MAGIC

Next-generation models and genetic therapies for rare neuromuscular diseases

Grant agreement No. 101080690

D8.2 Data Management Plan

Lead partner	DDF
Other contributors	UCL, INSERM, MHH, UPEC, NUIM, BIOND, VIVE, DNMQS, CRICK, KCL

Dissemination level	Public (PU)
Contractual date of delivery	M6, 30/11/2023
Actual date of delivery	M7, 01/12/2023
Work Package	WP8
Type	Report
Version	V1.0

Copyright

Disclaimer

Acknowledgment of funding

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 and HADEA.

This work is funded by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee grant numbers 10080927, 10079726, 10082354 and 10078461.

This work has received funding from the Swiss State Secretariat Funding for Education, Research and Innovation (SERI).

History

Version	Date	Description	Reviewers
V0.1	2/10/2023	Draft for internal review	Francesco Muntoni (UCL), Karen English (NUIM)
V0.2	3/11/2023	Addressing Comments	Paraskevi Sakellariou (DDF)
V0.3	10/11/2023	Pre-final version	Mario Amendola (INSERM), Eduard Ayuso (DNMQS), Marie Fertin (VIVE), Angelo Raggiolo (RT)
V1.0	01/12/2023	Submission to EU	

Author list

Organisation	Name Surname	Email
DDF	Paraskevi Sakellariou	sakellariou.elvina@duchennedatafoundation.org

Table of Contents

Abbreviations and Acronyms

Acronym	Description
APIs	Application Programming Interfaces
CC 0	Commons Public Domain Dedication
CC BY	Creative Commons Attribution International Public License
DAC	Data Access Committee
DCAT	Data Catalog Vocabulary
DDF	Duchenne Data Foundation
DMP	Data Management Plan
DOIs	Digital Object Identifiers
EB	Executive Board
EGA	European Genome-phenome Archive
ENA	European Nucleotide Archive
FAIR	Findable, Accessible, Interoperable and Reusable
FDP	FAIR Data Point
GEO	Gene Expression Omnibus
hiPSC	Human induced pluripotent stem cell
hPSCreg	Human Pluripotent Stem Cell Registry
IP	Intellectual Property

IPC	Intellectual Property Committee
MAGIC	Next-Generation Models And Genetic Therapies for Rare Neuromuscular Diseases
PMID	PubMED Identifier
RDF	Resource Description Framework
TTL	Terse RDF Triple Language
URI	Uniform Resource Identifier
VMs	Virtual Machines
WP	Work Package
YAML	Ain't Markup Language

Executive Summary

About the MAGIC project

The MAGIC project addresses a critical challenge in the field of genetic therapies for neuromuscular and musculoskeletal disorders, particularly focusing on severe genetic conditions like muscular dystrophy. The primary issue at hand is the lack of advanced, humanized models for developing effective therapies, hindering the translation of genetic treatments into clinical practice. Currently, treatments for these disorders primarily manage symptoms, with only a few gene therapies or genome editing strategies approved.

MAGIC's primary goal is to pioneer cutting-edge models and therapeutic approaches to overcome these limitations. This involves the creation of innovative models for human skeletal muscle disorders and the development and validation of new gene therapy vectors and genome editing strategies. The ultimate objective is to reduce premature mortality and the healthcare burden associated with these disorders by providing precision medicine tools to healthcare systems.

The project brings together a multidisciplinary consortium from seven European countries, the UK, Switzerland, and the USA, encompassing expertise in muscle development, disease, and regeneration, as well as various related fields such as bioengineering, modelling, viral vector engineering, gene therapy, and patient advocacy.

In summary, the MAGIC project aims to revolutionize the treatment landscape for severe muscular dystrophies and myopathies by equipping healthcare systems with advanced precision medicine tools, emphasizing patient engagement, and fostering knowledge dissemination to drive future advancements in the field.

About this deliverable

This report is the first deliverable of Task 8.3 “Data management” and summarizes the data management plan (DMP) in the MAGIC project, following the Horizon Europe DMP template.

The first version of the MAGIC DMP was designed based on the datasets that will be collected, generated and re-used during the project and followed the EU legislations. The purpose of the DMP is to address the relevant aspects of making data findable, accessible, interoperable and reusable (FAIR), including what data the project will collect and generate, whether and how it will be made accessible for verification and re-use, how data will be interoperable between individuals and institutions and countries as needed, and how it will be curated and preserved. Details on how the data will be shared and any restrictions these may need will be also stated. Data collected and generated by the

project will be placed in a secure database, catalogued and made available for access to consortium partners to analyze when needed, as well as to interested stakeholders for future related projects.

The DMP is a live document that will evolve during the lifespan of the project, to be updated at least for each periodic review. Each updated version will contain a detailed and updated description of the MAGIC datasets and any potential changes in MAGIC Consortium agreed-upon processes and policies related to data management both at the administrative and technical level.

Data Summary

1.1 Will you re-use any existing data and what will you re-use it for?

The types and purpose of existing data that will or may be (re-)used are the following:

- Relevant publication work from the Consortium partners
    - The UCL team at the Institute of Child Health has previously described a number of mutations in the COL6A1, 2 and 3 genes (PMID: 12011280; PMID: 16075202; PMID: 15955946; PMID: 15689448; PMID: 22480491; PMID: 18366090; PMID: 25204870; PMID: 24706943; PMID: 30895940). Including mutations for which RNA therapies could be viable (PMID: 36401040; PMID: 32865794; PMID: 32585628; PMID: 30895940 PMID: 28918041). Further, established immunohistochemical and circulating biomarkers that could be used for monitoring success of therapeutic intervention using genetic therapies (PMID: 32390640; PMID: 26945058; PMID: 25211533; PMID: 24223098; PMID: 22075033).
    - The UCL/Crick team has generated several databases of transcriptional, imaging and other disease-associated cellular phenotypes from different models of muscle diseases, including: PMIDs: 36989620; 36792780; 36161772; 32087527; 30745033; 30405424; 29669293; 29242210; 26042384; 22745439
- GenBank gene and genome sequences for designing and selection of target sides for gene editing and off target activity.
- Genomic and Transcriptomic Annotations coming from the ENCODE data to analyse the transcriptional and epigenetic status of disease genes.

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

The typical types of the data generated and re-used are:

- Imaging data
- Muscle physiology data
- Genetic data
- Genomic data
- Molecular data
- Biochemical data
- Cellular data (re-used)
- Immunocytochemical / protein expression data
- Bioprocessing data

The typical formats of the data generated and re-used are:

- Images will be stored in JPEG or TIFF formats
- Videos will be stored in AVI, MOV, VLC formats
- Genetic data will be stored in APE and DNA formats
- Numerical data will be stored in spreadsheet format (Excel)
- Software/Equipment specific data will be stored in CTL, FCS, ELN and Excel formats

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

The purpose of the data collection, generation and re-use is:

- To generate new knowledge
- To make informed decisions
- To combine with other data
- To share information and data
- To develop a product

This project aims at designing new vectors for safe and efficacious neuromuscular gene therapy with high efficacy and selective tropism coupled with low toxicity and limited immunogenicity. The consortium will develop a limited number of lead candidates that will be validated at preclinical level and will be generating readiness for clinical translation. Collection of data is essential to inform on the design characteristics of these vectors and to determine their biological effect.

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

Kilobytes to gigabytes depending on the types of generated and re-used data.

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

Primary data and secondary data.

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

This data collected might be useful for other scientists working on advanced therapies and bioengineering fields. The data might be useful for pharmaceutical industry if the project succeeds proving the efficacy and safety at preclinical stage of the newly developed vectors for gene therapy and/or genome editing.

FAIR data

2.1. Making data findable, including provisions for metadata.

2.1.1 Will data be identified by a persistent identifier?

MAGIC datasets will be described with rich metadata:

- Metadata is assigned a globally unique and persistent identifier based on standard ontologies.

- Metadata are registered or indexed in a searchable resource (e.g. FAIR Data Point).

2.1.2 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.

To ensure data discoverability, all MAGIC datasets stored in the DDF repository will be described using the Dublin Core Metadata standard. The datasets, when generated, will encompass essential components and contribute to the metadata information model associated with the data:

1. Digital Object Identifiers (DOIs).
2. A custom metadata structure adhering to open linked data standards.
3. For open MAGIC data housed in external public databases, they will be identifiable through a persistent Uniform Resource Identifier (URI).

4. A general metadata schema, based on the Dublin Core Metadata standard, will be employed for datasets created by the project and deposited in the DDF repository. Extensions of this metadata schema, utilizing standard vocabularies, can be implemented for specific MAGIC datasets, as needed. The repository's metadata is exportable in YAML format.

For external public repositories, respective data descriptors will be completed.

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

Search keywords will be provided together with the necessary metadata description for the MAGIC datasets.

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

DDF FAIR Data Point (FDP) is already continuously indexed by https://home.fairdatapoint.org. All the metadata for the MAGIC datasets will be publicly available.

2.2. Making data accessible – Repository.

2.2.1. Will the data be deposited in a trusted repository?

There are two types of repositories to be used:

1. The DDF repository as the central storage of the MAGIC datasets. The DDF Repository includes:

- Private Access Website. Each user will receive an account to be able to operate and share information and datasets (https://repository.duchennedatafoundation.org);
- Private GraphDB Access, the local storage of metadata for FDP (https://graphdb.duchennedatafoundation.org/login) offered using a secured connection;
- Public FDP access (https://fdp.duchennedatafoundation.org) offered using a secured connection.

2. External public repositories (e.g. EGA, ENA, GEO and hPSCreg).

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

For the DDF repository: Yes.

For the external public repositories: Standard procedures will be followed.

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

For each MAGIC dataset deposited in the DDF repository a URI will be assigned within the context of the repository. Metadata will be resolved to DOIs within the context of an FDP publication.

2.3. Making data accessible – Data.

2.3.1 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.

For certain MAGIC datasets publication (see examples in 2.6.2) through the DDF repository and other external public repositories will depend on the data maturity, typically following scientific publications, for utilization by stakeholders beyond the consortium.

The Executive Board (EB), consisting of one representative of all partners, will ensure data is released in accordance with the obligation to disseminate results (Article 17 of Grant Agreement), and that the MAGIC consortium adheres to agreed publication policies and processes while also ensuring that:

- Intellectual Property (IP) issues are considered, by taking advice from the MAGIC Exploitation Manager and Intellectual Property Committee (IPC).
- Publication impact is maximized, which could involve delaying publication.

However, metadata (machine-readable data description) will be made accessible through DDF FDP.

2.3.2 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.

An embargo period will be imposed for certain datasets by the partners who own these and in collaboration with IPC, when needed, respecting any potential confidentiality and IP restrictions. The embargo period will be set during the project's lifetime and at least up to one year after the end of the project- with possibility of extension- to allow proper exploitation of the MAGIC outputs by the partners. Relevant information on the embargoed MAGIC data (metadata), including the details of when the data will become available, will be published through the DDF FDP.

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

MAGIC datasets deposited in the DDF repository will be available through standard Application Programming Interfaces (APIs) and a secure web user interface depending on the user access level control. Each dataset will provide its own metadata structure using standard ontologies and represented through RDF-TTL (Resource Description Framework) formats which will be used to make it openly accessible through DDF FDP. DDF FDP will also provide long-term indexing of the project's metadata even in the case the data is no longer available.

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

During the project, data will become accessible only to authorized users. Access will be restricted to MAGIC partners, with personal accounts, requiring the authentication of a username and a password. Restricted access allows the datasets, and relevant data files, created by individual team members in the MAGIC consortium to become only visible to the owners/providers. Users from the same or different team(s) need to request access to view the data.

Secure logon processes have been developed and adapted for managing overall access control. Data access can also be provided by using private API keys. Requests for secondary use of MAGIC datasets made by stakeholders beyond the consortium will require approval from the Data Access Committee (DAC) respecting confidentiality, IP and contractual (as described in 2.3.6).

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

Each MAGIC consortium partner will hold a unique user account and an API key for data access by which they can be identified.

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

The current composition of the project's EB will reflect the structure of the DAC in order to review and assess potential data access requests beyond the MAGIC consortium. The DAC will ensure that requests for MAGIC datasets re-use will align with the publication policies and relevant data requirements as imposed by the consortium. Appropriate documents will be developed, including Data Access Policy, Data Access Request Form and Data Sharing Agreement. Only users authorized by the DAC will have access to and re-use the MAGIC datasets.

2.4 Making data accessible – Metadata.

2.4.1 Will metadata be made openly available and licensed 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?

Metadata will contain information to enable the user to request access to the MAGIC datasets.

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

The repository and external public repositories (e.g. EGA, ENA, GEO and hPSCreg) suggested for the storage of the MAGIC datasets are by definition long-term repositories facilitating data publication and data sharing, mostly when projects are completed. The MAGIC datasets that are deposited in the DDF repository will remain usable for a minimum of 5 years after the completion of the project.

DDF FDP will provide long-term indexing of the project's metadata even in the case the data is no longer available.

2.4.3 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)?

To access MAGIC datasets, deposited as data files in the DDF repository, no special methods or software tools needed. The repository is based on open-source CKAN

software and is licensed specifically under the terms of the Affero GNU GPL v3.0. GraphDB, an open-source software, will be used for metadata access. Documentation on CKAN and GraphDB is available on the following URLs:

- https://docs.ckan.org/en/2.10/
- https://graphdb.ontotext.com/documentation/10.0/using-the-graphdb-rest-api.html

2.5 Making data interoperable.

2.5.1 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?

Datasets produced in the project will be interoperable using different standards following multiple processes:

- Unstructured data: will be structured to a metadata level (RDF/OWL standard ontologies).
- Data mappings will be provided (in the form of an ontology) for creating dataset metadata structures (mapping further to DCAT models).
- Make data FAIR: specify metadata at the layer of dataset.
- Using generic vocabularies to semantically describe the concepts of the datasets generated, including NCIT, SIO, EDAM, SNOMED, etc.

2.5.2 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?

Currently, MAGIC intention is not to create new specific ontologies or vocabularies.

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

[1]*A qualified reference is a cross-reference that explains its intent. For example, X is regulator of Y is a much more qualified reference than X is associated with Y, or X see also Y. The goal therefore is to create as many meaningful links as possible between (meta)data resources to enrich the contextual knowledge about the data. (Source:* https://www.go-fair.org/fair-principles/i3-metadata-include-qualified-references-metadata/*)*

Meaningful qualified references to the generated MAGIC datasets will be generated in the context of FAIRification of the metadata. Description of and/or complementary information on relevant data to the MAGIC datasets will be collected as described in 2.1.1 and published via the DDF FDP.

2.6 Increase data re-use.

2.6.1 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.)?

A custom metadata schema for each MAGIC dataset will be developed collecting appropriate information in a spreadsheet (Excel) format to validate data analysis and facilitate data reuse.

2.6.2 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?

Data usage licenses defining the access rules for the datasets produced during the project will be determined by the data owners (partners of the consortium). The following types of data have been identified according to their availability outside the scope of MAGIC:

- Private: Available to the data owners;
- Available to MAGIC partners: Not public, rather available solely to partners subject to the terms & conditions set in the Consortium Agreement;
- Available to the wider scientific community and interested stakeholders: Not public, rather available to third parties subject to the specific permission and the underlying terms and conditions as set by the DAC following specific data access requests.
- Open Data: Both public and available under an open license as set out in Article 17 of the Grant Agreement following the principle 'as open as possible as closed as necessary' (e.g. Creative Commons Attribution International Public License (CC BY) or Creative Commons Public Domain Dedication (CC 0)).

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

The intention is that MAGIC datasets will become re-usable by third parties after the end of the project. In the event of particular confidentiality obligations and/or IP issues, access to specific restricted datasets may be granted upon request as described in 2.3.6. The means to ensure legacy data remains accessible and has appropriate governance beyond

the project lifetime will be defined by the MAGIC consortium and implemented before the end of the project.

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

Yes.

2.6.5 Describe all relevant data quality assurance processes.

A data manager will be appointed by each participating institution for the datasets produced. The data manager will be responsible for cleaning and curating the data, as well as uploading data to the repository and external public repositories in collaboration with local data stewards. For preclinical studies, SOPs will be developed and shared among labs. In addition, reproducibility between labs will be assessed. These processes will be followed to evaluate the data quality generated within different WPs. Industry partners of the MAGIC consortium have Quality Management Systems in place to guarantee the data quality assurance process.

Other research outputs

3.1 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.).

MAGIC research outputs, other than research data, will be also subject to FAIR principles workflows as described throughout the document and sharing requirements as imposed by the consortium aligned with the Exploitation Plan to ensure that legal and IP aspects have been addressed properly.

Allocation of resources

4.1 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.)?

The infrastructure (e.g. cloud computing and data storage, secure File Transfer Protocol backup) and maintenance (e.g. updates, data migration) costs associated with the DDF repository are recognized and managed through the project.

4.2 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)

The costs will be covered by the funding grant and the use of the DDF infrastructure.

4.3 Who will be responsible for data management in your project?

Responsibilities for data management could be defined based on the type of data storage:

1. Local Repositories: Local repositories are under the management of the participating institutions, and this will involve local staff and local resources. An appointed data manager should be appointed to manage data produced locally and be the contact person for the communication with the data managers from other sites, the DDF repository, and the external public databases.
2. DDF repository: DDF will provide services for the management of the deposited datasets (e.g. development of metadata schemata based on standard ontologies and metadata catalogues).
3. External public repositories: Third-party repositories (e.g. e.g. EGA, ENA, GEO and hPSCreg) will be responsible for data management, in interaction with the team that generated the data in case corrections are needed. DDF and MAGIC partners will collaborate with these external teams.

4.4 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)?

There is no cost associated with depositing and retrieving data from the DDF repository. The DDF repository will provide persistent storage, primarily for processed data during the lifetime of the project. However, it remains to be determined how the costs of maintaining the MAGIC database in the DDF repository after the completion of the project will be covered.

Data security

5.1 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)?

Regarding data stored in the DDF Repository's database, which includes MAGIC datasets, relevant data files, and metadata, DDF, as the last data processing resort within MAGIC, will ensure through internal policies that persistent or transient data is secured. The Information Systems Security Policy, including control points, procedures and reports, will be governed by a set of key principles that should be complied with by all roles involved in the use, management and development of DDF information systems.

A cloud-based Computing and Data storage has been selected for the DDF repository and the MAGIC database and is compliant with the requirements as stated in the security standard: ISO/IEC 27001:2013. The website, the applications and the database for the MAGIC data, part of the DDF Repository, will be hosted in virtual machines (VMs). A snapshot-based backup system is utilized that creates a point-in-time image based on the current state of each VM related to the different parts of the DDF Repository, which includes the MAGIC database. Backups are taken once per week, minimum, and each backup is retained for 4-6 weeks. Backups are stored in the same datacenter as the corresponding VMs.

Ethics

6.1 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).

Under Task 9.5 of WP "Exploitation and Ethics" all possible ethical matters will be monitored and addressed as needed, including policies relative to data management aligned with confidentiality and IP obligations. An internal ethics response board, chaired by an external advisor, will be formed to oversee all MAGIC's ethical matters.

Most human cell lines to be used are already available through UCL Biobank with ethics to use them for the described studies already in place, including for sharing with EC and non-EC countries. A unique identifier will be allocated to each cell culture available, and only the gene mutation of the affected individual will be shared within the consortium (personal identifiers removed). Laboratories in collaborating institutions involved in experiments using human cell lines, will require local ethical approval for the use of these cells. All hiPSC lines to be used in MAGIC will be registered in the European Human Pluripotent Stem Cell Registry (hpscreg.eu)

Other issues

7.1 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)?

All partners do follow their own data management procedures to the extent these do comply with the common technical and organizational measures as described herein.