residues Connectivity Titration Chains Annotations on domains or functional sites Initial Coord.	Force field and potentials Temperature + thermostat Pression + barostat pH Solvent Ionic concentration Box shape and dimensions	Why was the simulation set-up? Related publications Points to be aware of Interpretation

There is a lot of metadata to provide for each MDS. We are evaluating the possibility of using third-party software such as AIIDA-Gromacs. It captures the full data provenance for simulations made with GROMACS. MDDB itself should however be code-agnostic, but it could be adapted to other software.

We will initially store data as free text, but in parallel define strict ontologies to describe simulation metadata. Searches among these metadata will be done via strictly determined keywords with case-sensitivity, and with ontologies in place it will be possible to use complex queries of combined conditions. The vocabulary could be inspired and associated with the PDBx/mmCIF dictionary (and integrated to it). Expert users will also be able to propose new flags, keywords and ontology extensions where we will establish a process for the community to review these for a new version (e.g. yearly) …

As an example, we will also implement support at least for GROMACS to *automatically* generate complete submission metadata records that can be tested against (evolving) MDDB quality standards locally already when starting a simulation. This will both enable users to apply the same standards for provenance in all their stages of data

analysis, and get an early “green”, “yellow” or “red” flag for whether their setup is compliant with community high-quality simulation standards.

Thus, MDS metadata will be able to be deeply explored and provide valuable information. For example, MDS is usually done at a specific temperature. Through the exploration of the metadata by a keyword/flag *“#temperatures”*, it will be possible to get some statistics about temperature. Then, it will allow us to see if some temperatures are over-explored while others are not. From a point of view of biopharmaceuticals manufacturing it is important to explore the temperature impact on a large panel of proteins. Indeed, biopharmaceuticals are produced in a large range of temperature (253K to 310K) and their stability can be impacted by such variations.

Making MDS data FAIR is a topic currently being explored and initial approaches already exist which the project will build on. For instance, there is the Simulation Foundry that is a modular workflow for the automated creation of molecular modelling (MM) data. The SF makes MM repeatable, replicable, and findable, accessible, interoperable, and reusable (F.A.I.R.).

3.2 Making data accessible

Will the data be deposited in a trusted repository?

Have you explored appropriate arrangements with the identified repository where your data will be deposited?

Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?

Will all data be made openly available? If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why, clearly separating legal and contractual reasons from intentional restrictions. Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.

If an embargo is applied to give time to publish or seek protection of the intellectual property (e.g., patents), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

Will the data be accessible through a free and standardized access protocol?

If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?

How will the identity of the person accessing the data be ascertained?

Is there a need for a data access committee (e.g., to evaluate/approve access requests to personal/sensitive data)?

Will metadata be made openly available and licenced under a public domain dedication CC0, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?

How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?

Will documentation or reference about any software be needed to access or read the data be included? Will it be possible to include the relevant software (e.g., in open-source code)?

MDDB repository is largely envisioned as an Open Science infrastructure through a free and standardised access protocol following the tradition in the structural field. MDDB © 2023 is marked with CC0 1.0 Universal by the consortium of IRB, BSC, KTH, EMBL-EBI, University of Oxford, Nostrum Biodiscovery and EPFL-CECAM. Since there is increasing awareness of value in data itself (not least for AI/ML training), we might also explore possibilities to allow different types of data/reuse licensing, which possibly could help with MDDB sustainability by providing additional income, and enabling users to e.g. search only for simulations that can be reused commercially.

Data will be made available via the MDDB platform using a user-friendly front end that allows data visualisation, as a first proof-of-concept towards complete Virtual Research Environments. No specialised software will be needed to access the data, usually just a modern browser. Access will be possible through web interfaces. In this context, a quality trust must be established. Each MDS data will be associated with a DOI number and metadata will provide a detailed provenance. All (meta)data will be released with a clear and accessible data usage marked with CC0 1.0 Universal in human and machine-readable format. In that way, the project will draw on the work of the community of RO-Crate and Biocompute objects.

MDDB will have RESTful APIs. the database will integrate other APIs to enable data manipulation for analysis, modelling, etc.

As far as possible, data will be made openly available. Completely anonymous access is normally allowed. However, some MDS data come from private industries or related to articles that have not yet appeared. Thus, some datasets cannot be shared (or need to be shared under restricted access conditions). Access authorizations can be set up according to the login (for private industries) or an embargo can be put in place until the article is published. The person accessing the data will be identifiable, but we will work on technologies to enable an intermediate shield layer, for instance to share data with paper reviewers known to an editor, but not the author. These accesses will be managed by the authors of the data through the appropriate mechanisms that will be made available to the requesters. Technologies currently being applied to the access to sensitive genomics data (ex. Passport/Visa std. By GA4GH) will be considered.

Regarding the metadata, it will be accessible without restrictions and after the deletion of the described data. Also, metadata will contain cross-links to used software's documentation, and with software versions as well as complete settings metadata stored, in the future it might also be possible to extend this to with bug tracking reports from codes to flag simulations that might have been influenced by serious bugs in algorithms or implementations.

3.3 Making data interoperable

What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?

In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining or extending them?

Will your data include qualified references [1] to other data (e.g., other data from your project, or datasets from previous research)?

MDS data will be useful for a large number of experiences/inexperienced communities. The Life Sciences Data Ecosystem (LSDE) includes a lot of biological databases like PDB, Uniprot, GenBank, PDBe-KB, Binding DB. The Protein Data Bank uses PDBx/mmCIF as its master format to represent structural data. The format employs a text-based file format to encode both data and metadata, utilising data items organised into distinct categories. Within this framework, the PDBx/mmCIF dictionary establishes a standardised collection of categories and data items, along with controlled vocabularies and explicit linkages among various categories and data items. The PDBx/mmCIF format has already been successfully employed to incorporate cross-references and annotations from several prominent sources such as UniProt, GenBank, Pfam. Moreover, the format is highly extensible as demonstrated by IHM extension for models from integrative/hybrid methods or ModelCIF extension for computationally predicted models. This inherent extensibility of the format enables the seamless inclusion of new data elements and classifications. Leveraging this adaptability, it becomes feasible to seamlessly include MDS data. This integration offers the potential for comprehensive storage of results and insights derived from MDS trajectories. This inclusion will significantly enhance the biological context and understanding of biomolecular systems. MDDB will use existing de-facto standards for formats wherever available to promote interoperability. Even where large-scale data is highly compressed or contains specific descriptions (such as connectivity and metadata), MDDB will aim to also enable extraction of individual representative frames from each dataset such that users can instantly visualise (online or offline) data without requiring specific codes.

Ontologies (EDAM) of formats from MD engines already exist (here for Gromacs and here for Amber). It will be enriched and used as a start to create our own ontology and data management system. In particular, we will extend ontologies to be able to provide the rich references to other data sets and versioning both of datasets and software producing them to drive interoperability. All such output will not only be shared with the community, but the project also aims to develop complete libraries to read and write such data, with liberal licences such that it can be linked into any code. The project has already initiated collaborations with the largest-scale community resources e.g., for analysis and visualisation of simulation data, to ensure these projects are instantly able to read & write the relevant data sets, metadata, and ontologies.

The integration of biological data can be overwhelming for the user due to the amount of information it contains. Thus, the integration will follow the lazy approach and will contain some translation layers to give clear information to the user.

With MDS results stored in such standardised format, researchers can readily access and reuse previously conducted analyses. This not only streamlines research processes but also ensures that valuable insights gained from MDS trajectories are readily available, enabling researchers to build upon existing knowledge and delve deeper into the biological implications of dynamic structural changes.

3.4 Increase data re-use

How will you provide documentation needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?

Will your data be made freely available in the public domain to permit the widest re-use possible? Will your data be licensed using standard reuse licenses, in line with the obligations set out in the Grant Agreement?

Will the data produced in the project be reusable by third parties, in particular after the end of the project?

Will the provenance of the data be thoroughly documented using the appropriate standards?

Describe all relevant data quality assurance processes.

Further to the FAIR principles, DMPs should also address research outputs other than data, and should carefully consider aspects related to the allocation of resources, data security and ethical aspects.

MDS data represents a highly valuable source of information, often employed to address specific scientific investigations. These investigations not only produce results but also encompass a series of analyses conducted by the authors. These analyses will be meticulously documented and enriched with metadata, enabling through review and interpretation. In particular to encourage re-use and meta-analysis, each data set will contain complete metadata sufficient to exactly reproduce the simulations, using strict ontologies, which will also enable metadata mining e.g. of all simulations including lipids.

We expect data sets will contain some basic analyses - such as potential energies, temperatures, pressure and other features that are naturally calculated and stored during standard simulations. The decision regarding the inclusion of author-specific analyses, along with accompanying scripts in the MDDB requires careful consideration within the scientific community. While MDDB primarily might include topology, coordinates and metadata (including energy/data files), the inclusion of specific analyses and associated scripts could provide a deeper understanding. The merits of preserving these analyses will be discussed collaboratively, ensuring that the chosen approach aligns with the needs and preferences of the community at large.

The data generated can be reused to answer many other questions. To be reused, data must be clearly described by generous metadata (especially for provenance) but also

be of high quality. In this context, documentation on best generation and analysis practices will be available and standards will be created. On the other hand, the quality of the deposited data will be checked by automatic scripts - and the specifications used to assess the metadata will also be implemented on a trial basis in molecular simulation codes such that users get early warnings or recommendations about how to configure their simulations *before starting their projects* in order to enable later data deposition. These quality requirements need to be discussed with the communities, and are likely to naturally evolve over years - and establishing procedures to set these standards is another goal of the project likely to improve both documentation and quality of data in general, and establish the trust required for reuse of existing data sets in particular.

In the same way, Virtual Research Environment (VRE) based on OpenVRE will be implemented to build user friendly portals. Such VRE would provide access to the repository data (integrated with other bioinformatics databases), a comprehensive offer of analysis and visualisation tools, and procedures for advanced users to incorporate their own tools to the system. A layer of pre-calculated analyses performed automatically in any newly deposited data. This will include data validation and quality control as an initial step on accepting new depositions. Additionally, traditional analysis as essential dynamics, contacts conservation or helical parameters in the case of Nucleic Acids, among others, will be added to the repository.

Users will be able to reuse the various data under a standard reuse licence to conduct their own research and publish their results, citing the author and the project - by default the project will suggest Creative-commons based licences. However, we will consider the possibility of letting authors choose a more protective licence if justified (industry, ethics, security, etc.).

4 Other research outputs

In addition to the management of data, beneficiaries should also consider and plan for the management of other research outputs that may be generated or re-used throughout their projects. Such outputs can be either digital (e.g., software, workflows, protocols, models, etc.) or physical (e.g. new materials, antibodies, reagents, samples, etc.).

In addition to the MDS data, other output data will be produced, and MDDB will use the same FAIR approach e.g., to code developed as the simulations deposited. We already have a track record where e.g., each patch release of key software has a patch-specific DOI tied to non-mutable Zenodo repositories. Table 2 summarises the type of data, its format and how it will be managed according to the FAIR principles. Most of the other research outputs could be stored on the MDDB website. Molecular dynamics is a widely used method in scientific research. It is often linked to scientific publications. However, due to its complexity and the size of the data generated, it is difficult to share the data with reviewers and to reproduce it. The MDDB project will bring a lot to the various studies involving simulation data by creating standards and making the MDS FAIR.

Table 2: Other research outputs

DATA TYPE	DATA FORMAT	MANAGEMENT
Code	Text/Script	Public code repositories and released versions deposited with DOIs.
Analysis	JSON/BSON	Associated to the corresponding MDS and code on github/gitlab
Documentation on best practices for analysis of MDS data	Rich Text	On the website of the project and RDMkit
Documentation on best practices for generation of MDS data	Rich Text	On the website of the project and RDMkit
Scientific article	PDF	Managed by journal associated to a DOI
Meetings	Agenda/Report/Record	On the website of the project
Workshops	Book/Record	On the website of the project
Conferences	Book/Record	On the website of the project
Flyers	PDF	On the website of the project
Tweet	Text	Tweeter
LinkedIn post	Text	LinkedIn
Outreach documents	PDF	On the website of the project

5 Allocation of resources

What will the costs be for making data or other research outputs FAIR in your project (e.g., direct and indirect costs related to storage, archiving, re-use, security, etc.)?

According to the grant proposal, the infrastructure cost will be 250 000 €.

How will these be covered? Note that costs related to research data/output management are eligible as part of the Horizon Europe grant (if compliant with the Grant Agreement conditions)

It will be covered by the grant.

Who will be responsible for data management in your project?

BSC, KTH and EBI-EMBL will be responsible for data management.

How will long term preservation be ensured? Discuss the necessary resources to accomplish this (costs and potential value, who decides and how, what data will be kept and for how long)?

IRB and NBD will seek the advice of subcontractors to leverage the technical and legal requirements defined in previous work and draw a plan for long-term financial sustainability. Aspects like the legal status, policy of the participation in research projects, opportunities of economic return from data analysis services, consultancy or training will be evaluated.

IRB together with all partners and subcontractors will lead the design for the MDDB governance structure. Taking as reference other large-scale repositories and infrastructure, the necessary components will be defined, and their activities and procedures outlined. At this point components like a Steering committee, a Management board, Technical and Data Management boards can be envisioned. We will follow the engagement activities to ensure representation of communities in the governance structure. All partners will work together to create a strong community engagement strategy.

6 Data security

What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?

Will the data be safely stored in trusted repositories for long term preservation and curation?

MDDB is envisioned in the long term as a distributed repository, as the volume of stored data is expected to be significant, including even enabling users to link to locally stored data sets if they are too large or otherwise not possible to deposit. Institutions holding MDDB data will be mainly those with the technical resources to produce large scale simulations, typically HPC facilities. MDDB security will be based on the actual security protocols accepted in those infrastructures for their normal activities.

To ensure end-to-end data integrity, the project will use user-exposed hash summation of all data, and also investigate the usage of public key digital signatures to be able to guarantee the origin and provenance of data. For data recovery, MDDB will initially use backups, but also start work on investigating the usage of distributed replication and/or erasure coding over multiple sites, which would enable data to be continuously available even if one site is offline or has undergone a storage disaster.

For concrete data safeguarding, taking BSC as a representative example, usual procedures include: physical access to the facilities is restricted to operations personnel; user access to any computational resource is based on standard SSL protocols; user credentials are always personal with full identification of the user's identity and signed legally-bound agreements. For web-based resources (e.g., analysis environments), authentication protocols based on OpenID Connect and Oauth2 are in place, and all transactions are made through HTTPS. In terms of network connectivity, HPC and data holding computers have limited outbound connectivity, all network transactions should be initiated from the outside, through the indicated SSL based protocols. Specific network routes are set on a per-service basis to link HPC and cloud resources. Database systems are also isolated, and accessed only by operations

personnel, all data transactions are made through programmatic interfaces. Stored data have periodic full backups allocated in an HSM tape-based system, and for critical projects, a second set of complete backups are kept in a different center within the Spanish HPC network. It is worth noting as a reference that BSC is maintaining, for instance, a full copy of the European Genome-Phenome Archive, holding highly sensitive data like health-related genomics data. Experience of MDDB partners (KTH, BSC holding supercomputing facilities, and EMBL-EBI the most relevant institution in maintaining biological databases) will be leveraged to define MDDB security standards.

7 Ethics

Are there, or could there be, any ethics or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

We do not anticipate ethical or legal issues with data sharing. In terms of ethics, since this is generated data, originally based on publicly available structures/proteomics data, there is no need for an ethics committee.
However, beyond traditional ethical considerations, the project will take a far-reaching responsibility to ensure users worldwide have equal opportunities to provide suggestions and requests e.g. for definitions of good practice in MD and metadata standards, and all produced libraries for data interfaces will be made available under highly permissible licences (BSD) that enable inclusion in all codes, including commercial ones, to avoid creating unfair privileges for certain groups of users.

Will informed consent for data sharing and long-term preservation be included in questionnaires dealing with personal data?

Yes, all users will be informed about our GDPR rules and data licences. The only personal data that will potentially be stored is the submitter name and affiliation in the metadata for data. In addition, personal data will be collected for dissemination and communication activities using specific methods and procedures developed by the MDDB partners to adhere to data protection.

8 Other issues

Do you, or will you, make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones (please list and briefly describe them)?

Yes, theMDDB will use common Research Data Management (RDM) tools. The repository will be submitted to the Core Trust Seal group and will get a certificate. It will be registered on Re3data and FAIRsharing to allow users or other repository creators to get all information about our data management, standards and best practices. Finally, we will be active in different communities such as Global Biodata Coalition and Research Data Alliance.

9 Conclusion

In a first version, this document presents the different methods, approaches and ambitions of the MDDB project concerning data management. It may evolve over time to clarify or detail the different approaches, or adapt them to the state of the art.