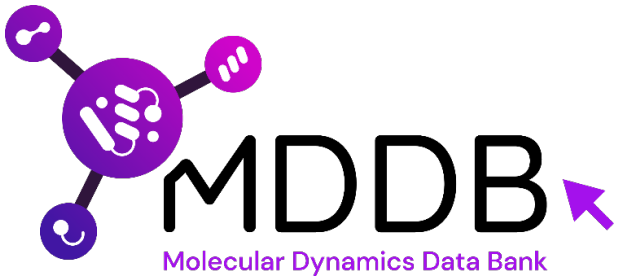

Project Acronym: MDDB
Project title: Molecular Dynamics Data Bank. The European Repository for Biosimulation Data
Call: HORIZON-INFRA-2022-DEV-01
Topic: HORIZON-INFRA-2022-DEV-01-01- Research Infrastructure Concept Development
Project Number: 101094651
Project Coordinator: Institute for Research in Biomedicine (IRB Barcelona)
Project start date: 01/03/2023
Project end date: 28/02/2026

Deliverable 4.3: Preliminary report on sustainability models for the MDDB infrastructure

Work Package: WP4 – Implementation, sustainability and community engagement
Lead beneficiary: IRB-CERCA
Dissemination level: PUBLIC

Due date: 31/03/2024
Actual submission date: 04/04/2024

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	Contributor(s)	Partner	Date	Comments
0.1	Anna Montras, Lejla Bektic	IRB-CERCA	19/02/2024	First draft
0.2	Adam Bellaiche, Preeti Choudhary	EMBL-EBI	26/03/2024	Second iteration
0.3	Adam Hospital, Daniel Beltran	IRB-CERCA	27/03/2024	Revision and contribution to the document
0.4	Sameer Velankar	EMBL-EBI	29/03/2024	Revision and contribution to the document
0.5	Josep Lluís Gelpí	BSC-CNS	30/03/2024	Revision and contribution to the document
0.6	Erik Lindahl, Magnus Lundborg	KTH	31/03/2024	Revision
0.7	Adam Bellaiche, Anna Montras, Lejla Bektic	EMBL-EBI IRB-CERCA	03/04/2024	Incorporated feedback
1.0			04/04/2024	Final version

Table of contents

Executive summary

This report serves as a preliminary examination of sustainability models for MDDB. Section 2 outlines the importance of data policies and strategies essential for integrating MDDB into the life sciences ecosystem. In section 3, we explore the sustainability aspects concerning both software and hardware components, demonstrating the resilience and efficiency of MDDB. Community engagement and collaborations with respect to development and sustainability of MDDB are addressed in section 4. Section 5 investigates the various sustainability models and legal frameworks commonly adopted by projects within the Horizon Europe programme. Through a rigorous analysis, we seek to identify the most suitable model for MDDB, ensuring its long-term viability and impact, aligning its trajectory with the overarching European objectives.

1 INTRODUCTION

Sustainability within research infrastructure encompasses multiple dimensions, including robust data management practices, technical design, community involvement, suitable governance models, and financial viability. We acknowledge the significance of addressing these elements as each one of them plays a crucial role in ensuring the longevity and impact of research initiatives. Our objective extends beyond technical proficiency; we seek to understand the intricate ecosystem in which MDDB aspires to operate and meet the varied needs of the stakeholders across academia and industry setting.

Here we present our preliminary exploration of sustainability models tailored to the MDDB context. We examine components necessary for effective data management, quality control mechanisms, paths towards integration into the life sciences ecosystem, and strategies for software and hardware sustainability. Furthermore, we outline our approach for community assessment and propose options for ensuring financial sustainability.

Our aim is to lay a solid foundation for MDDB that not only supports cutting-edge research but also fosters multidisciplinary collaborations, innovation, and provides societal benefit. Our commitment remains to excellence and responsible resource management.

2 ESTABLISHING DATA MANAGEMENT STANDARDS

2.1 Data deposition process and release policies

We have established a comprehensive data management plan (Deliverable 4.1) to provide guidance for the development, implementation, support, and upholding of MDDB. Regarding its data lifecycle, the data will undergo a continuous process starting with data collection/deposition, curation, quality check, release, integration, updates, archiving, back-up, dissemination, and, in some rare cases, deletion (metadata will still be findable and accessible).

2.2 FAIR data

Respecting the FAIR principles is fundamental for ensuring the sustainability of a biological database. We have already addressed these principles in Deliverable 4.1, and our commitment to upholding them will remain throughout the growing and the development of MDDB.

However, there are two crucial aspects in the early steps that we are studying:

1. The elaboration of a back-end for the federated infrastructure and process to ensure the security and integrity of the data stored within MDDB nodes: An initial proposal for both long-term and analysis-ready storage is being prepared and tested on initial pilot installations.
2. Collaboration with the Physical Sciences Data Infrastructure (PSDI) project[1] to ensure FAIR compliance from data generation to data deposition and sharing. Responsible for developing AIIDA-GROMACS[2], a tool designed to track the provenance of molecular dynamics (MD) simulation data, PSDI is a central point of making MDDB a FAIR database. The project is also working on a community-agreed way to include provenance records in all simulation output files. Starting with the GROMACS[3] MD engine (Deliverable 1.1), the goal for most of the existing MD tools is to follow this path and integrate provenance data in the generated output files.

[1] https://www.psdi.ac.uk
[2] https://aiida-gromacs.readthedocs.io/en/latest/
[3] https://www.gromacs.org/

We will provide guidelines to our users on how to create and generate MD simulation data almost ready to be deposited using the mentioned tools to ensure FAIR data. Having explicit processes for data generation, tools to track data provenance and back-up infrastructure will contribute to the sustainability of MDDB. Comprehensive documentation will serve as guidelines for developers, stakeholders, contributors, and users facilitating understanding, maintenance, updates, deposition and utilisation.

2.3 Data quality

Through robust repository processes and release policies that include data assessment processes, MDDB aims to provide quality-assured data. This is an essential step in making MDDB trustworthy, credible and more widely used by the community.

To achieve this goal, two axes are considered: (1) automated checks on the data using precise in-house tools or (2) manual data curation. By combining automated verification with well-established curation processes, we aim to maintain the highest standards of data integrity and usability within MDDB.

A list of checks has already been established: i) topology and trajectory coordinate matching; ii) periodic boundary conditions or imaging problems; iii) topology problems (e.g. incorrect number of bonds); iv) Root Mean Square deviations (RMSd, RMSd per residue, pairwise RMSd); v) Radius of Gyration (Rgyr); vi) atomistic fluctuations; vii) Principal Component Analysis (PCA); and viii) Solvent-Accessible Surface Area (SASA) (Deliverable 2.1). Additionally, a list of extra checks including chain, residue and atom uniqueness, bond distances, periodic boundary conditions, and more, will be used to identify unusual behaviours, in which case the gathered information will be communicated to the authors who should confirm if they were expected due to the type of simulation. Future tests will be implemented as accepted by the community.

2.4 Integration into the life sciences ecosystem

Integrating data from MDDB into the life sciences ecosystem and vice versa is crucial for enhancing its impact and ensuring its continued relevance and utility. By providing access to integrated data, MDDB promotes collaboration and knowledge sharing within the scientific community, leading to a cycle of contributions, updates and improvements as well as facilitating informed research decisions and driving innovation.

2.4.1 ELIXIR

MDDB aims to establish itself as a trusted repository in the community for molecular dynamics simulation data and become integrated into the life sciences ecosystem. Within this ecosystem, various infrastructures must be considered, with our primary target being ELIXIR[4].

ELIXIR serves as a vital infrastructure facilitating the coordination and development of life science resources across Europe. It offers essential guidelines and best practices for data management, and software development. Furthermore, it has a set of stringent criteria for incorporation in its Core Data Resources list[5], including key indicators, governance structures, and processes. Being part of this curated list would enhance MDDB's credibility, utilisation, integration, dissemination, and support sustainability within the life science community.

[4] https://elixir-europe.org/
[5] https://elixir-europe.org/platforms/data/core-data-resources

2.4.2 Global Biodata Coalition and Research Data Alliance

Once MDDB is included in the ELIXIR Deposition Database and the Core Data Resources list, our next objective is to align with the Global Core Biodata Resources (GCBRs)[6]. This coalition serves as a worldwide alliance with dual aims: (1) facilitating discussions among funders of biodata resources to enhance coordination, sharing, and development approaches, and (2) ensuring stable and sustainable financial support. Additionally, it sets data management standards to maintain the excellence of biological resources.

Joining GCBRs would not only reinforce MDDB's data management, governance, and processes in line with global standards, but it would also amplify its dissemination on a global scale. This move has the potential to position both MDDB and Europe as leaders in the field of storing and sharing molecular dynamics data.

Similarly, MDDB aspires to become a part of the Research Data Alliance (RDA)[7], a pivotal organisation within the research community, dedicated to raise social and technical connections to facilitate open data sharing. At its core, RDA envisions a future where researchers and innovators openly exchange data across diverse technologies, disciplines, and geographical boundaries to tackle society's most pressing challenges. As an example, the BioExcel-CV19 database[8], a database of molecular dynamics related to the COVID-19 structure data, has been helpful in supporting efforts to understand SARS-CoV2 protein function. As MDDB will contain data valuable for development of new biological therapies and advancement of personalised medicine, it has a potential to become an important platform in medical and pharmaceutical innovation and discoveries.

2.4.3 Other infrastructures and labels

Following the initial phase of becoming integrated into ELIXIR and GBCRs, we aim to get the CoreTrustSeal[9] certificate, thus establishing MDDB as a sustainable and trustworthy data infrastructure that meets the mandatory criteria. This certification remains valid for a period of three years encompassing 16 requirements. Obtaining it not only enhances credibility but also ensures that the database remains aligned with current standards, with periodic updates every three years.

Furthermore, in addition to adhering to established standards, MDDB emphasises openness, accessibility, knowledge/experience transfer and transparency regarding its standards and processes. To achieve this, we have identified three key infrastructures:

- **Re3data**[10]**:** A global registry of research data repositories/databases spanning various academic disciplines. It facilitates permanent storage and access to datasets for researchers, funding bodies, publishers, and scholarly institutions. Re3data supports a culture of data sharing, increased accessibility, and improved visibility of research data.
- **FAIRsharing**[11]**:** Serving as a database of databases, FAIRsharing provides users with comprehensive insights into data standards, best practices, and policies associated with each database, utilising an intuitive graph view. This tool is instrumental in ensuring effective management and comprehensibility of every repository. Additionally, as a FAIR website, it assigns a DOI to each database, enhancing its discoverability and citability.

[6] https://globalbiodata.org/what-we-do/global-core-biodata-resources/
[7] https://www.rd-alliance.org/
[8] https://bioexcel-cv19.bsc.es/
[9] https://www.coretrustseal.org/
[10] https://www.re3data.org/
[11] https://fairsharing.org/

- **RDMKIT (Research Data Management toolkit for Life Sciences)**[12]**:** This toolkit offers best practices and guidelines for achieving FAIR data in various domains within the life sciences. It operates on a collaborative basis, allowing contributions from diverse stakeholders. With first contribution already made by MDDB members, contributing to the simulation data section of RDMKIT is a noteworthy accomplishment, further enriching the resources available within the life sciences community thus spreading and transferring the knowledge and experience acquired during the construction and the life of MDDB.
- Create an MD ontology taking inspiration from AMBER[13] and GROMACS ontology terms from EDAM[14] (BioExcel).

2.4.4 Current progress

Until now, we have worked on the data management plan that aligns with FAIR principles, established data and metadata formats, listed ontology and analysis for integration, and written a first draft for release policies. In parallel, we have studied the standards and requirements from ELIXIR, GBCRs and Core Trust Seal. Use cases (WP3) have been used to study required metadata information regarding distinct molecular systems and simulation methods. Missing EDAM ontology terms have been identified and will be defined and contributed to the next releases. We are working on a new community-agreed proposal to standardise MD data exchange formats (WP1) with a highly efficient trajectory coordinate compressor that includes simple system specifications (such as atom/residue names and connectivity) and key-value trees storing high-level and full simulation settings metadata.

One of the proposed MDDB use cases, the database of COVID-19 related MD simulations, has been published and documented in FAIRsharing. The database contains more than 10K MD simulations, divided into 11 SARS-CoV-2 protein units including the well-known Spike, Angiotensin Converting Enzyme 2 (ACE2) and Receptor Binding Domain (RBD), with a total accumulated time of more than 10 milliseconds. This project has been fundamental for the identification of missing requirements, especially regarding the distinct MD data types (trajectories, ensembles, replicas) and how to integrate them into a single infrastructure in a seamless way. Encountered challenges have already been faced by the BioExcel-CV19 project, and will be the starting point for the integration into the MDDB prototype (Deliverable 3.1).

The next step is to cross correlate all the standards and requirements collected in the first phase of the project and work on their implementation into the first MDDB prototype.

3 INFRASTRUCTURE

3.1 Software sustainability

3.1.1 Version control

Regarding the MDDB proposed use cases and prototype, software version control is handled by the git version control system. Some code repositories are already open like the MD data processing and analysis workflow[15]. Other repositories such as the different web clients or the REST API are not yet public since they are still undergoing rapid changes and have not been cited in a publication.

[12] https://rdmkit.elixir-europe.org/
[13] https://ambermd.org/
[14] https://bioportal.bioontology.org/ontologies/EDAM?p=classes&conceptid=format_3880
[15] https://mmb.irbbarcelona.org/gitlab/d.beltran.anadon/MoDEL-workflow

3.1.2 Third party software dependencies

MD data processing requires the use of several third party software. The current workflow relies on state of the art tools that are also a standard in the field: Gromacs, MDtraj[16], PyTraj[17], MDAnalysis[18], VMD[19], scikit-learn[20], etc. These tools also support generic analyses while more specialised analyses are usually performed by additional dedicated software:

- Electrostatic and Van Der Waals energies between two interacting molecules are calculated with an in-house tool called CMIP[21].
- Conformation of canonical B-DNA molecules is analysed by Curves+[22].
- Binding pockets are dynamically tracked by MDpocket[23].

While most data processing and analysis is run in Python, data-sharing software is programmed mostly in JavaScript (JS). Use case databases are based in MongoDB[24], which is natively handled in JS. The REST API is made within the Express[25] framework and kept alive with the process manager PM2[26]. The web client is made with React[27], which combines JS with HTML and CSS.

3.1.3 Visualisation components

In terms of visualisation, MDDB aims to use a modern web-based, lightweight, open-source visualisation tool Mol*[28], where PDBe has created PDBe-Mol* wrapper to make it easy to use. The objective is to provide to MDDB users with: (1) a remote/local way to visualise trajectories, (2) a way to run analysis, (3) an overview of metadata highlighted on 3D structure (from MDDB and from data integration) and (4) explore the results in one place.

Using an external resource such as Mol*, Unitymol3D[29] or VTX[30] makes MDDB dependent on the continued availability and maintenance of these resources by their developers but at the same time, it could enhance its sustainability. Mol* is widely used in wwPDB databases, UniProt, Ensembl, Interpro and others biology databases and it is becoming a standard in visualisation. Creating customised Mol* functionalities tailored for MD analysis could further solidify its status as a standard tool in the field. Given the lack of consensus in MD visualisation methods currently, this presents an opportune moment to establish one and thus contribute to the sustainability of MDDB.

3.1.4 Unit testing and continuous integration

Some parts of the code are supervised by tests that are automatically run when changes are pushed to the remote repository. We plan to further establish more tests to ensure a stable deployment. The COVID-19 database has a development version that allows us to test changes in both the REST API and the web client before we deploy them in the production service.

[16] https://www.mdtraj.org
[17] https://amber-md.github.io/pytraj/latest/index.html
[18] https://www.mdanalysis.org/
[19] https://www.ks.uiuc.edu/Research/vmd/
[20] https://scikit-learn.org/stable/
[21] https://mmb.irbbarcelona.org/gitlab/gelpi/CMIP
[22] https://doi.org/10.1093/nar/gkp608
[23] https://doi.org/10.1093/bioinformatics/btr550
[24] https://www.mongodb.com/
[25] https://expressjs.com/
[26] https://pm2.keymetrics.io/
[27] https://react.dev/
[28] https://molstar.org/
[29] https://glycopedia.eu/resources/article-presentation-2/
[30] https://vtx.drugdesign.fr/

3.2 Hardware sustainability

3.2.1 Federated database infrastructure, computational resources, back-up and security

MDDB is conceived as a federated infrastructure. This implies that most of the computational burden required will remain in the contributing data nodes, while central infrastructure will act as a relay allowing users to access data and metadata in uniform manner but requiring just a fraction of the computational resources of the complete infrastructure. The necessary agreements to enrol MDDB nodes will assure the commitment of the participating data centres regarding the assignment and maintenance of the necessary computational resources to the project. MDDB will rely on the standards on security and procedures of the participating centres.

3.2.2 Authentication and authorization infrastructure (AAI)

Although most of the data available at MDDB will be on the public domain, an authentication and authorization infrastructure will be provided. To assure long-term sustainability. AAI will be based on standards widely accepted on the Life Sciences domain (OpenID Connect), and will accept LS Login as identity provider, which in turn relies on the researcher's institutional authentication protocols.

4 COMMUNITY ENGAGEMENT

4.1 Key stakeholders

MDDB aims to develop a database that harmonises all existing efforts aimed at storing and sharing MD simulation data.

The following stakeholders will be essential at different stages and for different aspects of development of the MDDB infrastructure (namely design, prototype testing, and eventual scale-up and operation as a pan-European infrastructure):

- MDDB Pilot cases: Work Package 3 leverages the experience of five different dataset use cases, bringing in technical requirements from different types of datasets.
- MD code developers: their involvement in the co-creation process is essential to reach consensus on standards and interoperability between codes.
- Infrastructure providers: dialog with representatives from other key infrastructures to ensure interoperability and proper integration with other databases.
- National and EU-wide policy-makers: support for future development, funding, etc.
- End-users, including both academia and industry, who will interact with the future infrastructure either uploading or downloading data, need to be thoroughly characterised and consulted on their needs.

During the first year of the project, efforts have been focused towards defining the strategy for the validation of the widespread need for the MDDB infrastructure and setting up the tools to engage all necessary stakeholders.

MDDB's engagement strategy, developed under WP4 (Deliverable 4.2), identifies the key activities that the project will undertake to (i) actively involve key stakeholders in the co-design of the infrastructure and (ii) characterise the potential end-user community (Fig. 1).

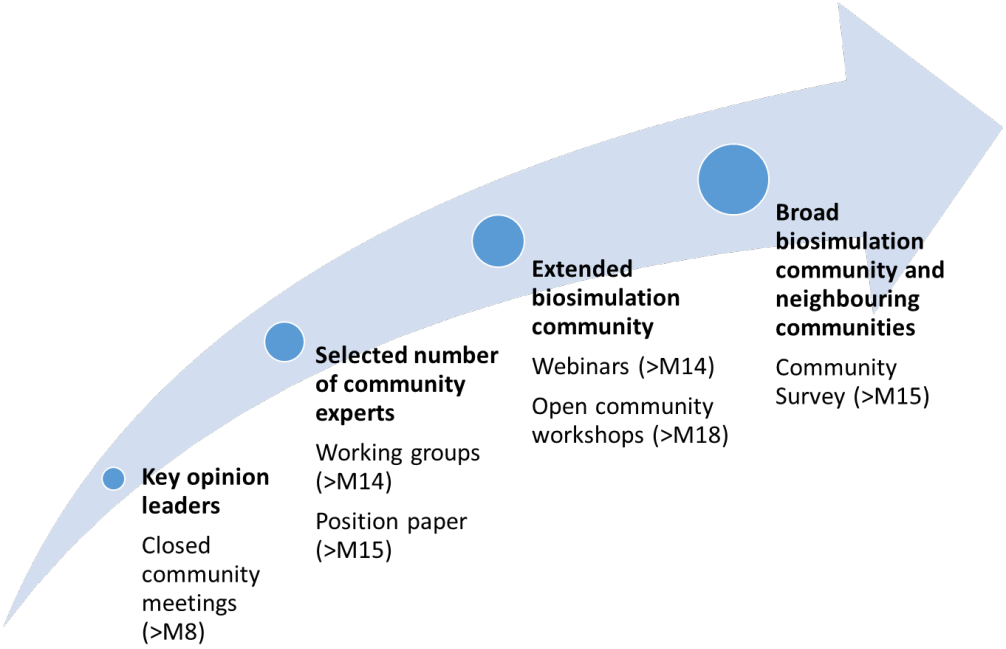

Fig. 1: MDDB's approach towards (i) collecting requirements and (ii) characterising the end user community. The diagram represents the involvement of representatives from the different stakeholder segments.

4.2 Design and co-creation of the MDDB infrastructure

Involving all relevant stakeholders in the development of MDDB since its conceptual design will be essential to facilitate its long-term acceptance and sustainability as a European (and ultimately global) research infrastructure. The following tools and actions have been defined with the aim to engage the key stakeholders in the co-creation (incl. design and validation) of the MDDB infrastructure concept.

- MDDB has started to organise **closed community meetings**, in which the consortium invites key stakeholders to take part in discussions regarding data management, policies, etc., and provide feedback on the infrastructure.
    - The first community meeting (Oxford, UK, October 2023) brought together representatives of the key MD biosimulation codes (11 participants) and MDDB members (10 participants) to seek consensus on MD standards, formats and data interoperability.
    - Three additional topics will be discussed in high-profile community meetings: (i) Best practices and metadata in MD simulations (the meeting will be co-located with a BioExcel gathering of key stakeholders in the biosimulation community in October 2024), (ii) Structural databases in Exascale and (iii) bridging the gap between communities in MD data sharing.
- **Working groups**: One of the outcomes of the first community meeting (Oxford, UK, October 2023) was the definition of a series of working groups in which external community members will be invited to participate to enhance the community and stakeholder engagement. The set-up of the working groups is foreseen between M13 and M18. The following working groups have been identified: (i) Data quality and QC analysis, (ii) Definition of new potential file format, (iii) Data integration, (iv) Technical installation and (v) engagement and sustainability.
- A **position paper** signed by a number of key opinion leaders in the field of molecular simulation is in preparation.

4.3 Characterization of end-user communities

The segmentation and size assessment of the community is a necessary step towards defining the potential sustainability models for the MDDB infrastructure. Starting in M13, efforts devoted to characterising the composition and size of the end-user community will be initiated, which, together with the outcomes of WP2, will be essential to outline the most suitable business model for the operation of the MDDB infrastructure and to pursue the necessary national supports for the inclusion of MDDB in the respective roadmaps.

- **Community survey**: Building on the insights gathered from the closed community meetings, we will design and conduct a survey to gather input from a broader audience, including researchers, practitioners, and other stakeholders. The survey will aim to answer a number of key questions, such as, the size of the community, what are the requirements from different community groups, list of services expected to be offered by MDDB, etc.

  We will utilise a multi-channel approach to distribute the survey, starting with hosting it on the MDDB website with the information on its objectives and instructions for participation. MDDB consortium members and supporters will be encouraged to share the survey through mailing lists and social media. Related communities such as ccp-biosim, pdb-l, ELIXIR 3D-bioinfo, etc. will be instrumental to ensure this multiplying effect. We will encourage community members contacting us through the website to complete the survey and we will use X (Twitter) and LinkedIn to reach a broader audience. Additionally, we will promote it during relevant conferences and workshops, and leverage our professional and collaborator networks (BioExcel, GROMACS, EMBL, etc) to distribute the survey.

- **Webinars and open community workshops**: We are organising a workshop titled "Pushing the frontiers of molecular dynamics simulations"[31], planned for October 7-9, 2024 in Lausanne, Switzerland, where 23 experts in the field together with a broader community will jointly set out a community roadmap for key issues to work on.

5 GOVERNANCE NEEDS

The conclusion from the community characterisation will be essential to define different scenarios for the exploitation and governance of the resulting infrastructure.

Variables such as the size of the community (including the validation of the interest of other communities beyond the biosimulation field), the criteria for the installation of the infrastructure nodes (e.g. by type of data, by country, etc.) and other variables that will arise from the activities undertaken as described in section 4 above, will define the actors that need to be involved in the governance. Such actors will also pose the constraints with regards to the exploitation route.

6 FINANCIAL STABILITY

6.1 National roadmaps

Engaging with national roadmaps of the different Member States (MS) and Associated Countries (AC) targeted as hosts of the nodes of the distributed infrastructure is imperative within the expansive framework of European research infrastructures. These roadmaps hold a pivotal role in shaping investments and priorities, providing insights into scientific needs, technological capabilities, and socio-economic priorities at national levels. Recognition at this level not only paves the way for

[31] https://www.cecam.org/workshop-details/pushing-the-frontiers-of-molecular-dynamics-simulations-1275

inclusion in the European Strategy Forum on Research Infrastructures (ESFRI)[32] roadmap but also opens the door to European funding and resources and validates MDDB's contributions to the wider scientific community. To this end, the efforts are currently being directed towards engaging with key stakeholders that will unlock the access to such support from the MS governments.

6.2 The ESFRI roadmap

Aligning with the ESFRI roadmap is crucial for research infrastructure projects aspiring to integrate into the European research landscape, thus increasing access to funding opportunities. We have already begun considering the requirements and gathering the necessary information for MDDB's application. The publication of the next ESFRI roadmap, although not yet confirmed, is anticipated around 2026. By then, we aim to secure official support from three MS and/or AC, namely, Spain, where MDDB's statutory seat and one of the initial nodes will be located, along with Sweden and the UK, where other initial nodes will be established. Additionally, we are actively exploring collaborations with supercomputing centres in other countries, aiming to gather their national support as well. Expanding the support network to other countries beyond the drivers of the MDDB project is essential to obtain a multiplying effect in terms of future funding sources.

6.3 Legal entities

The majority of research infrastructures enter the ESFRI Roadmap without a defined legal entity and often adopt one once they achieve landmark status. Understanding the importance of a solid legal foundation early on, we have actively examined several potential frameworks for MDDB. Moving forward, we are committed to selecting and implementing the most suitable legal entity that aligns with MDDB's objectives and needs, securing its long-term success and sustainability.

- **Consortium agreement under legal entity of one of the partners:** This is an interesting prospect for MDDB due to its cost-effective, rapid and straightforward establishment process. It offers a high degree of flexibility in governance structures, customisation of terms and conditions, as well as adaptability, allowing for seamless adjustments of terms and structures in response to evolving needs, priorities, and circumstances. The potential challenge may arise in coordinating efforts, setting up governance structures, and ensuring enforceability, especially when compared to more formal legal entities such as ERICs and EDICs.
- **International non-profit association (Association Internationale Sans But Lucratif or AISBL):** Governed by Belgian law, it offers notable advantages such as legal recognition, international scope, and flexibility in membership[33]. Despite its merits, the AISBL presents certain drawbacks, such as the requirement for the head office to be located in Belgium, which may pose logistical challenges considering that none of the current MDDB partners is based in Belgium. Additionally, adherence to Belgian law may introduce complexities related to legal jurisdiction and law interpretation, particularly given MDDB's multi-country operations with varying legal systems across jurisdictions.
- **European Research Infrastructure Consortium (ERIC):** It offers many advantages, including legal recognition and protection, facilitation of cross-border collaboration, increased international visibility and impact, and improved long-term sustainability[34], making it an appealing option for MDDB. By attaining ERIC status, MDDB would showcase its excellence

[32] ESFRI. "Strategy Report on Research Infrastructures Roadmap 2021 Public Guide". September 25, 2019

[33] Service Public Fédéral Justice. "AISBL". Accessed on February 27, 2024. https://justice.belgium.be/fr/themes_et_dossiers/societes_associations_et_fondations/associations/aisbl

[34] Directorate-General for Research and Innovation (European Commission). "ERIC Practical guidelines, Legal framework for a European Research Infrastructure Consortium". March 2015.

and leadership in the field, attracting attention from researchers and stakeholders. This achievement would pave the way for new collaboration opportunities, enhancing MDDB's influence and relevance within the scientific community.

- **European Digital Infrastructure Consortium (EDIC):** As a newly proposed legal framework by the European Commission aimed at supporting the development of digital infrastructures in Europe, it presents an appealing option for MDDB due to its simplified governance tailored for digital infrastructures. By attaining EDIC status, MDDB would demonstrate its dedication to digital transformation, open data sharing, and impactful solutions for scientific and societal challenges, aligning with the Decision (EU) 2022/2481[35]. Although we are diligently working on expanding our network and operational reach across multiple Member States, we are still one member short of meeting the national backing criterion for EDIC accreditation. Nonetheless, we are actively addressing this challenge and remain committed to fulfilling all necessary requirements for the success of MDDB.
- **Company:** Establishing MDDB as a company would provide flexibility in decision-making and resource allocation, enhancing responsiveness and effectiveness in delivering value to stakeholders. It would facilitate engagement with the private sector and industries, potentially leading to industry-driven research activities, technology transfer opportunities, and commercialisation, and thus improving financial sustainability. However, transitioning to a company structure may divert focus and resources from research activities to revenue generation and pose conflict of interest. Careful consideration of these factors is essential before proceeding with the decision.

6.4 Beyond Europe

MDDB is conceptualised as a European repository for biosimulation data but has garnered interest from institutes outside of Europe since the start of the project. We are currently exploring collaboration opportunities with these institutes, considering their potential incorporation as nodes within MDDB. Such collaboration would bolster MDDB's sustainability through additional funding and offer Europe the chance to lead in global data management practices and guide them towards the common goal while upholding European values. As research is a global endeavour, MDDB will continue to explore global collaboration to guarantee its sustainability and openness.

7 CONCLUSIONS

Effective data management is fundamental, covering the entire lifecycle of data from collection and deposition to curation, integration, archiving, backing up, and in rare cases even deletion. Adhering to FAIR data principles and best practices throughout this lifecycle, MDDB aims to establish itself as a trusted resource across various scientific communities, promoting data reuse and minimising resource waste. To this end, integration into the broader life sciences ecosystem is imperative for MDDB's continued relevance and utility. We are committed to integrating MDDB with renowned initiatives such as ELIXIR, Global Biodata Coalition, and Research Data Alliance, and pursuing CoreTrustSeal certification to enhance MDDB's credibility and stay up to date with the current data repository standards.

Technical sustainability is equally crucial for a research infrastructure, both in terms of software and hardware design. We are building MDDB as a federated database, accessible through one central node

[35] Decision (EU) 2022/2481 of the European Parliament and of the Council of 14 December 2022 establishing the Digital Decade Policy Programme 2030 (Text with EEA relevance), 2022 O.J. (L323) 4.

with data distributed across multiple additional nodes, employing open-source software for data processing and analysis, supplemented by well-established third-party tools.

Community engagement serves as the cornerstone of our efforts, guiding the design and development of MDDB to meet the diverse needs of the stakeholders. Through targeted engagement with key opinion leaders and broader community members via closed meetings, working groups, webinars, and open workshops, we seek to foster collaboration, gather feedback, and build consensus on data management policies and infrastructure development. Currently, we are focusing our efforts on the biosimulation communities, but we expect to assess the interest and extend our collaboration with other communities, such as material sciences.

In terms of financial sustainability, alignment with national and ESFRI roadmaps is essential. We are also exploring various legal frameworks and aim to select the most suitable one for MDDB's long-term financial sustainability. Furthermore, engagement with stakeholders outside of Europe presents opportunities for additional funding while providing Europe with global leadership in data management practices.

In summary, our overarching goal is to establish MDDB as a resilient, impactful, and globally recognised research infrastructure, driving innovation and collaboration for the betterment of scientific endeavours in Europe and beyond.