



Universiteit
Leiden
The Netherlands

The FAIR journey of a patient-driven registry: reflections and practical solutions from the Duchenne Data Platform FAIRification experience

Lalout, N.; Wilkinson, M.D.; Wandrei, D.; Tassoni, A.; Atalaia, A.; Prieto, M.; ... ; Vroom, E.

Citation

Lalout, N., Wilkinson, M. D., Wandrei, D., Tassoni, A., Atalaia, A., Prieto, M., ... Vroom, E. (2025). The FAIR journey of a patient-driven registry: reflections and practical solutions from the Duchenne Data Platform FAIRification experience. *Journal Of Neuromuscular Diseases*. doi:10.1177/22143602251382969

Version: Publisher's Version

License: [Creative Commons CC BY-NC 4.0 license](https://creativecommons.org/licenses/by-nc/4.0/)

Downloaded from: <https://hdl.handle.net/1887/4284018>

Note: To cite this publication please use the final published version (if applicable).

The FAIR journey of a patient-driven registry: Reflections and practical solutions from the Duchenne Data Platform FAIRification experience

Journal of Neuromuscular Diseases

1–12

© The Author(s) 2025

Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/22143602251382969

journals.sagepub.com/home/jnd



Nawel Lalout^{1,2,3} , Mark D. Wilkinson^{4,5} , Dagmar Wandrei⁶ , Adrian Tassoni⁶ , Antonio Atalaia² , Mario Prieto^{4,5} , Alberto Camara^{4,5} , Eduardo Quemada⁵ , Mirjam Franken¹, Anneliene H. Jonker¹ , Georgios Paliouras^{7,8} , Sergiu Siminiuc⁷ , Claudio Carta⁹, Bruna Dos Santos Vieira¹⁰ , Marco Roos¹¹ , Rajaram Kaliyaperumal¹¹, Teresinha Evangelista², Peter A.C. 't Hoen¹⁰ and Elizabeth Vroom^{2,12}

Abstract

Background: Since 2018, World Duchenne Organization, Dutch Duchenne Parent Project, and Duchenne Data Foundation have been championing efforts to make Duchenne-related data reusable in combination with data contained in other registries. Transforming human language into a coded language that machines can understand (“FAIRification”; FAIR, Findable, Accessible, Interoperable, Reusable) offers a solution.

Purpose: To recount and reflect on the process and challenges encountered during the FAIRification of a patient-registry, the Duchenne Data Platform.

Methods: The FAIRification plan was developed by a multidisciplinary team that was coordinated by a FAIR project manager. It focused on FAIRifying common data elements for rare disease registrations and patient-related outcome data. Protecting patient privacy and autonomy were at the forefront throughout the process. FAIR data transformation was accomplished through a combination of open-source and custom-written software. Data access for federated exploration was enabled through a privacy-preserving “data-visiting” approach.

Results: The plan consisted of 10 main steps and addressed social, legal, ethical, and technical issues. Proof-of-concept testing for interoperability between the Duchenne Data Platform and four other registries demonstrated that FAIR data discovery and reuse was possible. Misconceptions about FAIR data persist, which act as barriers to scaling-up community-level FAIR efforts. Suggestions for overcoming these barriers are provided.

Conclusions: Data visiting and federated analyses between registries is possible. Actions to help mitigate hesitation to implement FAIR in practice include seeking out existing FAIR training opportunities, addressing misconceptions as needed, contacting FAIR experts for advice and using the open-source resources that we have shared.

Keywords

rare diseases, Duchenne muscular dystrophy, FAIR data, FAIRification, federated analyses, data visiting

Received: 11 April 2025; accepted: 1 September 2025

¹Duchenne Parent Project, Veenendaal, The Netherlands

²ERN EURO-NMD, University Hospitals Pitié Salpêtrière, Paris, France

³Radboudumc Institute for Patient Care, Radboud University Medical Center, Nijmegen, The Netherlands

⁴Departamento de Biotecnología-Biología Vegetal, Escuela Técnica Superior de Ingeniería Agronómica, Alimentaria y de Biosistemas and Centro de Biotecnología y Genómica de Plantas (CBGP, UPM-INIA), Universidad Politécnica de Madrid (UPM) – Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA), Madrid, Spain

⁵FAIR Data Systems S.L., Madrid, Spain

⁶Clinical Trials Unit, Faculty of Medicine, Medical Center – University of Freiburg, University of Freiburg, Freiburg, Germany

⁷Duchenne Data Foundation, Veenendaal, The Netherlands

⁸National Center for Scientific Research “Demokritos”, Agia Paraskevi, Greece

⁹Istituto Superiore di Sanità, Rome, Italy

¹⁰Department of Medical BioSciences, Radboud University Medical Center, Nijmegen, The Netherlands

¹¹Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands

¹²World Duchenne Organization, Veenendaal, The Netherlands

Corresponding author:

Nawel Lalout, Radboudumc Institute for Patient Care, Radboud University Medical Center, Nijmegen, The Netherlands.

Email: nawel.lalout@radboudumc.nl



Introduction

Duchenne muscular dystrophy (DMD) is a rare genetic disorder characterized by progressive muscle degeneration and weakness due to the alterations of a protein called dystrophin.¹ Becker muscular dystrophy (BMD) is a mild form of DMD. DMD affects boys with a prevalence of approximately 1 in 5000 newborn boys.²

For the DMD/BMD community, an ongoing concern is how the development of innovative diagnostics, therapies, and healthcare policies to improve care and quality of life are impeded by fragmented and inaccessible health data.³ Since 2018, the World Duchenne Organization, Dutch Duchenne Parent Project, and the Duchenne Data Foundation have been championing efforts to address the various issues underlying this barrier so that existing data can be reused.^{4,5} Patients value the ability to reuse their data as a means of gaining greater control over their care and advocating for their health needs; they also support having their data be reused by health care providers and researchers to benefit all affected.⁶ Issues that limit data reuse include data ownership, data protection, data security, informed consent, and technical characteristics of the data and registries themselves (e.g., type, format).⁷

An approach that helps mitigate barriers to reusing data is “FAIR”; that is, data within a given registry are *Findable* and *Accessible* by machines, data are *Interoperable* under well-defined conditions so that information about the same topic can be combined, and data are thus *Reusable*. Briefly, this means that data should be uniquely and permanently identified, be supported by a wide range of metadata to support both citation and discovery, should take advantage of machine-readable formats and globally accepted vocabularies and ontologies to represent both data and metadata, and should follow community standard.⁸

Another primary limiting factor in implementing FAIR data principles is the rigorous curation of data, which is essential to ensure that datasets are both unambiguously interpretable and of high quality. This includes activities such as consistent use of controlled vocabularies and ontologies, standardization of metadata, validation of data entries, and alignment with common data models. Without such thorough curation, the risk increases for misinterpretation, duplication, or incomplete reuse of the data, thereby undermining its potential value for research, clinical decision-making, and cross-registry interoperability.

A variety of standards and technologies exist to make data FAIR. Examples of standards include the Data Catalog Schema and Vocabulary⁹ for metadata about data repositories; Web Ontology Language¹⁰ for formal knowledge representation; and Resource Description Framework¹¹ and “Linked Data”¹² for data and knowledge publication on the web. Technologies supporting FAIR data include amongst others, Triplestore as an approach to store Linked Data, and the SPARQL Protocol and RDF Query

Language¹³ to query Linked Data. Within the field of rare diseases, the European Joint Programme on Rare Diseases (EJP-RD)¹⁴ has developed FAIR data models for a set of common data elements (CDEs) that should appear in all rare disease registries (See Table 1).¹⁵ These FAIR data models integrate biomedical ontologies such as the Human Phenotype Ontology (HPO),¹⁶ the Orphanet Rare Disease Ontology (ORDO),¹⁷ and the National Cancer Institute’s Thesaurus and Ontology (NCIT).¹⁸

In addition to making data within a given registry FAIR, making it easier to understand and reuse by the patients and local curators, FAIR data also ensures that a given registry will be able to communicate (i.e., be interoperable) with other FAIR-enabled registries within a federated infrastructure. The process of “FAIRification” involves transforming human language into widely accepted identifiers (“codes”) for all concepts such that machines can also unambiguously understand.^{19,20} To illustrate, information about the same disease cannot be combined by machines if, for example, the information exists in French in one registry, German in another, and Swedish in a third. The information in each registry needs to be transformed from the given human language into the same machine coded language.

The first action taken to overcome fragmented and inaccessible data was to design and develop a patient-driven registry known as the Duchenne Data Platform (DDP).²¹ In 2021, FAIRification of the DDP began, and much progress has been made from that moment. However, the journey is at a crossroads and fulfilling the goal of overcoming fragmented DMD/BMD data is only possible if other patient-led and clinical registries embrace and implement FAIR data solutions in practice. DDP’s FAIRification process represents a practical implementation, developed for a specific context and use case. This example should serve to illustrate core FAIR principles in practice, not prescribe a universal method. The purpose of our paper is to recount and reflect on the FAIRification process of the DDP; discuss misconceptions that act as barriers to FAIRification efforts; and propose actions to help other DMD/BMD and rare disease patient organizations and stakeholders to FAIRify their data registries in practice. In doing so, we hope to contribute to the ultimate aim of improving clinical care and quality of life for persons with DMD/BMD.

Methods

FAIR project team

Core members of the FAIR project team were a project manager, two leading academic researchers in bioinformatics (Radboud University Medical Center²²) and biosemantics (Leiden University Medical Center²³), Duchenne domain experts (i.e., patients, patient organization representatives, clinicians), and FAIR consultants and software engineers

Table 1. Overview of the common data elements collected by the participating NMD registries.

#	Common Data Element	EURO-NMD	DM-Scope	DDP	SMARtCARE	CRAMP
1.1.	Pseudonym	✓	✓	✓	✓	✗
2.1.	Date of birth	✓	✓	✓	✓	✓
2.2.	Sex	✓	✓	✓	✓	✓
3.1.	Patient's status	✓	✓	✓	✓	✗
3.2.	Date of death	✓	✓	✓	✓	✗
4.1.	First contact with specialised centre	✓	✓	✓	✓	✗
5.1.	Age at onset	✓	✓	✓	✓	✗
5.2.	Age at diagnosis	✓	✓	✓	✓	✗
6.1.	Diagnosis of the rare disease	✓	✓	✓	✓	✓
6.2.	Genetic diagnosis	✓	✓	✓	✓	✗
6.3.	Undiagnosed case	✓	✗	✗	✗	✗
7.1.	Agreement to be contacted for research purposes	✓	✓	✓	✓	✗
7.2.	Consent to the reuse of data	✓	✗	✓	✓	✗
7.3.	Biological sample	✓	✗	✗	✗	✗
7.4.	Link to a biobank	✓	✓	✗	✗	✗
8.1.	Classification of functioning/disability	✓	✓	✓	✗	✗
	Number of CDEs (out of 16):	16	13	13	12	3

– FAIR Data Systems S.L., Madrid (FDS).²⁴ Additional project collaborators were FAIR data experts from the EJP-RD. The project manager had a central role in coordinating the multidisciplinary team of experts, formulating a detailed plan, defining concrete deliverables, and managing the budget.

Funding and ethical considerations

The Duchenne Parent Project in the Netherlands financed the costs of core FAIR project members and related international meetings and knowledge dissemination activities.^{4,5} The input of extended team members was provided in-kind.

While obtaining ethical approval for the FAIRification project was not necessary, protecting patient privacy and patient autonomy over their data was at the forefront throughout the process. We continually assessed compliance with the General Data Protection Regulation (GDPR) and reviewed whether patients enrolled in the DDP had consented to have their data reused for research purposes. We aimed to develop a FAIR solution based on “data visiting” as opposed to “data sharing” or “data harvesting”. This means registry data do not leave their location, analyses take place within a registry’s protected environment, and only non-identifiable data is transmitted back to the user.²⁵ In addition, when designing algorithms that limit data access to allow only certified queries, we consulted an ethics committee.

FAIR data visiting enables researchers and other authorized stakeholders to query distributed datasets without needing to move or centralize the underlying data. This approach is particularly well-suited for purposes where data privacy, governance constraints, or technical limitations make data sharing difficult. For example, FAIR data

visiting is highly applicable in feasibility studies, where aggregated insights—such as patient counts by phenotype or treatment status—can be drawn from multiple registries without accessing sensitive individual-level data. Similarly, it can support natural history research by allowing standardized queries across multiple datasets to detect patterns in disease progression. In the context of regulatory submissions, data visiting can enable federated analyses to meet evidence requirements while preserving compliance with national data protection laws.²⁶ Additionally, it holds potential for real-time patient feedback and outcome measurement by allowing institutions to compare metrics across sites or cohorts in a privacy-preserving way. However, limitations exist—especially in cases requiring highly detailed, patient-level data or advanced statistical modeling that cannot be performed via federated queries alone. The success of FAIR data visiting depends on the maturity of semantic interoperability, technical infrastructure, and governance frameworks among participating data custodians.

In line with sustainability and data minimization best practices, we used an incremental approach where we addressed a series of specific high-value questions by reusing FAIR implementation components that were developed with open-source resources. In turn, the coding from this project has been made available in the open-source commons.

Data registry description

Launched in 2019, the DDP was designed to allow patients with DMD/BMD to autonomously upload, store, and access their health data. That is, information collected during hospital visits or via wearables could be uploaded into a secure database, stored in a personal data “locker”, and accessed through a mobile phone application or laptop. The DDP

contains patient demographic and family history information, clinical data (e.g., diagnosis, genetic features, phenotypes, medications, clinical care/visits), and patient-reported outcome (PRO) data related to nutrition, bone health, and gastrointestinal issues. Most of this data is provided in structured formats—such as predefined categorical choices, numerical values, and standardized date/time fields—which align well with FAIR principles by enabling machine readability and interoperability. Healthcare data uploads into the DDP primarily consist of discrete data elements, rather than unstructured formats like PDF files.

Originally, the DDP was constructed as a so-called JSON document store within the Microsoft Azure cloud. This meant that for each data point (e.g., an observation or an answer to a PRO measure item), a document, in the structured JSON syntax, would be generated on a patient-by-patient basis, assigned an identifier, and stored. Most data within the record were textual (predominantly in the form of a choice from a selection of possible answers), numerical, or date/time. While these data formats are machine-readable, they are not machine-actionable (See Table 2). Thus, FAIRification of the DDP had to be undertaken largely from scratch. For the DDP FAIR project, mock data were used for testing purposes. These datasets were selected by clinicians and registry custodians to closely simulate real-world data. While not derived from actual patient records, the mock data accurately reflects the structure and characteristics of the data held in the live systems.

FAIRification plan

In general, guidelines from Jacobsen¹⁹ were used to help formulate the steps of the FAIRification plan for both data and metadata. The workflow consisted of mapping the concepts contained in the DDP into ontologies, constructing an ontology to represent the PRO-related data, and developing the technical solution that would perform automated transformations to make the DDP interoperable with other data registries.

We focused on FAIRifying two types of data: (1) the minimal set of CDEs for rare disease registrations (See Table 1) and (2) PROs. We reused components of an EJP-RD FAIR model for the CDEs and developed further techniques to annotate the PRO data elements, aligning with the Human Phenotype Ontology (HPO) and the National Cancer Institute's Thesaurus and Ontology (NCIT). Over the course of the project, we expanded the range of clinical observations that were FAIRified and added the ability to retrieve longitudinal records.

Given the DDP was already in use it was imperative that we not be disruptive to their ongoing activities. As such, we created the FAIR infrastructure as a “shell” around the

existing database infrastructure, such that it was not necessary to modify the source registry data, tools, or practices. The software used to achieve this was a DPP-specific customization of the open-source “FAIR-in-a-Box” (FiaB) software, created jointly by our project, the EJP-RD project, and FAIR Data Systems. FiaB enables three phases of FAIRification: extract, transform, and load (Figure 1). “Extract” refers to the registry data being exported from its native JSON format into a simple comma-separated values (CSV) file format, where the “ontologization” (transforming text into machine-readable codes) is part of the extraction process; “transform” converts the CSV file into output data in RDF format; and “load” refers to the transformed output being loaded into a new FAIR repository known as a triplestore. Further details of the methods, tools, and technologies to transform the CDEs and PRO-related data in the DDP are described in the Supplemental Appendix.

The methodology for preserving privacy while allowing federated access to the data involved, first, designing a technology (called “Shallot”) that allowed only pre-selected queries to be processed on the FAIR data (we will return to this topic in more detail later); and then establishing a peer-review and ethics approval process through which researcher-proposed queries could be evaluated for their: (1) utility and research-relevance, (2) accuracy, (3) privacy-preservation, and (4) patient-acceptance/agreement with the question being asked. To this end, we established a governance board including FAIR experts, clinical experts, privacy experts, and patient representatives, all of whom had to agree on a query before it could be deployed over the FAIR registry via Shallot. All participating registries that implement the same Shallot query then become interoperable with one another.

Results

FAIR outcomes

In 2021, the DDP was assessed using an automated open-source FAIR assessment tool (FAIR Evaluator) and scored 20 out of 22 points.³⁰ At that time, the DDP was the highest scoring digital resource that had ever been tested for FAIRness. In practical terms it meant that the DDP could be found by machines (Findability); machines could determine that the access to DDP metadata was possible (Accessibility); DDP data were ready to be linked to other registries' data (Interoperability); and finally, that once the DDP FAIR transformation workflow was on, machines would know which licenses were in place for reuse of data (Reusability). Six months later, interoperability with ERN EURO-NMD registry was demonstrated. By the end of 2023, privacy-preserving data analysis was demonstrated across multiple and independent FAIR

Table 2. Misconceptions and corresponding clarifications about FAIR data and FAIRification that apply broadly to the field of rare diseases.

Misconceptions	Facts
Conceptual	
Machine-actionability – the ability of a machine to correctly interpret the “meaning” of a piece of data it finds, and act accordingly – is an optional feature of FAIR data that can be ignored.	That data are 100% machine-actionable is the cornerstone of being FAIR. It is what drives interoperability and will allow users from the comfort of their desks to ask a question (i.e., make a query) to multiple registries simultaneously and receive answers within seconds (instead of the current months). Data being machine-actionable goes beyond being machine-readable (i.e., using a standard syntax like JSON); machine-actionable means that human intervention is not required during data discovery, retrieval, and integration. This is only achieved through semantic interoperability, and in fact, the clarity around the meaning of data elements also assists humans to understand the data. ²⁷ Without machine-actionability, FAIR has little purpose.
FAIR data sharing means that anyone can have access to data.	FAIR can be used with full attention to protecting patient privacy. In our example here, the concept of data sharing has been replaced by “data visiting”. With data visiting, data never leaves the registry within which it is housed – it is not “shared”. Also, visitors can be monitored so those who access the data are known (and potentially regulated). Access to data within a registry is controlled through carefully designed algorithms that contain queries that have been certified by an ethics committee. Data analysis is performed within the registry containing the data, and only the answer to the query is delivered back to the user.
Implementation-related	
FAIRification costs too much time/ is not worth the time.	FAIR and FAIR visiting is a sustainable approach that eliminates time-wasting tasks associated with the traditional approach of building a new database to address the issue of fragmented data arising from traditional federation approaches; that is: (1) double data entry; (2) need to harmonize data twice – at the source and the new database); and (3) solving synchronization issues. The FAIR approach is aligned with Open Science and inherently eschews wasting time by reinventing the wheel. Reusable FAIR implementation components are openly available (e.g., in “FAIR-in-a-box”). Other than the DDP-specific extraction library, all code developed in this project is open source and sufficiently generic that it can be immediately reused by any potential partner, with minimal effort/cost. FAIRification dramatically reduces the time required to merge data from multiple registries – improving from months to seconds.
FAIRification costs too much money.	A concrete business case for the DMD and other rare diseases community comparing the traditional approach with implementing FAIRification is still pending. There are implementation costs for both approaches; however, downstream, considerable cost saving can be expected with FAIRification. Moreover, the benefits from reducing the time to finding solutions to health issues thanks to interoperable data sources must not be overlooked. For some perspective, a 2019 report from SPARC Europe estimated that the annual cost of not making research data open and FAIR in Europe was a minimum of €10.2 billion/year. The European intergovernmental organization, ELIXIR, which aims to coordinate life science resources into a single infrastructure, estimates that their return-on-investment is €20 to €1. ²⁸
FAIRification can occur without any changes to status-quo: “I’m already FAIR.”	Institutional and perceived privacy boundaries can seem insurmountable, and lead to the conclusion that “we can’t get there from here”, or worse, claiming “I am already FAIR” as an excuse to not tackle difficult interoperability problems. It is important to be persistent and negotiate accommodations, share real experiences and examples, and openly and honestly address valid (but entrenched) concerns, to ensure that the journey can be initiated and completed.
Benefit-related	
Becoming FAIR is primarily for the benefit of others.	Having data in FAIR format has both direct and indirect benefits to the data controllers/custodians. First, there is a vast amount of publicly available data in the world that is already in FAIR format and can therefore be easily accessed and, because it is open access, dynamically pulled into your local data environment and merged with your local data to enrich the knowledge you already have, for your own use. Second, having data in FAIR format increases one’s visibility to third

(continued)

Table 2. Continued.

Misconceptions	Facts
FAIR does not benefit a world with Artificial Intelligence and Machine Learning, which is the basis of contemporary personalized medicine.	parties and opportunities for collaborations as interactions will have less friction and faster outcomes with less investment (on both sides). The field of Artificial Intelligence has two branches – one is based on logic and reasoning; the other is based on statistical modeling of data. FAIRification involves the pre-cleansing of raw data to put it into logical frameworks consistent with the first branch of AI. This is then fully optimized to be used by the second branch of AI, to build stronger and more accurate statistical models. So, “FAIR” also means “Fully AI Ready”. ²⁹
FAIR principles focus on data management and sharing, but not necessarily on ensuring data validity or quality control.	The FAIRification process not only enables broader data sharing and reuse but also serves as a mechanism for improving the intrinsic quality of the registry itself. A key step in this process involves making the meaning of each data element explicit —through clear definitions, controlled vocabularies, and mappings to established ontologies. This semantic clarification compels registry maintainers and clinical domain experts to revisit and validate the structure and content of their data. As a result, ambiguities are reduced, documentation is improved, and inconsistencies across data entries are identified and corrected. Moreover, this process supports better data governance by encouraging standard operating procedures for data entry, annotation, and curation. It also improves interoperability with external datasets and tools, thereby increasing the registry’s long-term value for research and clinical decision-making. In this way, FAIRification contributes not only to making data more findable, accessible, interoperable, and reusable, but also to enhancing its reliability, interpretability, and sustainability.

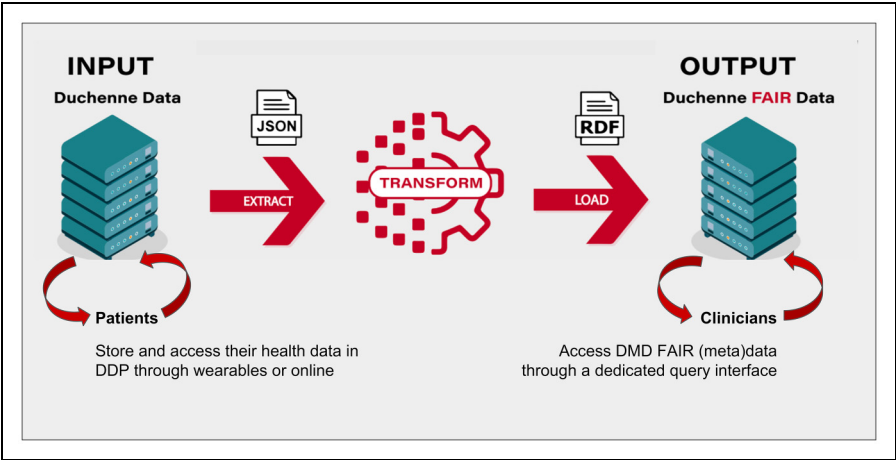


Figure 1. DDP FAIRification approach in three steps: extract, transform and load (ETL pipeline).

registries (i.e., the DDP, ERN EURO-NMD registry and three other neuromuscular registries; Figure 2).

Steps of the DDP FAIRification process

At project inception, there was very little guidance for how to undertake a FAIR initiative or a FAIRification project. The FAIR principles themselves provide no guidance, neither for specific expectations, nor the technology to achieve them. Several of the original authors stated, “how

the FAIR principles should manifest in reality was largely open to interpretation”.³³ Issues such as what data/metadata should be transformed and how; where to host the transformed data; how to update it; how to navigate the distinct privacy differences between FAIR metadata versus FAIR data; and what expertise would be needed, had barely become part of the public/scholarly discourse. Even today, novel approaches to achieve FAIR data transformations continue to appear in the literature.²⁹ The DDP’s FAIRification project consisted of 10 steps summarized in Table 3.

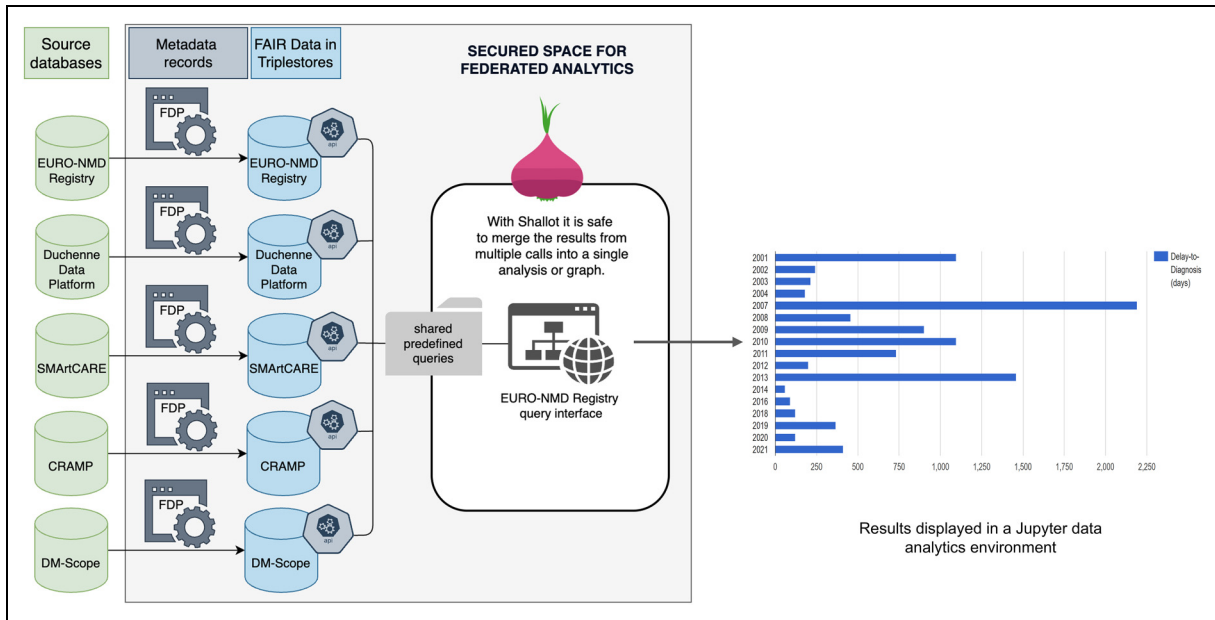


Figure 2. Schematic representation of how the privacy-preserving federation of five independent NMD registries works to provide output for the query “How many patients with DMD/BMD?”. Source databases (green) are untouched and retain their native data schemas and tooling. Within the secured space for federated analytics, a metadata record for each source (grey) is published as a FAIR Data Point (FDP)³¹; FDPs help with resource discovery and query matching. Source data is transformed in a fully or semi-automated manner to conform to the CARE-SM FAIR data model,³² and is stored in a triplestore, a repository for linked data (blue). Shallot services (hexagons) are deployed at each triplestore, where they dynamically import pre-defined and peer-reviewed queries from a central server, ensuring they are coordinated with one another. A Shallot query can be invoked over all participants, and the query results are easily merged for downstream visualization or analytics.

Data visiting infrastructure for federated analyses

When interoperability with other data registries is achieved, the power of FAIR data visiting is fully leveraged. Developing the infrastructure for data visiting involved creating a query component called “Shallot”. This component was developed by adapting a previous software library called “grlc”³⁴ by removing all code and functionalities that were inappropriate for highly sensitive data environments.

As noted above, queries were selected based on their scientific or administrative relevance. We established a formal governance process for how queries are accepted into the FAIR interoperability network, to ensure that the queries were (1) privacy-preserving, (2) accurate, (3) research-relevant, and (4) acceptable to patients. Any future queries, including those involving real data, will be subject to the same formal governance process to ensure consistency, compliance and data protection. The governance board includes FAIR experts, clinical experts, privacy experts, and patient representatives.³⁵

The data visiting process begins with a user selecting their query of interest from the pre-approved list. The queries can vary in complexity from, for example, a simple question such as “How many patients with BMD are wheelchair-bound?” to more difficult ones that require some local data processing (“What is the average age at

which Duchenne patients lose ambulation?”), to complicated analysis algorithms such as calculating the statistical significance of the difference in the age at which patients lose ambulation as a consequence of different corticosteroid treatment regimens using a Cox proportional hazard model.³

Shallot downloads the folder of pre-approved queries from a public repository and provides a public web address (URL) for each of them. Calling that URL triggers the execution of that query, and export of the results. All participants who activate the Shallot component will, therefore, share and publish the same set of pre-approved, privacy-preserving queries, and thereby become both federated and interoperable. Using mock data, proof-of-concept of this data-visiting approach was demonstrated between five participating neuromuscular diseases (NMD) registries: CRAMP (Computer Registry of All Myopathies and Polyneuropathies, Netherlands),³⁶ DM-Scope (National registry for Myotonic Dystrophies, France),³⁷ SMARtCARE (Clinical registry for Spinal Muscular Atrophy, Germany),³⁸ EURO-NMD registry⁷ and DDP.²¹

The selection of the five participating NMD registries was based on several criteria, including their willingness to engage in the FAIRification process, existing maturity of data infrastructure, and alignment with the project’s objectives. Each registry had a history of active data curation and governance, which allowed for a feasible

Table 3. The 10 steps in the FAIRification process of the duchenne data platform.

Step 1	Checkpoint: legal Legal documents were adapted to address the requirements of the FAIR project. Key stakeholders decided GDPR Status of team partners (data controller vs data processor).
Step 2	Review patient informed consent To ensure patients enrolled in DDP had consented to the reuse of their data for research purposes but also at European level (within a European Reference Network).
Step 3	Selected interesting research questions To provide “early wins” for the FAIRification team to show what could be solved through the simplified data integration behaviors enabled by FAIR data.
Step 4	Checkpoint: go-no-go decision point To validate previous decisions with the registry custodians. A reasonably clear estimate of the final project budget was reviewed.
Step 5	Processing and storage Chosen data elements were ready to be FAIRified. We applied the extract-transform-load automated transformation pipeline FAIR-in-a-Box. Storage location was defined.
Step 6	Deployed the question-answering infrastructure This was the stage where the FAIR data was queried and (potentially) integrated with other third-party data to answer selected questions.
Step 7	Checkpoint: validation of output This was used for quality control over (1) the FAIRification itself, (2) the question-answering infrastructure, (3) the selection of third-party data sources, and (4) the accurate federation of these third-party resources.
Step 8	Advertising and publishing It is the metadata record that will allow other users to discover and (where appropriate) reuse the registry FAIR data. Global and EU-recommended standards for metadata publishing were used.
Step 9	Federation FAIRification is generally undertaken with the objective of joining an existing or emergent network (e.g., ERN EURO-NMD) so federated analyses can be performed.
Step 10	Maintenance over the long-term: DDP is now generating FAIR data that can be used to answer interesting questions. If data changes frequently, it is advisable to run the FAIRification pipeline frequently.

post-FAIRification workflow. In terms of quality, the registries were assessed based on the completeness of key data elements, consistency of terminology, and adherence to pre-defined data standards. Coverage varied slightly across registries but collectively offered a broad representation across different NMD subtypes within European countries, enhancing generalizability. Representation was evaluated through the diversity of patient populations and longitudinal data availability. These characteristics were crucial for ensuring that federated analyses and data visiting would yield meaningful, unbiased insights. Furthermore, engagement with registry custodians throughout the process helped confirm the accuracy and interpretability of shared data assets, thus supporting reliable cross-registry research.

First, each participating registry implemented the FAIR transformation workflow and the Shallot data visiting infrastructure. Next, participants created their own CSV file, following the templates provided, and executed their data transformations. Then participants were able to share queries among themselves, “visit” each participant’s registry via their Shallot URLs and execute the same query at all sites (See Figure 2). In the end, all participating sites were able to create their CSV files in a format that matched the required templates. Thus, their data could be automatically transformed and integrated, via Shallot, to support federated analysis.

In a test of our governance mechanism, FAIR experts collaborated with DMD clinical experts to draft a query that would answer a Key Performance Indicator for DMD clinical care – the delay between symptom onset and final diagnosis. The draft query was rewritten several times through consultation with the governance body until it met their approval as being sufficiently aggregative to be privacy-preserving. Moreover, the FAIR experts engaged with the clinical experts on the governance board to design a filter for “data outliers” (i.e., data that *prima facie* could not be true, and likely represent a data entry error) to help improve the quality of the output results. This Shallot query was deployed over all registries participation in the proof-of-concept initiative, and we thereby demonstrated that automated, federated privacy-preserving analysis could be achieved thanks to FAIRification in a manner that generated robust, non-trivial output for use in research.

Building on the mock data analysis conducted across the five sites, the next phase involves operationalizing the FAIR principles in real-world settings. Each site will need to address local infrastructure capabilities, align on metadata standards, and implement governance workflows to support federated data access and query execution. Documenting this process will provide a practical

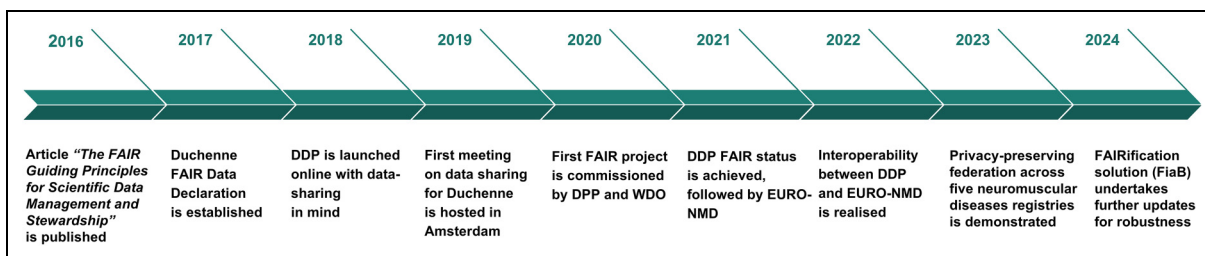


Figure 3. Key milestones during the Duchenne Data Platform FAIR journey.

demonstration of the system's scalability and interoperability under FAIR-aligned conditions.

Discussion

Since starting the FAIRification process of the DDP in 2021 with a multidisciplinary team of experts that was coordinated by a dedicated FAIR data project manager, we have achieved notable outcomes. The DDP was transformed into a high-quality FAIR registry (20/22 points on the FAIR Evaluator assessment tool) and 6-months thereafter, became interoperable with the EURO-NMD registry. In 2023, the DDP was part of a proof-of-concept that demonstrated privacy-preserving federated analysis across five independent registries was feasible.

The FAIRification process consisted of 10 steps and considered legal, ethical, technical, financial, sustainability, and maintenance issues. To protect patient privacy, we chose to design the DDP FAIRification process in terms of data visiting instead of data sharing or data harvesting. Data visiting is challenging, because in almost all cases, neither the researcher, nor the data steward (i.e., the patient, in this case), knows enough about the data structures and semantic content to assist the query/analysis tool in achieving its goal. As such, a data-visiting environment is very much a "black box" where everything that happens inside must be achieved by the computational agent fully autonomously. This is simplified in FAIR environments because FAIR is explicitly designed to aid machines in finding, exploring, and interpreting data on their own.

We used an incremental approach to FAIRification, which gave the FAIR team flexibility in maximizing investment returns at early stages, and more importantly, minimize the processing of data that is not going to be immediately used. We reused and customized previously developed tools and technologies for this FAIRification project; our innovations and technologies have been made open source to aid and encourage other DMD/BMD and rare disease patient organizations.

Throughout our FAIR journey, we have invited other DMD/BMD patient organizations from around the world to join our FAIR efforts and FAIRify their respective registries.⁵ To date, our encouragement has not resulted in new

participating FAIR projects, with hesitancy apparently related to misconceptions about FAIR and FAIRification efforts. Thus, our FAIR journey (Figure 3) is at a crossroads as the goal to overcome the problem of fragmented and inaccessible data is only possible if other patient-led and clinical registries embrace and implement FAIR data solutions in practice.

Misconceptions about FAIR and FAIRification can be classified as (1) conceptual, (2) implementation-related, and (3) benefit-related. Notably, these apply broadly to the field of rare diseases. An overview of the misconceptions and corresponding clarifications are presented in Table 2.

Proposed solutions

We propose five actions to help rare diseases patient organizations implement FAIRification solutions in practice. The first is to educate your directorship and IT staff; various levels of training opportunities exist.^{39,40} This will make them aware of the two FAIRification options: FAIR-by-design, or FAIR retrospectively. FAIR-by-design means that data are FAIRified at the entry point of any electronic data capture (EDC) system such as CASTOR²⁰ or RedCap.⁷ That is, the FAIRification process is incorporated into the chosen EDC software and as soon as a patient clinical data are entered, they are automatically transformed into FAIR data and stored (usually) within a local repository. In contrast, the retrospective approach – as used by the DDP – involves adding a FAIR layer on top of an existing infrastructure. These programs then extract data to be FAIRified, transform it into a machine-actionable format, and store it in a safe repository. At this point, data are ready to be queried by clinicians, researchers, representatives of patient organizations and patients themselves.

While this manuscript defines "FAIR-by-design" as initiating FAIRification at the entry point of any Electronic Data Capture (EDC) system, it is important to recognize that the FAIR principles can be applied even earlier—at the source of data generation. Embedding FAIR practices directly within primary data sources (e.g., clinical systems, laboratory instruments, or patient-generated devices) could further enhance data quality and interoperability by ensuring that data are consistently captured

with standardized formats, terminologies, and metadata from the outset. This upstream integration reduces the need for retrospective harmonization and curation and facilitates more seamless downstream reuse. Future developments in FAIRification should therefore explore strategies that enable FAIR data stewardship from the earliest stages of the data lifecycle.

Second, start identifying barriers and designing plans to address them. For example, attitudes may be modified, and readiness may be increased by educating all stakeholders on the concepts around FAIR data and the consequences of not making your data FAIR. The clarifications corresponding to misconceptions in Table 2 may help in this regard.

Third, start identifying sources of long-term support for FAIRification. Contact experts for advice on how to plan, prepare, and fund your own FAIRification project. Note that support is also available in the form of the open-source tools and technologies that we have provided.

Fourth, to increase the enthusiasm of peers, contribute new queries to the existing privacy-preserving environment that will be of particular interest to them and show them the generated output.

Lastly, remain aligned with the wider FAIR-oriented activities within the EU community. Key examples include the ELIXIR, a distributed infrastructure for biological data across Europe²⁸; the federated European Open Science Cloud⁴¹ and the European Health Data Space.⁴²

Conclusion

It is important to emphasize that FAIR is not an end in itself but rather a set of guiding principles that support broader research, clinical, and regulatory objectives. The utility of FAIR implementation lies in its capacity to enhance the quality, transparency, and reusability of data in ways that directly benefit patient outcomes, scientific discovery, and policy development. Therefore, FAIRness should be viewed as a flexible and iterative process that is aligned with the specific value proposition of a given registry.

First, there are multiple viable pathways to achieving FAIRness, ranging from prospective, “by-design” approaches to retrospective harmonization of existing datasets. Prospective efforts allow for early integration of FAIR elements into registry infrastructure, while retrospective initiatives may involve mapping legacy data to FAIR-compliant formats. Both approaches are valid and often complementary. Encouraging flexibility in how FAIR principles are applied can accelerate adoption across diverse settings and ensure inclusivity of existing data resources.

Second, FAIR implementation should not be treated as a one-size-fits-all process. Instead, strategies must be adapted depending on the specific aims of the data collection effort. In feasibility studies, a lightweight FAIR approach that

prioritizes findability and minimal metadata may be sufficient. Conversely, natural history studies, which often span decades, require robust governance frameworks to ensure interoperability and reusability over time. Recognizing these differences enables the design of proportionate and goal-driven FAIR strategies.

Lastly, clarifying the practical and strategic dimensions of FAIR implementation contributes to more effective and sustainable registry development. By aligning FAIR efforts with the real-world goals of researchers, clinicians, and patients, registries can be better positioned to serve as dynamic, interoperable platforms that support both short-term use cases and long-term data reuse. Ultimately, this alignment fosters a more impactful and responsive data ecosystem for rare disease research and care.

The FAIR journey of the DDP shows that the problem of fragmented and inaccessible DMD/BMD-related data can be overcome. However, further progress towards the ultimate objective of reusing globally available data for innovative research to improve care and quality of life of patients with DMD/BMD is dependent on scaling-up FAIR efforts throughout the community. Actions to help mitigate hesitation to implement FAIR in practice include seeking out existing FAIR training opportunities, identifying barriers in one’s particular context and addressing misconceptions as needed, contacting FAIR experts for advice, and using open-source resources that we have shared, contributing new queries, and remaining aligned with other FAIR activities.

Acknowledgments

The authors thank the DMD/BMD patient community for their advocacy and leadership to implement FAIR in practice, as well as Foundation 29, the first developers of the Duchenne Data Platform for their continuous technical support throughout the duration of these FAIR projects. The authors would also like to thank their collaborators who made the testing of the proof-of-concept on the federated analysis across five independent registries possible: Janbernd Kirschner (SMARtCARE); Guillaume Bassez (DM-Scope) and Robert Molthof (Dutch Hospital Data Foundation on behalf of CRAMP). Lastly, the authors thank Kimi Uegaki, PhD (iWrite) for providing editing support.

ORCID iDs

Nawel Lalout  <https://orcid.org/0000-0003-3707-8373>

Mark D. Wilkinson  <https://orcid.org/0000-0001-6960-357X>

Dagmar Wandrei  <https://orcid.org/0000-0002-8144-162X>

Adrian Tassoni  <https://orcid.org/0000-0001-5392-4125>

Antonio Atalaia  <https://orcid.org/0000-0002-9345-0953>

Mario Prieto  <https://orcid.org/0009-0001-1837-003X>

Alberto Camara  <https://orcid.org/0000-0001-5613-9704>

Eduardo Quemada  <https://orcid.org/0009-0002-4959-9053>

Anneliene H. Jonker  <https://orcid.org/0000-0001-5883-7610>

Georgios Paliouras  <https://orcid.org/0000-0001-9629-2367>

Bruna Dos Santos Vieira  <https://orcid.org/0000-0001-7893-0505>

Marco Roos  <https://orcid.org/0000-0002-8691-772X>

Sergiu Siminiuc  <https://orcid.org/0000-0002-1531-1560>

Ethical considerations

Not applicable.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

Authors' contributions

NL: Conceptualization, Project administration, Funding acquisition, Writing – original draft, Writing – review & editing; MDW: Conceptualization, Data curation, Formal analysis, Methodology, Software, Supervision, Validation, Writing – original draft, Writing – review & editing; DW, AT: Data curation, Formal analysis, Methodology, Investigation, Software, Writing – original draft, Writing – review & editing; AA, TE: Conceptualization, Validation, Writing – review & editing; MP, AC: Data curation, Methodology, Writing – review & editing; EQ: Project Administration, Writing – review & editing; MF: Conceptualization, Project administration, Writing – original draft, Writing – review & editing; AHJ: Writing – review & editing; GP, SS: Data curation, Writing – review & editing; CC: Conceptualization, Writing – original draft, Writing – review & editing; BDSV, RK: Conceptualization, Data curation, Resources, Methodology, Software, Validation, Writing – review & editing; MR, PACTH: Conceptualization, Methodology, Supervision, Writing – original draft, Writing – review & editing; EV: Conceptualization, Project administration, Funding acquisition, Resources, Supervision, Writing – original draft, Writing – review & editing.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Funding was provided by the Duchenne Parent Project and World Duchenne Organization, with in-kind contributions from Data Stewards at EJP-RD. Also, considerable idea-sharing and software co-creation occurred between this project, the EJP-RD Project, the research team of MDW, and FAIR Data Systems commercial partners. As such, additional support for this project was provided by Proyecto TED2021-130788B-I00 financed by MCIN/AEI/10.13039/501100011033 and by the European Union Next Generation EU/ PRTR, and the EJP-RD funded by European Union's Horizon 2020 Research and Innovation Programme under grant agreement N°825575.

Declaration of conflicting interests

The authors declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of

this article: MDW, AC, MP and EQ all work for FAIR Data Systems S.L., Madrid, who were commercial partners in the activities described in this work. The remaining author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Availability of data and materials

The DPP FAIR Data Point can be freely accessed at: <https://w3id.org/duchenne-fdp>. Access to the pre-approved queries, exposed using Shalot, are available at: <https://fairdata.services/api-local>. The FAIR-in-a-box pipeline, excluding the DPP-specific extraction layer, is available at <https://github.com/markwilkinson/FAIR-in-a-box>. Access to the FAIR data may be granted upon reasonable request.

Supplemental material

Supplemental material for this article is available online.

References

1. Hoffman EP, Brown RH Jr and Kunkel LM. Dystrophin: the protein product of the Duchenne muscular dystrophy locus. *Cell* 1987; 51: 919–928.
2. Mendell JR and Lloyd-Puryear M. Report of MDA muscle disease symposium on newborn screening for Duchenne muscular dystrophy. *Muscle Nerve* 2013; 48: 21–26. 20130529.
3. McDonald CM, Henricson EK, Abresch RT, et al. Long-term effects of glucocorticoids on function, quality of life, and survival in patients with Duchenne muscular dystrophy: a prospective cohort study. *Lancet* 2018; 391: 451–461.
4. Verhaart IEC, t Hoen PAC, Roos M, et al. Meeting on data sharing for Duchenne 21–22 March 2019 Amsterdam, The Netherlands. *Neuromuscul Disord* 2019; 29: 800–810. 20190828.
5. van Lin N, Paliouras G, Vroom E, et al. How patient organizations can drive FAIR data efforts to facilitate research and health care: a report of the virtual second international meeting on Duchenne data sharing, March 3, 2021. *J Neuromuscul Dis* 2021; 8: 1097–1108.
6. Courbier S, Dimond R and Bros-Facer V. Share and protect our health data: an evidence based approach to rare disease patients' perspectives on data sharing and data protection – quantitative survey and recommendations. *Orphanet J Rare Dis* 2019; 14: 175. 20190712.
7. Atalaia A, Wandrei D, Lalout N, et al. EURO-NMD registry: federated FAIR infrastructure, innovative technologies and concepts of a patient-centred registry for rare neuromuscular disorders. *Orphanet J Rare Dis* 2024; 19: 66. 20240214.
8. Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR guiding principles for scientific data management and stewardship. *Sci Data* 2016; 3: 160018. 20160315.
9. Riccardo Albertoni (Invited Expert / CNR – Consiglio Nazionale delle Ricerche I, Refinitiv.com) DBIEPa, CSIRO) SJDCIEPa, et al. Data Catalog Vocabulary

- (DCAT) – Version 3, <https://www.w3.org/TR/2024/REC-vocab-dcat-3-20240822/> (2024, accessed 22 August 2024).
10. W3C Semantic Web Standards. OWL, <https://www.w3.org/OWL/> (2012, accessed 18 March 2025).
 11. W3C Semantic Web Standards, Resource description framework (RDF), <https://www.w3.org/RDF/> (2014, accessed 18 March 2025).
 12. Bizer C, Heath T and Berners-Lee T. Linked data – the story so far. *Int J Semant Web Inf Syst (IJSWIS)* 2009.
 13. W3C Semantic Web Standards, SPARQL query language for RDF, <https://www.w3.org/TR/rdf-sparql-query/> (2008, accessed 18 March 2025).
 14. European Joint Program – Rare Diseases. The journey doesn't stop here!, <https://www.ejprarediseases.org/> (2024, accessed 20 February 2025).
 15. European Commission. Set of common data elements. https://eu-rd-platform.jrc.ec.europa.eu/set-of-common-data-elements_en (2018, accessed 18 March 2025).
 16. Robinson PN, Köhler S, Bauer S, et al. The human phenotype ontology: a tool for annotating and analyzing human hereditary disease. *Am J Hum Genet* 2008; 83: 610–615. 20081023.
 17. Vasant DCL, Malone J and Hanauer M. Ordo: an ontology connecting rare disease, epidemiology and genetic data. *Proc Estonian Acad Sci Biol Ecol* 2024.
 18. Golbeck JFG, Hartel F, Hendler J, et al. The National Cancer Institute's thesaurus and ontology (2003). DOI: 10.2139/ssrn.3199007.
 19. Jacobsen A, Kaliyaperumal R and da Silva Santos LOB. A generic workflow for the data FAIRification process. *Data Intelligence* 2020; 2: 56–65.
 20. Kersloot MG, Jacobsen A, Groenen KHJ, et al. De-novo FAIRification via an electronic data capture system by automated transformation of filled electronic case report forms into machine-readable data. *J Biomed Inform* 2021; 122: 103897. 20210826.
 21. Health-RI. Duchenne Data Platform, <https://www.health-ri.nl/en/duchenne-data-platform-ddp> (2025, accessed 21 February 2025).
 22. Radboud University Medical Center, <https://www.radboudumc.nl/en/patient-care> (accessed 21 February 2025).
 23. Leiden University Medical Center, <https://www.lumc.nl/en/> (accessed 21 February 2025).
 24. FAIR Data Systems, <https://fairdata.systems/home/> (2025, accessed 21 February 2025).
 25. van Soest J, Sun C, Musmann O, et al. Using the personal health train for automated and privacy-preserving analytics on vertically partitioned data. *Stud Health Technol Inform* 2018; 247: 581–585.
 26. Hallock H, Marshall SE, t Hoen PAC, et al. Federated networks for distributed analysis of health data. *Front Public Health* 2021; 9: 712569.
 27. RDM Kit. Machine actionability: what does machine-readable, machine-actionable and machine-interpretable mean?, https://rdmkit.elixir-europe.org/machine_actionability (accessed 21 March 2025).
 28. Elixir Europe: Distributed Infrastructure for Biological Data, <https://elixir-europe.org/about-us> (2025, accessed 18 March 2025).
 29. Schultes E. FAIR data and data that are fully AI ready, https://eosc.eu/wp-content/uploads/2024/10/Schultes_Examples-of-EOSC-scientific-and-societal-impact-from-GO-FAIR.pdf (2024, accessed 18 March 2025).
 30. FAIR Evaluation Services. Resources and guidelines to assess the FAIRness of digital resources, <https://w3id.org/AmIFAIR> (accessed 18 March 2025).
 31. Santos LOBD, Burger K, Kaliyaperumal R, et al. FAIR data point: a FAIR-oriented approach for metadata publication. *Data Intelligence* 2023; 5: 163–183.
 32. Alarcón-Moreno P and Wilkinson MD. Take CARE of the patient data. Clinical and registry entries (CARE) semantic model. In: SWAT4HCLS, Leiden, February 26–29, 2024, CEUR Workshop Proceedings.
 33. Mons B, Neylon C, Velterop J, et al. Cloudy, increasingly FAIR; revisiting the FAIR data guiding principles for the European open science cloud. *Inf Serv Use* 2017; 37: 49–56.
 34. grlc, <https://grlc.io/> (accessed 21 February 2025).
 35. ERN EURO-NMD. Registry hub for rare neuromuscular diseases, <https://ern-euro-nmd.eu/reghub/> (accessed 18 March 2025).
 36. van Engelen BG, van Veenendaal H, van Doorn PA, et al. The Dutch neuromuscular database CRAMP (computer registry of all myopathies and polyneuropathies): development and preliminary data. *Neuromuscul Disord* 2007; 17: 33–37. 20061201.
 37. De Antonio M, Dogan C, Daidj F, et al. The DM-scope registry: a rare disease innovative framework bridging the gap between research and medical care. *Orphanet J Rare Dis* 2019; 14: 122. 20190603.
 38. Pechmann A, König K, Bernert G, et al. SMARtCARE – a platform to collect real-life outcome data of patients with spinal muscular atrophy. *Orphanet J Rare Dis* 2019; 14: 18. 20190121.
 39. Health-RI. FAIR Data, <https://www.health-ri.nl/en/about/organisation/fair-data> (2025, accessed 18 March 2025).
 40. FAIRsFAIR. FAIR frameworks & training programmes. (2020, accessed 21 March 2025).
 41. European Commission. European Open Science Cloud (EOSC), https://research-and-innovation.ec.europa.eu/strategy/strategy-research-and-innovation/our-digital-future/open-science/european-open-science-cloud-eosc_en (2022, accessed 18 March 2025).
 42. European Health Data Space Regulation (EHDS), https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space-regulation-ehds_en (2025, accessed 18 March 2025).