



# A Resource for Guiding Data Stewards to Make European Rare Disease Patient Registries FAIR

RESEARCH PAPER

PHILIP VAN DAMME

PABLO ALARCÓN MORENO

CÉSAR H. BERNABÉ

ALBERTO CÁMARA BALLESTEROS

CLÉMENCE M. A. LE CORNEC

BRUNA DOS SANTOS VIEIRA

K. JOERI VAN DER VELDE

SHUXIN ZHANG

CLAUDIO CARTA

RONALD CORNET

PETER A.C. 'T HOEN

ANNIKA JACOBSEN

MORRIS A. SWERTZ

MARCO ROOS

NIRUPAMA BENIS

\*Author affiliations can be found in the back matter of this article

ubiquity press

## CORRESPONDING AUTHOR:

**Philip van Damme**

Amsterdam UMC location  
University of Amsterdam,  
Department of Medical  
Informatics, Meibergdreef 9,  
Amsterdam, NL; Amsterdam  
Public Health, Digital Health &  
Methodology, Amsterdam, NL

[p.vandamme@  
amsterdamumc.nl](mailto:p.vandamme@amsterdamumc.nl)

## ABSTRACT

**Objective:** This paper reports on the development of a dynamic data management planning questionnaire to guide data stewards of the European Reference Network (ERN) rare disease patient registries to make their data findable, accessible, interoperable, and reusable (FAIR). As part of this work, the questionnaire was validated through expert review and aligned with existing resources on rare diseases and FAIR data management.

**Materials and Methods:** The questionnaire was developed for the Data Stewardship Wizard, a tool for data management planning. Knowledge sources on FAIR data, ERN patient registries, and data management were used to compose questions. Ten domain experts validated the questionnaire. The topics in the questionnaire were aligned with existing knowledge bases.

**Results:** A total of 57 questions were included in the questionnaire. Twenty-three references to the FAIR Cookbook and Research Data Management toolkit for Life Sciences were added. Expert validation provided a total of 166 comments on content, structure, and software-related issues. A public instance of the Data Stewardship Wizard was deployed for use by data stewards of ERN patient registries.

**Discussion:** The questionnaire addresses issues that ERNs encounter when making their registries FAIR and follows the implementation choices made by the European rare disease community. A challenging task for future research is to extend the questionnaire to other types of registries and to validate with users.

**Conclusion:** This smart questionnaire is the first model created for the Data Stewardship Wizard that helps ERN patient registries with making their data FAIR. It will assist data stewards in aligning their efforts and providing guidance on FAIR data.

## KEYWORDS:

FAIR data stewardship; data management planning; rare diseases; European Reference Networks; FAIR data; patient registries

## TO CITE THIS ARTICLE:

van Damme, P, Moreno, PA, Bernabé, CH, Ballesteros, AC, Le Cornec, CMA, dos Santos Vieira, B, van der Velde, KJ, Zhang, S, Carta, C, Cornet, R, 't Hoen, PAC, Jacobsen, A, Swertz, MA, Roos, M and Benis, N. 2023. A Resource for Guiding Data Stewards to Make European Rare Disease Patient Registries FAIR. *Data Science Journal*, 19: 12, pp. 1–12. DOI: <https://doi.org/10.5334/dsj-2023-012>

Up to 36 million people are affected by a rare disease in the European Union (EU), which is around 8% of the total EU population at the time of writing (European Commission 2022b). Like rare disease patients, data about rare diseases are often geographically fragmented. To organize the highly specialized care that patients with a rare disease need, the EU has set up European Reference Networks (ERNs) (European Commission 2017). By exchanging knowledge and information among health care providers, these networks aim to improve access to accurate diagnosis, timely treatment, and appropriate care for people living with rare diseases in Europe.

Members of an ERN share expertise on a specific group of diseases (e.g., rare bone or rare kidney diseases). According to the European Medicines Agency, patient registries collect uniform data over time about a population defined by a particular disease, condition, or exposure (European Medicines Agency 2022). A key task of ERNs is setting up and managing patient registries, which are valuable for research, treatment and outcomes monitoring, drug development, and improving quality of care (Boulanger et al. 2020). Standardizing data management practices to allow for data linking and reuse has been known to increase the benefits of rare disease patient registries (Boulanger et al. 2020; Fink et al. 2017; Kodra et al. 2018). As a result, improving the alignment between ERNs is one of the objectives of the European Joint Programme on Rare Diseases (EJP RD), a project with over 130 institutions from 35 countries, including representatives of all 24 ERNs, designed to establish a self-sustaining infrastructure for rare disease research and care (Inserm 2022). The EJP RD has been supporting patient registries, managed by ERNs, in making informed choices about their data management and in harmonizing choices among registries (Dos Santos Vieira, Bernabé, and Zhang et al. 2022).

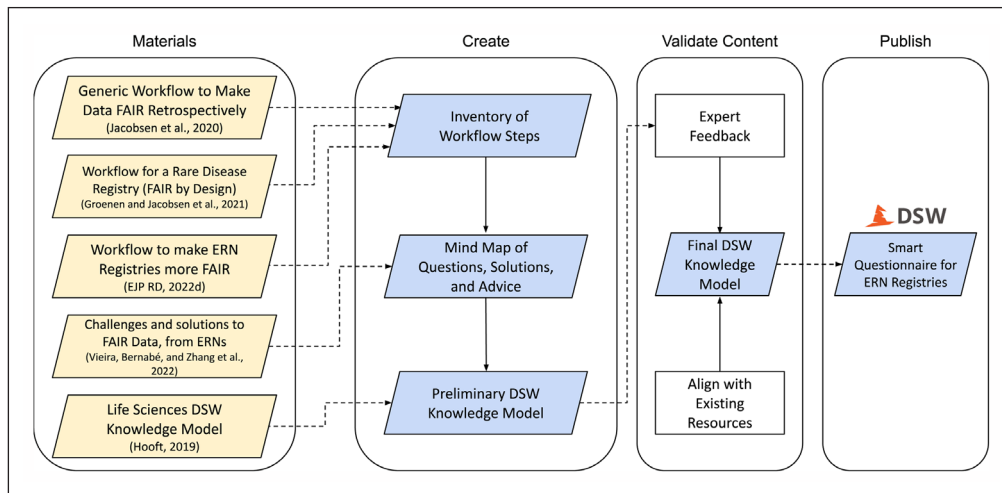
Wilkinson et al. (2016) introduced the findable, accessible, interoperable, and reusable (FAIR) principles, a set of high-level guidelines for research data management, stating that data should be FAIR for humans and computers. Dos Santos Vieira et al. (2022) provided insight into common challenges ERNs encountered when making their patient registry data FAIR and put forward a list of solutions that may help solve those challenges. To obtain these insights, a team of data stewards specialized in FAIR data have been working closely with ERN patient registries.

Hudson-Vitale and Moulaison-Sandy (2019) reviewed research on data management plans (DMPs) and reported that DMPs support data sharing and reuse; however, DMPs were often found to be static documents, making them less effective. Williams, Bagwell, and Zozus (2017) presented a framework for DMPs that covers topics such as personnel planning, data elements, data models, software, privacy, and data sharing practices. Since their introduction, the FAIR principles have been a staple for leveraging DMPs that should produce FAIR data. A tool for assembling DMPs is the Data Stewardship Wizard (DSW) (Pergl et al. 2019). The DSW uses dynamic questionnaires that provide context-dependent guidance, can generate DMPs from prebuilt templates, and provide metrics for compliance with the FAIR principles. Moreover, the DSW includes an expertcurated knowledge model, which represents a questionnaire, for creating DMPs for life sciences projects. In addition, Mons (2018) published a book titled *Data Stewardship for Open Science* that encourages readers to create their own DMPs using the DSW.

While the default knowledge models of the DSW are successful in helping data stewards create DMPs for projects in the life sciences domain, they fail to cover the domain-specific requirements of ERN rare disease patient registries. In addition, the European rare disease community, represented by the EJP RD, has made implementation choices for the FAIR principles specific to their domain (Dos Santos Vieira, Bernabé & Zhang et al. 2022). These choices should be reflected in DMPs of rare disease patient registries. Furthermore, sustaining human support for ERNs is challenging, as requirements evolve over time. Hence, there is a strong need for a maintainable data management planning tool tailored to ERN rare disease patient registries.

This paper reports on the results obtained from (1) the creation and validation of a data management planning questionnaire to guide ERN patient registries in making their data FAIR and (2) its integration with existing infrastructures around rare diseases and (FAIR) data management.

We developed a smart questionnaire for the DSW, that is, a questionnaire with mostly closed-ended questions that adapts follow-up questions based on previously given answers. This questionnaire—to guide data stewards of ERN patient registries to make their registry data FAIR—was built in four stages: (1) collect and analyze relevant knowledge sources; (2) construct a hierarchical mind map from which a DSW knowledge model was built; (3) acquire feedback from domain experts on the different topics in the questionnaire and DMP and align the content with existing tools for FAIR data management planning; and (4) set up a public instance of the DSW with the questionnaire preloaded. Figure 1 summarizes these stages.



**Figure 1** Overview of the stages performed to develop a smart questionnaire for the Data Stewardship Wizard (DSW): gathering relevant knowledge sources, developing a DSW knowledge model (questionnaire), validating the questionnaire, aligning with the Research Data Management toolkit for Life Sciences (RDMkit) (ELIXIR-CONVERGE 2022) and FAIR Cookbook (FAIRplus 2022), and publishing the questionnaire in a DSW instance.

Abbreviations: Data Stewardship Wizard (DSW); European Reference Networks (ERNs); findable, accessible, interoperable, and reusable (FAIR).

Legend: input (yellow), output (blue).

**MATERIALS**

First, we gathered relevant sources that provide information or knowledge on making ERN patient registries FAIR. These sources guided our subsequent decisions on the topics and questions to be included in the questionnaire.

- Generic workflow (Jacobsen, Kaliyaperumal et al. 2020): a step-by-step workflow to make data that has already been collected FAIR
- Rare disease registry workflow (Groenen and Jacobsen et al. 2021): a workflow designed to make the data of a rare disease patient registry on vascular anomalies FAIR from the moment it is collected
- Workflow to make ERN registries FAIR (EJP RD 2022c): a workflow designed to help ERN patient registries with making their data FAIR
- Challenges and solutions from ERN patient registries (Dos Santos Vieira, Bernabé, and Zhang et al. 2022): an extensive list of 41 challenges and proposed solutions that ERN patient registries encountered when making their data FAIR
- DSW knowledge model for the life sciences domain (Hoofst 2019): a knowledge model that includes expert content on data management planning for the life sciences, structured around the research data life cycle

**CREATE**

We created a preliminary knowledge model in three steps. First, we made an inventory of the steps and implementation choices within the three workflows mentioned under ‘Materials’. Second, we built a mind map based on these workflows and ERN challenges. And third, we converted the mind map into a DSW knowledge model. The DSW can export a DMP from a filled-in questionnaire. Questionnaires are generated from knowledge models, which are ordered collections of linked items. A knowledge model contains all the information necessary for generating a questionnaire, such as chapters, questions, descriptions, answer options, and advice bound to answers.

**Mind map**

We used the workflows for making data FAIR as the basis for a hierarchical mind map. A mind map was considered an appropriate intermediate step before building a knowledge model

because they provide a similar hierarchical structure. This mind map laid the groundwork for what would later become the smart questionnaire. We collaboratively populated the mind map with questions, answer options, and solutions. Solutions were written advice, software tools, standards, references to internal (i.e., EJP RD) and external resources, or other technical solutions for making data of ERN registries FAIR. Questions and solutions were derived from the work of Dos Santos Vieira et al. (2022). We used MindMeister (MeisterLabs 2022), a cloud-based online mind mapping tool.

## Knowledge model

After completing the mind map, we converted it to a DSW knowledge model. This step is composed of transferring elements from the mind map and adding additional information (such as answer types, descriptions, and titles) using the DSW's built-in knowledge model editor. In addition, to further enrich our model, we reused all seven chapter names and some relevant questions from the life sciences knowledge model of Hooft (2019). These chapters, based on the research data life cycle (Pergl et al. 2019), were found to be a good fit for structuring the content of our mind map. Hence, our knowledge model was built upon the following chapters: administrative information, reusing data, creating and collecting data, processing data, interpreting data, describing data, and giving access to data. Chapters represent sections of a knowledge model. We restructured questions from the mind map to match the chapters when necessary. Additionally, we added tags to questions that addressed a technical implementation choice for findability, accessibility, interoperability, or reusability. Tags are a feature of the DSW that can be used to organize questions, such as to select questionnaire subsets.

## VALIDATE CONTENT

To validate the correctness of the content of the questionnaire, we approached 10 domain experts and asked for their feedback. Among the invited experts were data scientists, project managers, senior researchers, and software engineers. All experts were affiliated with or involved in the EJP RD. Experts had expertise on authentication and authorization, biobanks, data querying, ERNs, FAIR data, patient consent, privacy legislation, project management, rare diseases, rare disease patient registries, record linkage, semantic models, and software architecture. Experts were asked to only appraise content relevant to their expertise. For example, an expert on patient consent would be asked to review all questions related to consent. Experts reviewed individual questions, the structure of the knowledge model, and additional information presented along with the questions and answers. Feedback was collected through a spreadsheet form, via video call, or both. We curated the received expert reviews to remove duplicate comments and to clarify what changes should be made to the knowledge model. We divided the curated feedback into three categories: textual change (question, answer (option), or description), structural change (e.g., change the question order), or software issue. We then updated the knowledge model according to the feedback.

Finally, we aligned the questionnaire with two existing resources that offer a plenitude of knowledge on how to make data FAIR. That is, the Research Data Management toolkit for Life Sciences (RDMkit) and the FAIR Cookbook (ELIXIR-CONVERGE 2022; FAIRplus 2022). We added references to pages from the RDMkit or recipes from the FAIR Cookbook to any description or advice in the questionnaire that mentioned a topic also covered by one or both resources.

## PUBLISH

Publishing involved hosting a public instance of the DSW with our knowledge model preloaded. We hosted this instance on the servers of ELIXIR's Czech Republic node, which also manages support and operation of the DSW (Czech National Infrastructure for Biological Data 2022). Existing privacy policies apply to this instance. The knowledge model source files were made available on a public repository (see 'Data Availability' section).

## RESULTS

The inventory of workflow steps to make data FAIR and implementation choices suggested by the EJP RD comprises nine steps, 19 topics related to those steps, and 12 implementation choices (e.g., a certain tool or standard). Table 1 shows an overview of this inventory.

| WORKFLOW STEP                               | RELATED TOPICS                              | IMPLEMENTATION   |
|---|---|--|
| 1. Identify FAIR objectives and expertise   | a. Defining objectives                      |  |
|   | b. Giving training                          |  |
|   | c. Hiring of personnel                      |  |
| 2. Define data elements to be collected     | a. Common data elements                     | CDE core elements (European Commission 2019)   |
|   | b. Data dictionary                          |  |
|   | c. Central metadata repository registration | ERDRI.mdr (European Commission 2022a)  |
| 3. Define metadata elements to be collected | a. Machine interpretable metadata           | EJP RD metadata model (EJP RD 2022d)   |
|   | b. Metadata store                           | FAIR data point (Bonino da Silva Santos et al. 2023)   |
| 4. Create a semantic data model             | a. Reuse of existing model(s)               | CDE semantic model (Kaliyaperumal et al. 2022)<br>CDISC ODM (CDISC 2022)<br>HL7 FHIR (HL7 2022)<br>OMOP CDM (OHDSI 2022) |
|   | a. Standardized informed consent form       | ERN ICF (EJP RD 2022b)   |
|   | 5. Obtain consent                           |  |
|   | 6. Enter (FAIR) data                        | a. Electronic data capture systems   |
| 7. Standardize metadata                     | a. Metadata model(s)                        | EJP RD metadata model (EJP RD 2022d)   |
|   | b. Standard terminology                     | CDE semantic model terminology (EJP RD 2022e)  |
| 8. Transform (meta)data to RDF              | a. Data transformation                      | CDE in a box (EJP RD 2022a)  |
|   | b. Terminology mappings                     |  |
| 9. Manage authentication and authorization  | a. Authorization roles                      |  |
|   | b. Access conditions                        |  |
|   | c. Data pseudonymization                    |  |
|   | d. Querying                                 |  |

**Table 1** Overview of the workflow steps and inventory of topics and implementations.

Abbreviations: common data elements on rare disease registration (CDE); European Platform on Rare Disease Registration metadata repository (ERDRI.mdr); European Reference Network (ERN); Health Level 7 Fast Healthcare Interoperability Resources (HL7 FHIR); Clinical Data Interchange Standards Consortium Operational Data Model (CDISC ODM); Observational Medical Outcomes Partnership Common Data Model (OMOP CDM); findable, accessible, interoperable, and reusable (FAIR); informed consent form (ICF); Resource Description Framework (RDF).

Steps, topics, and implementations were translated into questions, answers, and advice. For example, the topic ‘defining objectives’ was rephrased as ‘Have you defined objectives?’ Eventually, the mind mapping process resulted in 22 out of 41 ERN challenges identified by Dos Santos Vieira et al. (2022) being included as a question, answer, advice, or a combination of the three. Challenges that were categorized under ‘community,’ that is, alignment between ERNs, were not included as questions but were indirectly addressed by using the DSW. That is to say, enabling ERN data stewards to use the DSW with our questionnaire untangles those challenges in part. For example, ERNs found that they were unaware of choices other ERNs made, which they can now share through the DSW. Similarly, one challenge categorized under ‘training’ was also indirectly addressed (need for more information on activities of the EJP RD).

Table 2 shows the number of challenges included in the questionnaire and the motivation for why some were not. The full list of challenges and can be found in the original publication (Dos Santos Vieira, Bernabé, and Zhang et al. 2022).

The mind map was converted into a preliminary DSW knowledge model. That is, questions, answers, and advice were added to the knowledge model based on the mind map. We reused one question from the life sciences knowledge model of Hooft (2019): ‘Who is a contributor to the DMP?’ Once this preliminary version of the questionnaire was available in the DSW, we started the validation process.

Validating the correctness of the content of the questionnaire resulted in an updated version of our knowledge model. A total of 10 experts reviewed the content of the questionnaire. We received a total of 166 comments. Each chapter was assigned at least seven experts. All experts reviewed the questions in ‘Processing data’ and ‘Interpreting data.’ Table 3 shows an overview



| CATEGORY       | DIRECTLY INCLUDED | INDIRECTLY INCLUDED | MOTIVATION   |
|----------------|-------------------|---------------------|--|
| Community      | 0 out of 7        | 7 out of 7          | All challenges addressed a lack of alignment between registries. The DSW questionnaire solves this issue.                        |
| Implementation | 7 out of 9        | 0 out of 9          | Two not-included challenges were irrelevant at the time of developing the questionnaire.   |
| Legal          | 3 out of 5        | 0 out of 5          | Two not-included challenges addressed a tool that was not relevant for developing the questionnaire.                             |
| Modeling       | 3 out of 5        | 0 out of 5          | Two not-included challenges addressed issues that were too specific.<br>Five not-included challenges addressed irrelevant tools. |
| Training       | 9 out of 15       | 1 out of 15         | One indirectly covered challenge was not mentioned specifically in the questionnaire but could be deducted from the information. |
| All categories | 22 out of 41      | 8 out of 41         |  |

**Table 2** Challenges and categories from Dos Santos Vieira et al. (2022) that were included in the questionnaire during the mind mapping phase. Challenges marked as indirectly covered are not specifically mentioned in the questionnaire but were solved solely by the use of the Data Stewardship Wizard (DSW) and the questionnaire.

| CHAPTER                      | TEXTUAL CHANGES | STRUCTURAL CHANGES | SOFTWARE ISSUES |
|------------------------------|-----------------|--------------------|-----------------|
| Administrative information   | 18              | 5                  | 3               |
| Reusing data                 | 13              | 2                  | 0               |
| Creating and collecting data | 3               | 2                  | 2               |
| Processing data              | 19              | 3                  | 0               |
| Interpreting data            | 47              | 8                  | 0               |
| Describing data              | 10              | 1                  | 0               |
| Giving access to data        | 27              | 3                  | 0               |
| All chapters                 | 137             | 24                 | 5               |

**Table 3** Quantification of the received feedback per chapter. Feedback is categorized as textual change, structural change, or software issue.

of the comments per chapter and category. Duplicate comments often regarded textual issues on flow or clarity. Experts also provided references to additional resources, such as web pages with more information on a certain topic. Structural changes asked for moving a question up or down the hierarchy or to another chapter. Software issues were related to issues with using the DSW interface, such as a nonfunctional button or a page that would not load. These issues were solved by updating to the latest version of the DSW.

After processing the feedback and updating the knowledge model, the questionnaire has 57 questions. A total of 6 questions are open-ended, and 51 questions are closed-ended. In total, 10 references were added to recipes in the FAIR Cookbook and 13 references to pages of the RDMkit. Three questions were tagged as an implementation choice for findability, 6 to accessibility, 14 to interoperability, and 21 to reusability. Thirteen questions were not tagged because they did not cover implementation choices but rather aspects like training, objectives, or administrative topics. Table 4 shows the number of questions and external references per chapter.

| CHAPTER                      | TOP-LEVEL QUESTIONS | TOTAL QUESTIONS | REFERENCES TO FAIR COOKBOOK | REFERENCES TO RDMKIT |
|------------------------------|---------------------|-----------------|-----------------------------|----------------------|
| Administrative information   | 6                   | 15              | 1                           | 4                    |
| Reusing data                 | 2                   | 9               | 3                           | 3                    |
| Creating and collecting data | 2                   | 5               | 1                           | 1                    |
| Processing data              | 1                   | 5               | 0                           | 2                    |
| Interpreting data            | 2                   | 12              | 4                           | 1                    |
| Describing data              | 2                   | 4               | 0                           | 0                    |
| Giving access to data        | 4                   | 7               | 1                           | 2                    |
| All chapters                 | 19                  | 57              | 10                          | 13                   |

**Table 4** Questions and external references per chapter. Top-level questions are questions that precede all other questions and are always presented to a user. Abbreviations: findable, accessible, interoperable, and reusable (FAIR); Research Data Management toolkit for Life Sciences (RDMkit) (ELIXIR-CONVERGE 2022); FAIR Cookbook (FAIRplus 2022).

Figure 2 depicts a simplified view of our knowledge model and includes all topics covered by the questionnaire. This is the final model that was constructed after expert validation. The questionnaire covers a broad range of topics: building and training a team of professionals, defining data management objectives, (meta)data modeling, data elements, using common standards, using common terminology, data pseudonymization, electronic data capture, querying, metadata exposure, authentication and authorization, and informed consent. Figure 3 provides a screenshot of the chapters and top-level questions. Figure 4 provides a screenshot of how the questionnaire is presented to a user.

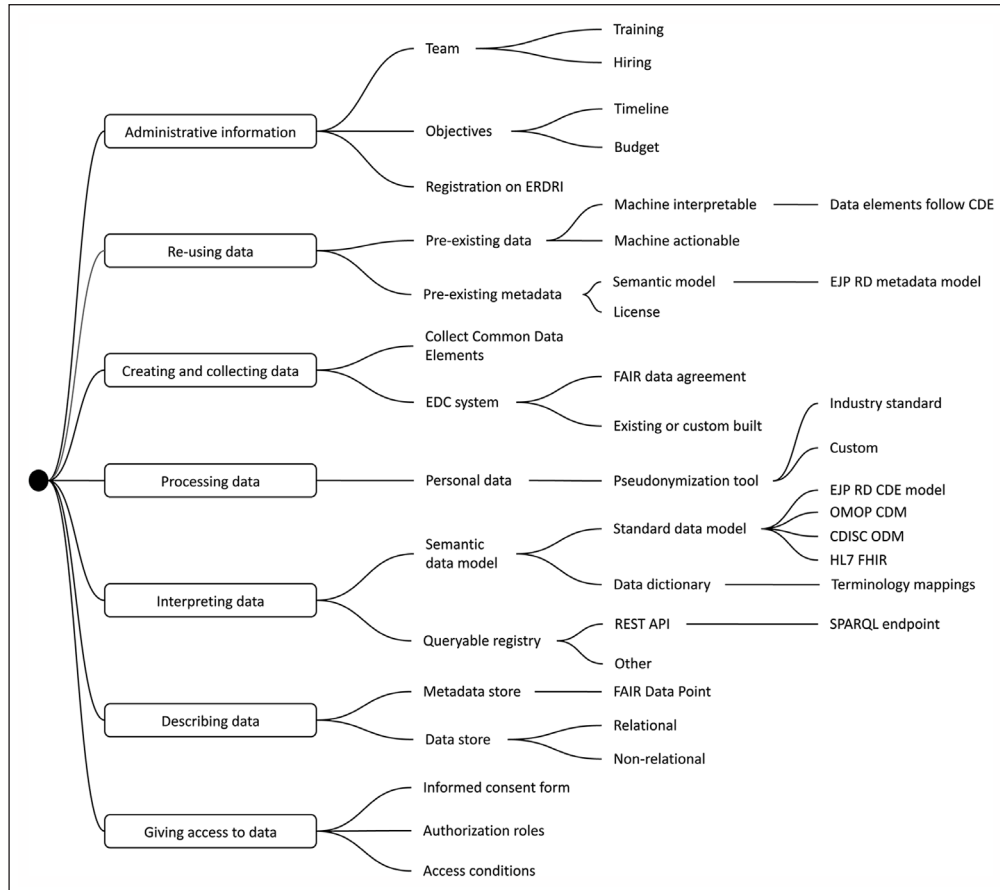


Figure 2 Simplified view of the knowledge model.

Abbreviations: common data elements (CDE); electronic data capture (EDC); European Joint Programme on Rare Diseases (EJP RD); European Platform on Rare Disease Registration (ERDRI); findable, accessible, interoperable, and reusable (FAIR); Health Level 7 Fast Healthcare Interoperability Resources (HL7 FHIR); Clinical Data Interchange Standards Consortium Operational Data Model (CDISC ODM); Observational Medical Outcomes Partnership Common Data Model (OMOP CDM); REpresentational State Transfer Application Programming Interface (REST API); SPARQL Protocol and RDF Query Language (SPARQL).

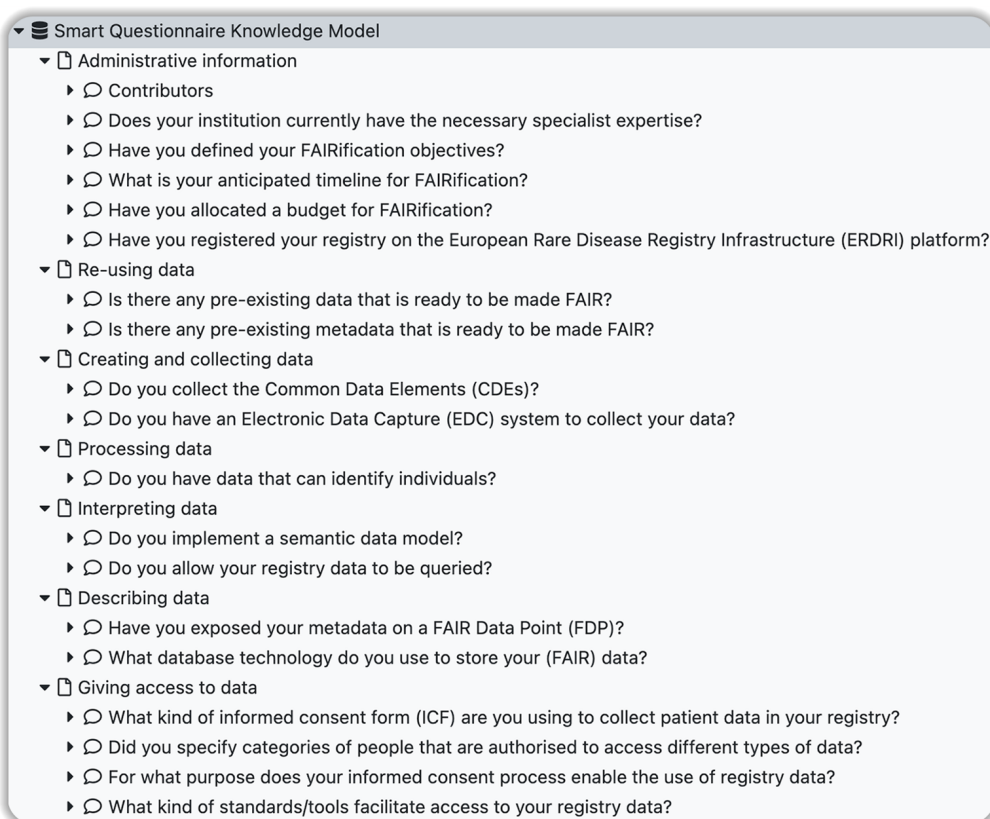
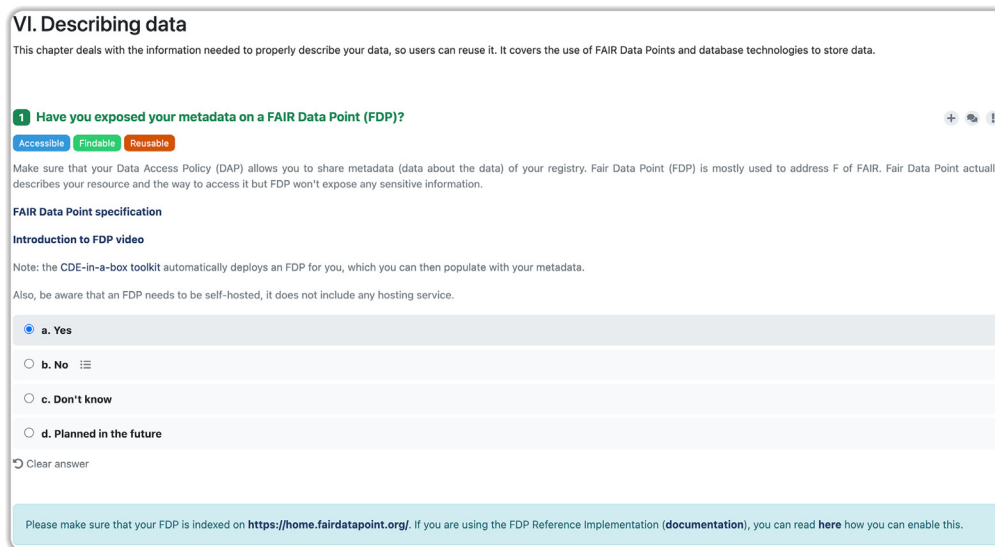


Figure 3 Screenshot of the knowledge model with top-level questions (Data Stewardship Wizard knowledge model editor module).



**VI. Describing data**  
This chapter deals with the information needed to properly describe your data, so users can reuse it. It covers the use of FAIR Data Points and database technologies to store data.

**1 Have you exposed your metadata on a FAIR Data Point (FDP)?**

Accessible Findable Reusable

Make sure that your Data Access Policy (DAP) allows you to share metadata (data about the data) of your registry. Fair Data Point (FDP) is mostly used to address F of FAIR. Fair Data Point actually describes your resource and the way to access it but FDP won't expose any sensitive information.

**FAIR Data Point specification**

**Introduction to FDP video**

Note: the CDE-in-a-box toolkit automatically deploys an FDP for you, which you can then populate with your metadata.

Also, be aware that an FDP needs to be self-hosted, it does not include any hosting service.

a. Yes

b. No

c. Don't know

d. Planned in the future

Clear answer

Please make sure that your FDP is indexed on <https://home.fairdatapoint.org/>. If you are using the FDP Reference Implementation (documentation), you can read [here](#) how you can enable this.

**Figure 4** Screenshot of the first question of the 'Describing data' chapter (Data Stewardship Wizard questionnaire module).

## DISCUSSION

The purpose of this work was to develop a smart questionnaire that guides data stewards working to make data of ERN rare disease patient registries FAIR. Data stewards of patient registries will increasingly have to manage data in ways that comply with implementation choices of the FAIR principles as recommended by their community. Standardizing data management practices of patient registries enables virtual pooling of otherwise sparse and geographically scattered rare disease data, increasing their usefulness for effective research and care. However, standardization for each of the FAIR principles in this domain is complex. ERNs face the challenge of registering data of thousands of diseases from many different sources and making that data as usable as possible within a global health data ecosystem. Our questionnaire addresses those challenges ERNs were known to commonly face and provides guidance according to the FAIR infrastructure set up by the EJP RD. The questionnaire acts as a checklist for making rare disease registries FAIR: data stewards can make sure that all boxes are checked.

Our questionnaire covers the process of making data in patient registries FAIR. Although the questions and advice are made for ERN patient registries, the questionnaire content was based on prior knowledge of FAIR data and experiences in the European rare disease community. For instance, annual Bring Your Own Data workshops have brought together FAIR data experts and rare disease data managers since 2014 (Roos et al. 2014). Workshops such as these have provided a valuable source of challenges regarding the FAIR guiding principles and proposed implementations thereof (e.g., Jacobsen, de Miranda Azevedo et al. 2020a). Furthermore, they affirmed our motivation for designing a smart questionnaire that enables data stewards to begin their FAIR journey from a variety of starting points. As a result of tailoring each topic and question to the unique needs of ERN patient registries, we have filled the gap in having a data management tool that is suitable for rare disease registries in Europe. Since this work is part of ongoing efforts of the EJP RD, integration of the DSW and questionnaire with the European infrastructure for rare disease research will be a natural next step.

Previous studies focused on DMP requirements (Williams, Bagwell & Zozus 2017) and DMPs for the life sciences domain (Pergl et al. 2019; Hooft 2019). Williams, Bagwell, and Zozus (2017) concluded that while most DMPs included components describing data reuse and sharing, few DMPs described data collection and processing practices. These last two are particularly hard to fix, as the quality of poorly collected data can most likely not be improved in retrospect. We were able to address all four topics in our questionnaire. Creating DMPs for projects in the life sciences was addressed by the original authors of the DSW. We found that by reusing parts of their knowledge model, we were able to structure our questionnaire according to a well-established model. Moreover, Jones et al. (2020) concluded that DMPs are essential for FAIR data stewardship. By adopting the DSW as a tool for making ERN patient registries FAIR, we believe our work aligns with that conclusion.



Our work has some limitations. We validated the content of the questionnaire for correctness through expert feedback, but we did not validate the impact of the questionnaire on its intended users. Therefore, further research is needed to determine whether ERN registry data stewards benefit from our tool. Furthermore, the questions and advice are specific to the situation of ERN patient registries and cannot be extrapolated to other registries or projects without modifications. Our work mainly focused on guiding ERN patient registries in making their data FAIR; nevertheless, there is clear value in aligning more types of registries as well. Registry types outside of rare diseases, as well as non-European rare disease patient registries, could fall into this category.

Our work also has several strengths. First, navigating through FAIR implementation choices via questions and answers is a different experience from filling out DMP checklists. It is anticipated that this will lead to an increase in the quality of DMPs. Second, through the DSW, data stewards can learn from the implementation choices of others. Thus, it complements in-person training and contributes to community convergence. Third, we created a single place where ERN data stewards can go for guidance on making their registry FAIR. This makes maintaining and updating the knowledge in the questionnaire easier compared to having various sources in different locations. The knowledge model can be improved by learning from users who will fill out questionnaires on the DSW platform. Moreover, the DSW software is being actively maintained, and hosting our instance on the ELIXIR infrastructure means that it can be sustained beyond the lifetime of the EJP RD.

The knowledge model we developed is publicly available (see ‘Data Availability’ section) and can be used by others to build upon or to reuse parts from. For exporting the DMP, we use a default DSW template, which we intend to customize in the near future. It may be possible to improve the guidance offered to ERN data stewards through further customization of this template. Additional research is needed to quantify the impact of our questionnaire on the (process leading to) ‘FAIRness’ of ERN patient registries. Another challenging task for further research is to extend the questionnaire to other types of resources by collaborating with resource owners and users.

## CONCLUSIONS

The developed smart questionnaire for the DSW is a promising method for guiding data stewards in making their registry data FAIR. It is the first model created for the DSW that helps to standardize data management practices among ERN patient registries. Future research should focus on user validation and extending the questionnaire beyond the realm of ERNs.

## DATA ACCESSIBILITY STATEMENT

The smart questionnaire is available at <https://smartguidance.ejprarediseases.org>. (Registration is required.) A user guide and the knowledge model source files are available at <https://github.com/ejp-rd-vp/smart-guidance>.

## ACKNOWLEDGEMENTS

We thank Marek Suchánek and the ELIXIR Czech Republic node for providing technical support and resources for hosting our public instance of the DSW. We thank Rajaram Kaliyaperumal for setting up and maintaining a local instance of the DSW for testing and development and for providing technical support. We thank Anthony Brookes, Esther van Enkevort, Tala Haddad, Rajaram Kaliyaperumal, Karl Kreiner, Nawel Lalout, Annalisa Landi, Yanis Mimouni, and David Reinert for providing feedback on the questionnaire.

## FUNDING INFORMATION

This work was supported by the EU’s Horizon 2020 research and innovation program under the EJP RD COFUND-EJP No., 825575.

The authors have no competing interests to declare.

## AUTHOR CONTRIBUTIONS

Philip van Damme: conceptualization, methodology, visualization, writing—original draft. Pablo Alarcón Moreno: conceptualization, methodology, writing—review and editing. César H. Bernabé: conceptualization, methodology, writing—review and editing. Alberto Cámara Ballesteros: conceptualization, methodology, writing—review and editing. Clémence M. A. Le Cornec: methodology, writing—review and editing. Bruna Dos Santos Vieira: conceptualization, methodology, writing—review and editing. Joeri van der Velde: methodology, writing—review and editing. Shuxin Zhang: conceptualization, methodology, writing—review and editing. Claudio Carta: writing—review and editing. Ronald Cornet: writing—review and editing. Peter A. C. 't Hoen: writing—review and editing. Annika Jacobsen: resources, writing—review and editing. Morris A. Swertz: writing—review and editing. Marco Roos: conceptualization, project administration, writing—review and editing. Nirupama Benis: conceptualization, methodology, supervision, writing—original draft

## AUTHOR AFFILIATIONS

**Philip van Damme**  [orcid.org/0000-0002-7124-8949](https://orcid.org/0000-0002-7124-8949)

Amsterdam UMC location University of Amsterdam, Department of Medical Informatics, Meibergdreef 9, Amsterdam, NL; Amsterdam Public Health, Digital Health & Methodology, Amsterdam, NL

**Pablo Alarcón Moreno**  [orcid.org/0000-0001-5974-589X](https://orcid.org/0000-0001-5974-589X)

Departamento de Biotecnología-Biología Vegetal, Escuela Técnica Superior de Ingeniería Agronómica, Alimentaria y de Biosistemas, Centro de Biotecnología y Genómica de Plantas. Universidad Politécnica de Madrid (UPM)–Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria-CSIC (INIA-CSIC), Campus Montegancedo 28223 Pozuelo de Alarcón (Madrid), Spain, ES

**César H. Bernabé**  [orcid.org/0000-0003-1795-5930](https://orcid.org/0000-0003-1795-5930)

Department of Human Genetics, Leiden University Medical Center, Leiden, NL

**Alberto Cámara Ballesteros**  [orcid.org/0000-0001-5613-9704](https://orcid.org/0000-0001-5613-9704)

Departamento de Biotecnología-Biología Vegetal, Escuela Técnica Superior de Ingeniería Agronómica, Alimentaria y de Biosistemas, Centro de Biotecnología y Genómica de Plantas. Universidad Politécnica de Madrid (UPM)–Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria-CSIC (INIA-CSIC), Campus Montegancedo 28223 Pozuelo de Alarcón (Madrid), Spain, ES

**Clémence M. A. Le Cornec**

Division of Paediatric Nephrology, Centre for Paediatrics and Adolescent Medicine, University of Heidelberg, Heidelberg, Germany, DE

**Bruna Dos Santos Vieira**  [orcid.org/0000-0001-7893-0505](https://orcid.org/0000-0001-7893-0505)

Center for Molecular and Biomolecular Informatics, Radboud University Medical Center, Nijmegen, NL; Department of Medical Imaging, Radboud University Medical Center, Nijmegen, NL

**K. Joeri van der Velde**  [orcid.org/0000-0002-0934-8375](https://orcid.org/0000-0002-0934-8375)

Genomics Coordination Center, University of Groningen and University Medical Center, Groningen, NL

**Shuxin Zhang**  [orcid.org/0000-0003-4715-9070](https://orcid.org/0000-0003-4715-9070)

Amsterdam UMC location University of Amsterdam, Department of Medical Informatics, Meibergdreef 9, Amsterdam, NL; Amsterdam Public Health, Digital Health & Methodology, Amsterdam, NL

**Claudio Carta**  [orcid.org/0000-0003-3545-198X](https://orcid.org/0000-0003-3545-198X)

Instituto Superiore di Sanità, Rome, IT

**Ronald Cornet**  [orcid.org/0000-0002-1704-5980](https://orcid.org/0000-0002-1704-5980)

Amsterdam UMC location University of Amsterdam, Department of Medical Informatics, Meibergdreef 9, Amsterdam, NL; Amsterdam Public Health, Digital Health & Methodology, Amsterdam, NL

**Peter A.C. 't Hoen**  [orcid.org/0000-0003-4450-3112](https://orcid.org/0000-0003-4450-3112)

Center for Molecular and Biomolecular Informatics, Radboud University Medical Center, Nijmegen, NL

**Annika Jacobsen**  [orcid.org/0000-0003-4818-2360](https://orcid.org/0000-0003-4818-2360)

Department of Human Genetics, Leiden University Medical Center, Leiden, NL

**Morris A. Swertz**  [orcid.org/0000-0002-0979-3401](https://orcid.org/0000-0002-0979-3401)

Genomics Coordination Center, University of Groningen and University Medical Center, Groningen, NL

**Marco Roos**  [orcid.org/0000-0002-8691-772X](https://orcid.org/0000-0002-8691-772X)

Department of Human Genetics, Leiden University Medical Center, Leiden, NL

**Nirupama Benis**  [orcid.org/0000-0002-2101-6154](https://orcid.org/0000-0002-2101-6154)

Amsterdam UMC location University of Amsterdam, Department of Medical Informatics, Meibergdreef 9, Amsterdam, NL; Amsterdam Public Health, Digital Health & Methodology, Amsterdam, NL

- Bonino da Silva Santos, LO, Burger, K, Kaliyaperumal, R and Wilkinson, MD.** 2023. FAIR data point: a FAIR-oriented approach for metadata publication. *Data Intelligence*, 5(1): 163–183. DOI: [https://doi.org/10.1162/dint\\_a\\_00160](https://doi.org/10.1162/dint_a_00160)
- Boulanger, V, Schlemmer, M, Rossov, S, Seebald, A and Gavin, P.** 2020. Establishing patient registries for rare diseases: rationale and challenges. *Pharmaceutical Medicine*, 34(3): 185–190. DOI: <https://doi.org/10.1007/s40290-020-00332-1>
- CDISC.** 2022. *Clinical Data Interchange Standards Consortium Operational Data Model*. <https://www.cdisc.org/standards/data-exchange/odm>
- Czech National Infrastructure for Biological Data** 2022. *ELIXIR CZ*. <https://www.elixir-czech.cz/>
- Dos Santos Vieira, B, Bernabé, CH, Zhang, S,** et al. 2022. Towards FAIRification of sensitive and fragmented rare disease patient data: challenges and solutions in European Reference Network registries. *Orphanet Journal of Rare Diseases*. DOI: <https://doi.org/10.21203/rs.3.rs-1572508/v1>
- EJP RD.** 2022a. *CDE-in-box*. <https://github.com/ejp-rd-vp/cde-in-box>
- EJP RD.** 2022b. *Semantic data model of the set of common data elements for rare disease registration*. <https://github.com/ejp-rd-vp/CDE-semantic-model>
- EJP RD.** 2022c. *Metadata for EJP rare disease patient registries, biobanks and catalogs*. <https://github.com/ejp-rd-vp/resource-metadata-schema>
- EJP RD.** 2022d. *FAIRopoly—FAIRification guidance for ERN patient registries*. <https://www.ejprarediseases.org/fairopoly/>
- EJP RD.** 2022e. *ERN registries generic informed consent forms*. <https://www.ejprarediseases.org/ern-registries-generic-icf/>
- ELIXIR-CONVERGE.** 2022. *The Research Data Management toolkit for Life Sciences (RDMkit)*. <https://rdmkit.elixir-europe.org/>
- European Commission.** 2017. *European Reference Networks: working for patients with rare, low-prevalence and complex diseases*. [https://health.ec.europa.eu/publications/brochure-european-reference-networks-rare-and-complex-diseases\\_en](https://health.ec.europa.eu/publications/brochure-european-reference-networks-rare-and-complex-diseases_en)
- European Commission.** 2019. *Set of common data elements*. [https://eu-rd-platform.jrc.ec.europa.eu/set-of-common-data-elements\\_en](https://eu-rd-platform.jrc.ec.europa.eu/set-of-common-data-elements_en)
- European Commission.** 2022a. *Rare diseases*. [https://health.ec.europa.eu/non-communicable-diseases/steering-group/rare-diseases\\_en](https://health.ec.europa.eu/non-communicable-diseases/steering-group/rare-diseases_en)
- European Commission.** 2022b. *European Platform on Rare Disease Registration metadata repository (ERDRI.mdr)*. <https://eu-rd-platform.jrc.ec.europa.eu/mdr/>
- European Medicines Agency.** 2022. *Patient registries*. <https://www.ema.europa.eu/en/human-regulatory/post-authorisation/patient-registries>
- FAIRplus.** 2022. *The FAIR Cookbook*. <https://faircookbook.elixir-europe.org/>
- Fink, AK, Loeffler, DR, Marshall, BC, Goss, CH and Morgan, WJ.** 2017. Data that empower: the success and promise of CF patient registries. *Pediatric Pulmonology*, 52: S44–S51. DOI: <https://doi.org/10.1002/ppul.23790>
- Groenen, KHJ, Jacobsen, A, Kersloot, MG,** et al. 2021 The de novo FAIRification process of a registry for vascular anomalies. *Orphanet Journal of Rare Diseases*, 16. DOI: <https://doi.org/10.1186/s13023-021-02004-y>
- HL7.** 2022 *Health Level 7 Fast Healthcare Interoperability Resources*. <https://www.hl7.org/fhir/>
- Hoofft, RWW.** 2019. Data stewardship mindmap. <https://doi.org/10.5281/zenodo.2614819>
- Hudson-Vitale, C and Moulaison-Sandy, H.** 2019. Data management plans: a review. *DESIDOC Journal of Library and Information Technology*, 39(6): 322–328. DOI: <https://doi.org/10.14429/djlit.39.06.15086>
- Inserm.** 2022. *European Joint Programme on Rare Diseases*. <https://www.ejprarediseases.org/>
- Jacobsen, A, de Miranda Azevedo, R, Juty, N,** et al. 2020. FAIR principles: interpretations and implementation considerations. *Data Intelligence*, 2(1–2): 10–29. DOI: [https://doi.org/10.1162/dint\\_r\\_00024](https://doi.org/10.1162/dint_r_00024)
- Jacobsen, A, Kaliyaperumal, R, Bonino da Silva Santos, LO, Mons, B, Schultes, E, Roos, M and Thompson, M.** 2020. A generic workflow for the data FAIRification process. *Data Intelligence*, 2(1–2): 56–65. DOI: [https://doi.org/10.1162/dint\\_a\\_00028](https://doi.org/10.1162/dint_a_00028)
- Jones, S, Pergl, R, Hoofft, R,** et al. 2020. Data management planning: how requirements and solutions are beginning to converge. *Data Intelligence*, 2(1–2): 208–219. DOI: [https://doi.org/10.1162/dint\\_a\\_00043](https://doi.org/10.1162/dint_a_00043)
- Kaliyaperumal, R, Wilkinson, MD, Moreno, PA,** et al. 2022. Semantic modelling of common data elements for rare disease registries, and a prototype workflow for their deployment over registry data. *Journal of Biomedical Semantics*, 13. DOI: <https://doi.org/10.1186/s13326-022-00264-6>
- Kodra, Y, Weinbach, J, Posada-de-la-Paz, M,** et al. 2018. Recommendations for improving the quality of rare disease registries. *International Journal of Environmental Research and Public Health*, 15(8): 1644. DOI: <https://doi.org/10.3390/ijerph15081644>

MeisterLabs. 2022. MindMeister. <https://www.mindmeister.com/>

Mons, B. 2018. *Data Stewardship for Open Science*. New York: Chapman and Hall/CRC. DOI: <https://doi.org/10.1201/9781315380711>

OHDSI. 2022. *Observational Medical Outcomes Partnership Common Data Model*. <https://www.ohdsi.org/data-standardization/>

Pergl, R, Hooft, R, Suchánek, M, Knaisl, V and Slifka, J. 2019. 'Data Stewardship Wizard': a tool bringing together researchers, data stewards, and data experts around data management planning. *Data Science Journal*, 18(1): 59. DOI: <https://doi.org/10.5334/dsj-2019-059>

Roos, M, Gray, AJG, Waagmeester, A, et al. 2014. Bring Your Own Data workshops: a mechanism to aid data owners to comply with Linked Data best practices. [https://ceur-ws.org/Vol-1320/paper\\_36.pdf](https://ceur-ws.org/Vol-1320/paper_36.pdf)

Wilkinson, MD, Dumontier, M, Aalbersberg, IJ, et al. 2016. Comment: the FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3. DOI: <https://doi.org/10.1038/sdata.2016.18>

Williams, M, Bagwell, J and Zozus, MN. 2017. Data management plans, the missing perspective. *Journal of Biomedical Informatics*, 71: 130–142. DOI: <https://doi.org/10.1016/j.jbi.2017.05.004>

van Damme et al.

*Data Science Journal*

DOI: 10.5334/dsj-2023-012

12

#### TO CITE THIS ARTICLE:

van Damme, P, Moreno, PA, Bernabé, CH, Ballesteros, AC, Le Cornec, CMA, dos Santos Vieira, B, van der Velde, KJ, Zhang, S, Carta, C, Cornet, R, 't Hoen, PAC, Jacobsen, A, Swertz, MA, Roos, M and Benis, N. 2023. A Resource for Guiding Data Stewards to Make European Rare Disease Patient Registries FAIR. *Data Science Journal*, 19: 12, pp. 1–12. DOI: <https://doi.org/10.5334/dsj-2023-012>

**Submitted:** 30 November 2022

**Accepted:** 26 April 2023

**Published:** 05 June 2023

#### COPYRIGHT:

© 2023 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

*Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.