FAIR Data in Medical Research

Incorporating the FAIR Principles in the Research Data Life Cycle

Kersloot, M.G.

Publication date
2022

Chapter 6

The de novo FAIRification process of a registry for vascular anomalies

Karlijn H.J. Groenen*
Annika Jacobsen*
Martijn G. Kersloot
Bruna dos Santos Vieira
Esther van Enckevort
Rajaram Kaliyaperumal
Derk L. Arts
Peter A.C. 't Hoen
Ronald Cornet
Marco Roos
Leo Schultze Kool

* These authors contributed equally

Orphanet Journal of Rare Diseases 16, 376 (2021).
DOI: 10.1186/s13023-021-02004-y

Abstract

Background

Patient data registries that are FAIR – Findable, Accessible, Interoperable, and Reusable for humans and computers – facilitate research across multiple resources. This is particularly relevant to rare diseases, where data often are scarce and scattered. Specific research questions can be asked across FAIR rare disease registries and other FAIR resources without physically combining the data. Further, FAIR implies well-defined, transparent access conditions, which supports making sensitive data as open as possible and as closed as necessary.

Results

We successfully developed and implemented a process of making a rare disease registry for vascular anomalies FAIR from its conception – *de novo*. Here, we describe the five phases of this process in detail: i) pre-FAIRification, ii) facilitating FAIRification, iii) data collection, iv) generating FAIR data in real-time, and v) using FAIR data. This includes the creation of an electronic case report form and a semantic data model of the elements to be collected (in this case: the "Set of Common Data Elements for Rare Disease Registration" released by the European Commission), and the technical implementation of automatic, real-time data FAIRification in an Electronic Data Capture system. Further, we describe how we contribute to the four facets of FAIR, and how our FAIRification process can be reused by other registries.

Conclusions

In conclusion, a detailed de novo FAIRification process of a registry for vascular anomalies is described. To a large extent, the process may be reused by other rare disease registries, and we envision this work to be a substantial contribution to an ecosystem of FAIR rare disease resources.

Background

Rare disease (RD) registries contain valuable information for improving diagnosis, treatment and event prevention [192]. For this reason, extensive research has been performed on setting up high quality and effective RD registries [193,194]. Using this information for research generally requires data from more than one registry, due to the low prevalence of RDs. However, RD registries are distributed across the world. Also, data from these registries are available in heterogeneous formats and multiple languages. As a consequence, optimising the use of RD registries for research requires substantial effort, and is further complicated by legal constraints and the need for proper precautions for protecting the privacy of the sensitive data. Kodra et al. [193] and Rubinstein et al. [195] mention the FAIR data principles as a means to make the use of distributed RD data as effective as possible.

The FAIR data principles aim to enable efficient analysis of data across multiple sources through enhancing their Findability, Accessibility, Interoperability and Reusability for humans and computers [18]. Data that are FAIR at their source are prepared for efficient computational analysis across multiple FAIR sources. For instance, multiple FAIR sources can be queried simultaneously to answer a research question in so-called 'federated queries' that do not require source data to be moved to one central location [196]. FAIR data are not open by definition. FAIR implies well-defined, transparent access conditions, which supports making data as open as possible and as closed as necessary [16]. By applying the FAIR principles to RD registries (here referred to as the data collected from RD patients), analysis across multiple RD registries and other relevant FAIR data is made possible, even when access criteria differ per source.

The added value of the FAIR principles for RD research led to early acknowledgement by the RD community, and in 2017 the FAIR principles became a recognised resource by the International Rare Disease Research Consortium (IRDiRC) [197]. For example, since 2014, "Bring Your Own Data" workshops (BYODs) have been held to accelerate the adoption of the FAIR principles [198–200]. This includes a series of annually recurring BYODs in the RD domain. Over the years, the FAIRification process applied in BYODs has been explored and refined, and finally described step-by-step in a generic workflow [19]. Other research communities have also developed similar workflows, such as the workflow for FAIRification of data for health research by Sinaci et al. [34]. An important step of this FAIRification process is to make data interoperable and machine-readable in a format that can be read and processed by computers. Data and data access protocols can be made machine-readable by annotating and structuring them with ontologies, which also ensures that data may be more easily analysed across RD registries using federated queries. IRDiRC has recognised ontologies to describe e.g., phenotypes (Human Phenotype Ontology - HPO [201]) and rare diseases (Orphanet Ontology for Rare Disease - ORDO [202]).

Another effort to further improve research across RD registries is the "Set of Common Data Elements for Rare Diseases Registration" (CDEs) released by the Joint Research Centre of the European Commission [203]. The set consists of 16 data elements that are con-

sidered to be essential for research. Next to this, the European Commission has set up the European Rare Disease Registry Infrastructure (ERDRI) to facilitate findability of RD registries [204], an important task for the European Reference Networks (ERNs) [205]. The ERNs are virtual networks at a European level, involving healthcare institutes recognised as expert centres for specific RDs. ERNs aim to facilitate discussion on complex or rare diseases and conditions that require highly specialised treatment. Also, they aim to concentrate knowledge and resources. To that end, ERNs have been provided funding to set up registries [206]. Minimum requirements include support for the CDEs, linking registries and making them interoperable. The European Joint Programme on Rare Diseases (EJP RD) further supports registries in implementing the FAIR principles. VASCERN is the ERN focusing on rare multisystemic vascular diseases [207]. VASCERN is subdivided into thematic working groups, one of which is the Vascular Anomalies working group, VASCA [208]. VASCA includes nine centres with individual databases and data collection processes.

Here, we describe how we set up a FAIR registry for vascular anomalies (hereafter referred to as the VASCA registry) in one of the VASCA centres, Radboud university medical center. The objectives were to 1) base our VASCA registry on the CDEs and the FAIR principles to enable it for analysis across RD registries, and 2) implement de novo FAIRification in our VASCA registry, where data are made FAIR automatically and in real-time upon collection. By doing all the hands-on work for the FAIRification before data collection, data are made FAIR through entering them into an Electronic Data Capture (EDC) system. This mitigates the need for post-hoc FAIRification operations, which include repeated, semi-manual conversions of the data collected into machine-readable data that is performed after data collection. The de novo approach saves time and budget for the actual FAIRification of the data in the VASCA registry. To our knowledge, this is the first attempt to create a de novo FAIR RD registry, and may therefore serve as an example for (and be reused by) other registries. This article focuses on the FAIR part of the registry and not on setting up a registry in general (for recommendations for setting up effective and high quality RD registries in general see for example Kodra et al. [193] and Stanimirovic et al. [194]). Therefore, this article describes the complete de novo FAIRification workflow, from identifying FAIRification objectives and required expertise to querying data over a FAIR Data Point. The technical implementation in the EDC system is described in detail in Chapter 7.

Methods

Workflow of the de novo FAIRification process

The workflow of the de novo FAIRification process for the VASCA registry developed and implemented in this project is divided into five phases: i) pre-FAIRification, ii) facilitating FAIRification, iii) data collection, iv) generating FAIR data in real-time, and v) using FAIR data (Figure 6.1). The phases are further divided into steps describing practical FAIRification tasks as explained throughout the following sections. Through phases I-IV, the

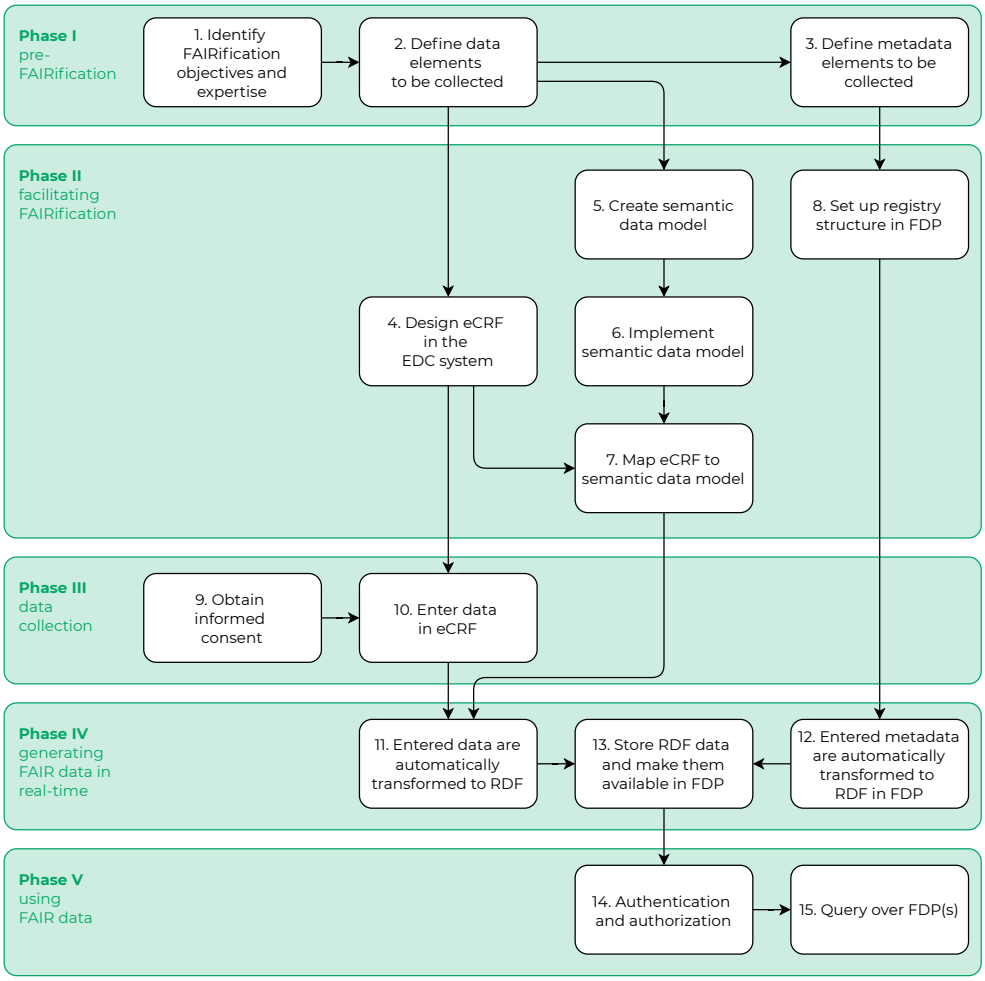


Figure 6.1: Workflow of the de novo FAIRification process of a registry for vascular anomalies

The workflow is divided into five 'phases': pre-FAIRification, facilitating FAIRification, data collection, generating FAIR data in real-time, and using FAIR data. The phases are further specified by 'steps' indicating practical FAIRification tasks. Abbreviations: electronic Case Report Form (eCRF), Electronic Data Capture (EDC), Resource Description Framework (RDF; machine-readable language), FAIR Data Point (FDP).

VASCA registry data and metadata become machine-readable. In the two final phases, the FAIR VASCA registry is made accessible (under well-defined conditions) and used for research.

Phase I: Pre-FAIRification

Pre-FAIRification pertains to the preparatory work before the actual implementation. Here, an inventory was made of everything necessary for developing and implementing

the FAIR VASCA registry in a de novo manner. This includes objectives, team requirements, and resources (data, tools, budget).

Step 1 - Identify FAIRification objectives and expertise

First, the FAIRification objectives for the VASCA registry were identified based on current challenges in RD. The objectives help to set a scope for the FAIRification work to be done and to plan the FAIRification process. In short, these were to 1) base our VASCA registry on the CDEs and the FAIR principles to enable it for analyses across RD registries, and 2) implement de novo FAIRification in our VASCA registry, where data are made automatically FAIR upon collection.

Second, the expertise required to achieve the objectives were identified. Conducting the FAIRification process requires a highly multidisciplinary team guided by a FAIR data steward [19]. The VASCA FAIRification core team consisted of a local data steward, an external FAIR data steward, and an EDC system specialist. Throughout the project, additional expertise was consulted, such as a clinician specialised in vascular anomalies, the Institutional Ethical Review Board, FAIR software developers, and researchers. A full overview of the different kinds of expertise and which part of the FAIRification process they contributed to can be found in Table 6.1. Note, in our project, the expertise in many areas were provided by the same person. Also, research expertise is applicable throughout Table 6.1 and is for simplicity not specified. The areas of expertise have been learned from a previous project [19], and further advanced here.

Step 2 - Define the data elements to be collected

As defined in the FAIRification objective, the data collected in the VASCA registry were based on the CDEs [203]. The CDE 'Classification of functioning/disability' was not added, because there were many uncertainties about its use (see Discussion). We formally defined what data to collect for each of the other CDEs by interpreting the meaning of the CDEs and how they relate to each other. This was captured in an extensive manual for data collection. This task was done by the core FAIRification team together with clinicians specialised in vascular anomalies and a patient advocate. This was an important preparatory step for designing the eCRF (step 4) and creating the semantic data model (step 5) in the 'facilitating FAIRification' phase. This also contributes to consistent data entry by data managers across institutes, step 10 in the 'data collection' phase.

Step 3 - Define the metadata elements to be collected

This step entails identifying what metadata (description of data) should be collected (e.g., license, owner, contributions statements, and description of use conditions and access of data) to comply with the FAIR principles. The World Wide Web Consortium (W3C) Data Catalog Vocabulary (DCAT) [209] is the default standard to predefine and structure metadata elements in FAIR Data Points (as defined in the FAIR Data Point specification; see [210]). We decided to make the VASCA registry findable and accessible (under well-

Table 6.1: Expertise required for the FAIRification of a registry for vascular anomalies (VASCA)

	What expertise is required?	Who provided this expertise?
a	On the data to be FAIRified and how they are managed	· Local and FAIR data steward · EDC system specialist · Clinicians specialised in vascular anomalies · Patient advocate for vascular anomalies
b	On the domain and the aims of the data resource within it	· Clinicians specialised in vascular anomalies · Patient advocate for vascular anomalies
c	On architectural features of the software that is (or will be) used for managing the data	· EDC system specialist · Software developer
d	On access policies applicable to the resource	· Local data steward · Clinicians specialised in vascular anomalies · Institutional Ethical Review Board
e	On the FAIRification process (guiding and monitoring it)	· Local and FAIR data stewards
f	On FAIR software services and their deployment	· EDC system specialist · Software developer
g	On semantic data modelling	· Local and FAIR data steward · Semantic data modelling specialists · Clinicians specialised in vascular anomalies
h	On global standards applicable to the data resource	· Local and FAIR data stewards · EDC system specialist · Senior healthcare interoperability expert
i	On global standards for data access	· Local data and FAIR stewards · EDC system specialist · Senior expert of standards for automated access protocols and privacy preservation

Research expertise is not specified as it is applicable throughout the table. The areas of expertise are inspired from previous FAIRification projects [19]. Abbreviation: Electronic Data Capture (EDC).

defined conditions) in a FAIR Data Point. Metadata elements described in the FAIR Data Point specification were therefore collected.

Metadata elements for the VASCA registry were also collected for ERDRI (ERDRI.dor and ERDIR.mdr). ERDRI.dor (European Directory of Registries) is a catalogue of RD registries. It contains the metadata related to the registry and includes 38 attributes, out of which 23 are compulsory. ERDRI.mdr (ERDRI Metadata Repository) contains detailed information about each variable collected in the registry including data type, description, and a list of permitted items.

Phase II: Facilitating FAIRification

In the second phase, the technical implementation in the EDC system was done to facilitate the de novo FAIRification. This pertains to e.g., the eCRF, the semantic data model, and the FAIR Data Point.

Step 4 - Design the eCRF in the EDC system

The eCRF was designed to collect data for the CDEs (described in step 2) in the Castor EDC system [50]. Several dependencies, e.g., only show 'Date of death' when the patient is deceased, and validations, e.g., validate whether the entered Online Mendelian Inheritance in Man (OMIM) genetic disorder code follows the OMIM standard, were included in order to collect high-quality data (the eCRF questions can be found in [211]). To this end, we mostly worked with closed questions and/or drop-down menus and prevented entering free text as much as possible. An example from the eCRF is shown in Figure 6.2A. The eCRF template containing the CDEs and the ontologies to annotate them (see step 5) was described in a codebook. This codebook was made openly available in ART-DECOR, a platform from Nictiz, the Dutch competence centre for electronic exchange of health and care information [212], and can be directly implemented in the Castor EDC system or other EDC systems using the openly available iCRF Generator tool [213].

Step 5 - Create the semantic data model

We created a semantic data model for the European Commission's recommended set of CDEs to be used for the VASCA registry. The model is openly available on GitHub [214]. A part of the model is shown in Figure 6.2B. First, our interpretations of the CDEs from step 2 were used to draw a conceptual model (listing the main concepts and relationships between the CDEs). This was done in close collaboration between the core FAIRification team, semantic data modelling experts and clinicians specialised in vascular anomalies to ensure that the intended meaning was captured. Later, machine processable ontologies were selected to replace the concepts in the conceptual model. A consequence of this process is that the representation of the data in the semantic data model is a product of our experts' interpretations of the CDEs. Currently, the semantic data model of the CDEs is assessed and further optimised by the RD community (see Discussion). The CDEs recommend diagnosis to be defined with the Orphanet Ontology (ORDO). However, the elements available in ORDO (ORPHAcodes) did not contain all terms in the ISSVA classification used clinically to classify diagnosis (i.e., some ISSVA terms were lacking in orpha). As a solution, we transformed the ISSVA classification into a machine-readable ontology and added mappings to ORDO. The more specific ISSVA terms not available in ORDO were mapped to more general available ORDO terms.

Step 6 - Implement the semantic data model

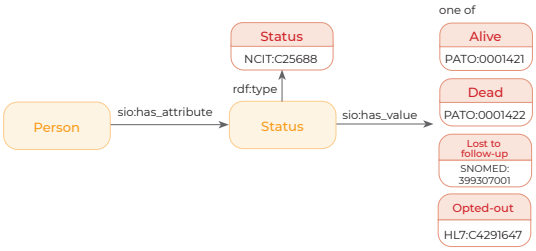
The semantic data model was implemented in a data transformation application in the EDC system, developed in this project (described in detail in Chapter 7). An example of

A. Electronic Case Report Form

2.10 **Patient's status**

- (●) Alive
- ( ) Dead
- ( ) Lost in follow-up
- ( ) Opted-out

B. Semantic data model



C. Implementation of the semantic data model in the data transformation application

Subject	Predicate	Object
Record	sio:SIO_000008	Status /status
Status /status	rdf:type	Status obo:NCIT_C25688 ↗
	sio:SIO_000300	Status Annotated value

D. Map eCRF structure to semantic data model

Object	eCRF question
Status Annotated value	2.10 **Patient's status**

Option	Ontology	Display name	Concept ID
Alive	Phenotypic Quality Ontology	alive	PATO:0001421
Dead	Phenotypic Quality Ontology	dead	PATO:0001422
Lost in follow-up	SNOMED CT	Lost to follow-up	399307001
Opted-out	HL7 RIM	opt-out with exceptions	C4291647

E. Rendered RDF

```
<.../NL-RAD-00001>          sio:SIO_000008      <.../NL-RAD-00001/status> .
<.../NL-RAD-00001/status>   a                   obo:NCIT_C25688 ;
                            sio:SIO_000300      <http://purl.obolibrary.org/obo/PATO_0001421> .
```

Figure 6.2: Schematic representation of the generation of machine-readable data in the Resource Description Framework (RDF)

The rendered RDF (E) is based on electronic Case Report Form (eCRF) data (A), a semantic data model (B), the implementation of this model in a data transformation application (C), and mappings to the eCRF (D).

this implementation is shown in Figure 6.2C. Specific elements in the model that should be filled with eCRF data were marked as elements that require a value. Based on mappings between the eCRF and the semantic data model (see step 7 and Figure 6.2D), the application converts the entered eCRF data to a machine-readable RDF representation (see step 11 and Figure 6.2E) and stores it in a triple store (see step 13). RDF is used since we work with ontologised data in this project and it allows machine-readable representation of ontologies.

Step 7 - Map the eCRF structure to the semantic data model

The eCRF structure was mapped to the semantic data model implemented in the data transformation application (Figure 6.2D). Specifically, the eCRF questions were mapped to the elements in the semantic data model that require an eCRF value (i.e., an object in Figure 6.2C). The eCRF values are linked to ontology concepts that are used as a machine-readable representation of the value in the rendered RDF (Figure 6.2E). For example, the object 'Status' in the semantic data model (Figure 6.2C) is mapped to the eCRF question 'Patient's status' and has a value of the eCRF that is an annotated value, which has one of the values listed under 'Option' in Figure 6.2D. If the patient's status has the value 'Alive', the ontology concept ID PATO:0001421 (Figure 6.2D) will be added to the RDF (Figure 6.2E). The implemented semantic model can be reused in other databases that are built in the EDC system by creating a mapping between the semantic data model elements and the eCRF questions.

Step 8 - Set up registry structure in the FAIR Data Point

The available semantic metadata model of the FAIR Data Point specification was used to describe the VASCA registry [210]. This model is based on the DCAT standard. The VASCA registry FAIR Data Point metadata are described in three layers: 1) catalog - a collection of datasets, 2) dataset - a representation of an individual dataset in the collection, and 3) distribution - a representation of an accessible form of a dataset, e.g., a downloadable file or a web service that gives access to the data for authorised users (Figure 6.3). A catalog may have multiple datasets, and a dataset may have multiple distributions. The VASCA registry described in this project (Registry of Vascular Anomalies - Radboud university medical center) is one of the datasets in the catalog (Registry of Vascular Anomalies). Other VASCA registries, from this or one of the other centers can also be described in this catalog. The semantic metadata model of the FAIR Data Point metadata specification was implemented in the Castor EDC's FAIR Data Point. The metadata that describe the catalog, dataset, and distributions of the VASCA registry described in this project, are publicly available and licensed under the CC0 license.

Phase III: Data collection

The third phase covers the actual collection of the clinical data including the process of obtaining informed consent.

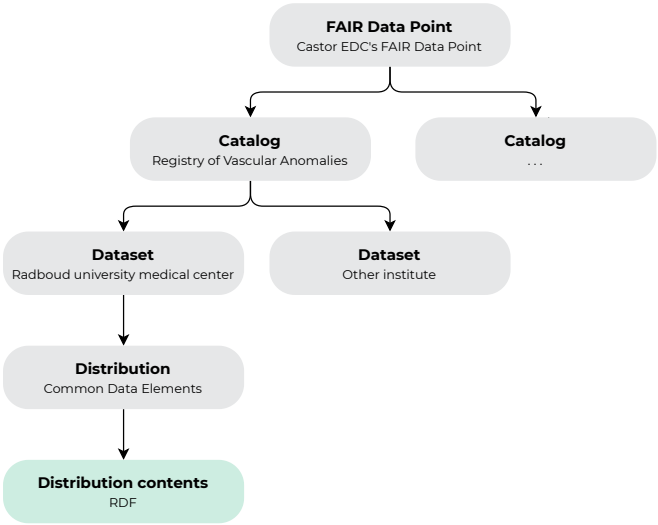


Figure 6.3: Metadata layers for the Registry of Vascular Anomalies (VASCA) in Castor EDC's FAIR Data Point

It consists of three layers: Catalog, Dataset and Distribution. The Distribution layer connects a machine-readable representation of the clinical data collected in the Resource Description Framework (RDF), only for authorised users.

Step 9 - Obtain informed consent

Our project was approved by the Radboud university medical center's Institutional Ethical Review Board. Informed consent was obtained for each patient. The ERN template for obtaining informed consent was not used. Instead, custom made patient information sheets and informed consent forms were applied (see Discussion). Informed consent includes the use of the patient's medical data for the VASCA registry as well as using this data in combination with data collected in other European registries or databases.

Step 10 - Enter data in the eCRF

Currently, data collection for the VASCA registry is a manual process, where data from the EHR is entered into the eCRF. Here, the symptoms described by the clinicians in natural language were manually converted into terms from HPO by using the HPO website [201].

The CDEs are static data elements, meaning that they do not include (changes over) time. Therefore, most data were collected and entered in the eCRF at the first contact in our centre. However, not all information is available at this point. For example, diagnostic imaging and genetic tests may still need to be performed. The results from these tests may provide new insights, thereby affecting CDEs such as genetic diagnosis, phenotype and age at which diagnosis was made. To include missing data or update data elements, we built in a six-month check, conducted six months after inclusion. At this point, data collected for the CDEs may be updated based on (newly) available information in the EHR.

Phase IV: Generating FAIR data in real-time

This phase entails the process of the actual de novo FAIRification of the VASCA registry. Here, the entered data and metadata are automatically converted into machine-readable representations. The machine-readable metadata constitutes the metadata in the FAIR Data Point. The machine-readable data are stored in a triple store (i.e., a specialised database to store and query RDF) and made available in the FAIR Data Point.

Step 11 - Entered data are automatically transformed to RDF

When the data are entered in the eCRF, they are automatically and in real-time converted into a machine-readable RDF representation by the data transformation application. Thus, the data are made machine-readable from the moment they are being collected: de novo FAIRification. This way, a periodic, manual conversion of the data into machine-readable language is not required, resulting in all data collected being available for reuse at any time. Also, updates in the semantic data model lead to automatic updates in the machine-readable RDF representations of data already collected. An additional benefit of this approach is that the people tasked with clinical care and data entry do not need this knowledge to generate FAIR data.

Step 12 - Entered metadata are automatically transformed to RDF in the FAIR Data Point

When the metadata are entered in the FAIR Data Point of the EDC system, they are represented in a human-readable format (a website, e.g., https://fdp.castoredc.com/fdp/catalog/vasca), and at the same time automatically converted into a machine-readable RDF representation, (e.g., the ttl format: https://fdp.castoredc.com/fdp/catalog/vasca?format=ttl).

Step 13 - Store RDF data and make them available in the FAIR data point

After transforming the eCRF data into a machine-readable RDF representation (step 11), they are stored in a triple store. This is done via the data transformation application upon data entry (collected or updated) in the EDC system (step 10). The URL providing access to the machine-readable data in the triple store is made available in the FAIR Data Point as an access URL in the Distribution layer (Figure 6.3).

Phase V: Using FAIR data

The final phase describes how the FAIR VASCA data available in the FAIR Data Point can be accessed and queried for research.

Step 14 - Authentication and authorisation

The VASCA registry metadata in the FAIR Data Point is open (CC0 license) and can be accessed by API calls. The actual registry patient data can only be accessed and queried by logging in with an authorised account of the EDC system (either viewing or exporting

the RDF or querying the data using SPARQL). The process of providing access, authentication and authorisation, is currently arranged in the EDC system. Users (currently only humans) can request access to the data by contacting a specific contact person for the VASCA registry (a Data Catalog Vocabulary contact Point) provided in the metadata. Evaluating requests for access is currently a manual process and follows the permission given for sharing and exchanging data by the patient on the informed consent form (see step 9). The contact person has the authority to decide if access is granted or not. If access is granted an authorised account is provided to the user.

Step 15 - Query over FAIR data point(s)

The machine-readable data are stored in a triple store and can, therefore, be queried using the query language SPARQL by users with access to the data (described in step 14). Query results can be displayed in multiple formats (e.g., JSON, XML, CSV or TSV). The SPARQL endpoint of the EDC system can be queried by using external SPARQL clients or by using a web-based version that is available in Castor EDC's FAIR Data Point. Currently, the web-based version can only query within a single database. Federated queries, therefore, need to be performed with external clients. These (federated) queries allow researchers to ask questions to the FAIR VASCA registry as well as other FAIR RD registries and data resources (multi-source analysis of FAIR data).

Results

We present a workflow for the de novo FAIRification process of the VASCA registry (Figure 6.1, see Methods section). In the following sections, we describe how our approach contributes to each of the four facets of FAIR as well as how it can be reused by other RD registries. For an automated assessment on the FAIRness [215] of our output, the FAIRified data and metadata, using the FAIR Evaluation Services [216] see Chapter 7.

Contribution to the four facets of FAIR

Contribution to F - Findable

We made the VASCA registry findable in searches for humans and computers by providing a description of the registry data ('metadata') relating to findability. The metadata was structured and made machine-readable using the Data Catalog Vocabulary (DCAT) standard (see details in Methods section - step 8). Each vocabulary term in the DCAT standard has a globally unique identifier with a machine-readable definition that can be found on the internet. The DCAT standard provides terms to denote metadata that facilitates findability such as: database title ('Registry of Vascular Anomalies'), description ('Databases of the ERN vascular anomalies'), and country ('The Netherlands'). The metadata was made available in a FAIR Data Point [210, 217] and represented for humans in a visual interface and for computers in the Resource Description Framework (RDF) [218]: http://purl.org/castor/fdp/catalog/vasca. Note that finding the VASCA FAIR

Data Point does not necessarily mean that the registry data can be accessed, interoperated and reused. This is covered in the following sections. VASCA registry metadata was also made findable for humans in the European Directory of Registries (ERDRI.dor) [219] under the namespace: "European Rare Vascular Anomalies Registry". Metadata in ERDRI.dor include the medical areas involved, rare diseases registered, characterisation of the registry, and affiliation to the ERN. At this time the metadata registered in ERDRI.dor is not yet findable for computers in a FAIR format such as DCAT.

Contribution to A - Accessible

We made the VASCA registry accessible by providing metadata in the VASCA FAIR Data Point that describes the access protocols for the registry data in the DCAT vocabulary's 'distribution' component within the FAIR Data Point record. Access protocols define where requests for access (calls) are sent and under which conditions a call will be accepted in order to gain access to data. In addition, calls will generally contain a payload that follows a particular interface format in order to be interpreted by the data endpoint. We accept the HTTP protocol for calls to the data endpoint and accept two interface formats - SPARQL Protocol and RDF Query Language (SPARQL) via HTTP GET, and simple HTTP GET. SPARQL allows for the execution of queries on data that is made available in RDF (similar to SQL for data in relational databases) and can also be used for federated querying across multiple data sources. To retrieve data, SPARQL queries use the semantic data model and ontologies that describe the data (see Methods section and Figure 6.3). Another access protocol was included for simple retrieval, to support viewing or exporting the machine-readable data and can, for example, be used to perform analyses on a local computer. Access to the VASCA registry data is managed by the authentication and authorisation system of the EDC system (see Methods section - step 14). Access can be granted to users (currently only humans) by the VASCA registry contact person specified in the metadata in compliance with the informed consent (see Methods section - step 9). Note, the 'Reusability' facet also relates to the access of registry data and metadata, but focuses on permission and trust, such as consent, license, and attribution.

Contribution to I - Interoperable

We made the VASCA registry machine-readable and interoperable at a number of levels. First, the metadata was structured using the DCAT vocabulary following the FAIR Data Point specification. This contributes to machines being able to query the existence of the registry and its content descriptions. Second, the registry patient data collected in the eCRF was structured using a semantic data model [214] constructed from terms and relations in commonly used ontologies (e.g., SNOMED CT and the IRDiRC recognised ontologies HPO and ORDO). Third, the VASCA registry was configured to collect data for the CDEs, and descriptions of these data elements were registered in the ERDRI Metadata Repository (ERDRI.mdr) [220] under the namespace: 'European Rare Vascular Anomalies Registry'. The CDEs do not directly address any FAIR interoperability principles but do increase the compatibility of data in registries for certain analyses. Using an ontological model to define the meaning of these data elements ensures that we give access

to a harmonised set of data elements and facilitate integration of CDEs from different registries, even across different ERNs. We note that without such an ontological model, computers cannot assess that common data elements are indeed common.

Contribution to R - Reusable

We made the VASCA registry reusable for humans and computers by providing metadata in the VASCA FAIR Data Point relating to reusability. Each metadata layer contains references to a license, the publisher (organisation and person), media type, version, and timestamp of the underlying data or metadata. More metadata are stored in the ERDRI.dor overview, but at this time this information can only be accessed after logging into the EU RD platform and is not yet accessible in a FAIR format. The VASCA registry collects clinical data, which contains privacy sensitive data. By making it FAIR, the registry data are as closed as necessary and as open as possible for other researchers (humans) and computers. The metadata in the VASCA FAIR Data Point are open with a CC0 license [23], whereas, the patient-derived data in the VASCA registry are only accessible to researchers that have been granted access by the registry contact person (see 'Accessibility' facet).

The metadata in the VASCA FAIR Data Point contains a reference to an RDF 'distribution' of the data that can be queried in terms of the CDE semantic model (see Methods section and Figure 6.3). An example ontological query could be: "List all phenotypes reported for patients diagnosed with any type of vascular anomaly or angioma from VASCA FAIR Data Points in France, Germany, and The Netherlands". These queries can span multiple databases, as ontologies are not bound to a single dataset, thereby enabling federated querying.

Reusability of the de novo FAIRification process

Several aspects of the de novo FAIRification process of the VASCA registry have been made available and can be reused by ERNs for setting up their FAIR RD registries that collect the CDEs. The workflow (Figure 6.1) and expertise (Table 6.1) used in our FAIRification projects can be reused for organisation and preparation of other projects. Likewise, other aspects developed for our project that can be reused are our interpretations, semantic data model, and eCRF of the CDEs, FAIR implementations in the EDC system, and structured metadata describing the VASCA registry.

We interpreted the CDEs in order to define what data should be collected in the registry (Methods section - step 2). In our opinion the CDEs are multi-interpretable, hence, downstream implementations depend on these. We therefore properly defined and made our interpretations reusable for others in an extensive manual (available upon request).

We created a semantic data model that describes the CDEs and their relation, and made it available on GitHub [214] (Methods section - step 5). Efforts to further develop and maintain the model are taking place [221] (also see Discussion). The goal of sharing the model is twofold: 1) Reuse: other ERNs can directly implement the model and

would only need to extend the model with elements that are not a part of the CDEs; 2) Improving interoperability: It is easier to perform analyses across datasets if they use the same semantic model (using different models requires ontology mapping to facilitate federated querying).

Castor EDC [50], the vendor of the EDC system used in our project, developed the technology to facilitate the de novo FAIRification of the VASCA registry (phase II in Figure 6.1). The eCRF designed for the CDEs, including the technology to translate to machine-readable format, are reusable (Methods section - steps 6 and 7). The eCRF can be copied directly to a new database within the EDC system, to initiate a new ERN registry. Some ERN-specific adaptations may be necessary. For instance, diagnosis is registered using a drop-down menu focusing on vascular anomalies and should therefore be adjusted for an ERN with a different focus. The ontologies used in the CDE semantic data model are not limited to an area of disease. The developed (eCRF to RDF) data transformation application (Methods section - step 6 onwards; Chapter 7) is generic and can be reused by other registries and clinical trials, ensuring that new FAIRification projects can easily be set up within the EDC system. Likewise, other registries in the EDC system can reuse the FAIR Data Point structure and query functionalities developed for the VASCA registry (Methods section - steps 8, 12, 13, 14, and 15). Furthermore, we have made our eCRF interoperable and reusable, as the codebook describing the eCRF templates containing the CDEs and the ontologies to annotate them is openly available in ART-DECOR [212]. Via the openly available iCRF Generator tool [213], the codebook can be directly implemented in other EDC systems such as OpenClinica and REDcap. Finally, structured metadata describing the VASCA registry on ERDRI.mdr and the FAIR Data Point are reusable. Structured record level metadata of the CDEs were included in ERDRI.mdr (name and descriptions of data collected for the CDEs). Other registries can clone and reuse the VASCA ERDRI.mdr metadata if they are setting up a registry according to the CDEs.

Discussion

This project aimed to 1) base our VASCA registry on the CDEs and FAIR principles to enable it for analysis across RD registries, and 2) implement de novo FAIRification in our VASCA registry, where data are made FAIR automatically and in real-time upon collection. With regard to this first objective, we created an ontology-based semantic model of the CDEs recognised by the European RD community and implemented this model in our eCRF. As a result, machine-readable data can be queried through a FAIR Data Point, thereby facilitating analysis across RD registries. Within this project, we opted for a de novo approach (objective 2). To this end, we developed software that converts 'normal data' entered in the eCRF automatically into machine-readable data, thereby following the semantic model implemented. This comes with the great advantage, that data are made FAIR and available for research upon data entry as well as that clinical people are not tasked with the technical data conversion steps. The step-by-step description provided in this paper, might help other ERNs and RD stakeholders setting up their own

FAIR registries. In the following sections, we discuss the lessons we learned during the project and describe our ideas for future developments.

Lessons learned

The interpretation and collection of the Common Data Elements

The CDEs include seemingly simple elements that turned out to be multi-interpretable. As an example, 'sex' can be interpreted as both genotypic sex and declared sex. Or, the element 'date of first contact with a specialised centre' requires a clear definition of a specialised centre; should it be a Healthcare Provider (HCP) that is a full ERN member, or can it also be an expert unit not being part of the ERN yet? In order to use a registry for research it is essential that it is clearly defined how the CDEs are interpreted for each registry to avoid the possibly false assumption that they are interpreted uniformly across registries. We recommend that all registries clearly document their interpretations of the CDEs, for instance in a manual such as the one created for our VASCA registry. Ideally, guidelines are provided on a European level.

Another issue regarding the CDEs is the discrepancy between data to be collected for the registry and data that is actually collected within the Electronic Health Record (EHR) in daily clinical practise. For example, the ORPHAcodes used to define the diagnosis are very extensive and include a hierarchy. In clinical practice, clinicians may not use ORPHAcodes to code diagnoses in a patient's medical record, nor use these detailed categories. Another example is the CDE 'disability'. The EU prescribes to operationalize the CDE 'disability' using the WHO Disability Assessment Schedule (WHODAS). WHODAS, however, is only validated for adults, whereas a significant part of patients suffering from rare diseases are children.

Furthermore, the CDEs form a static description, thereby not capturing changes in the patients' situation over time (follow-up). The data collected for the CDEs only represent their situation at the moment of data capture, but for some CDEs changes over time are likely to happen. For example, the execution of (new) diagnostic tests in a specialised centre or starting (new) treatments might very well affect the outcome of the disability score. Also, over time, new test results might become available (e.g., genetic tests, imaging), affecting the diagnosis. It is currently unclear in what cases and within what timeframe the information for already included patients should be updated. To this end, advice and alignment on when to assess and update the CDE data is needed.

The 16 CDEs form the core of the registries, but based on discussions with clinicians across Europe, we concluded that clinicians wish to extend the dataset with disease-specific elements that most probably differ between registries. This is, however, something that affects the work required for FAIRification, as the semantic data model should be extended with these disease-specific elements. Consequently, guidelines are required for extending the core CDE model with disease-specific elements. Also, coordination on data modelling is required between ERNs and/or registries to ensure compatible solutions (see also next section).

The semantic data model of the Common Data Elements

We learned that selecting ontologies can be difficult, as this process depends on the interpretation of the CDEs. When a CDE is interpreted similarly in different projects, it is recommended that the same ontology is used, as this prevents the need for mapping between ontologies. To this end, we recommend that a standard set of ontologies should be defined for ERN registries (in addition to HPO and ORDO) to enhance interoperability. When a CDE is interpreted differently in different projects, correct interpretation by FAIR should be facilitated: differences in interpretation are acceptable as long as these interpretations are explicit and represented in both human- and machine-readable formats.

In the current project, interpreting the CDEs and selecting the corresponding ontologies were handled as two distinct activities and to some degree performed separately and independently. As shown in Table 6.1 different expertises were required for interpretation of the data elements (clinicians specialised in and patient advocate for vascular anomalies) and generating a semantic data model (local and FAIR data steward, semantic data modelling specialist and clinicians specialised in vascular anomalies). To enhance efficiency and quality of the semantic data model, we recommend both expertise to be at the table when developing and discussing the semantic data model (at least in the conceptual modelling part).

During our FAIRification project, as expected, the semantic data model continued to evolve. We documented and implemented the first complete version of the model. Currently, the model is being further developed and optimised by ontology experts in EJP RD. Besides this, in future we foresee ongoing adjustments due to e.g., improvement of technologies, ontologies, as well as changes to the CDEs themselves. The question is if, how, and to what extent this would affect the interoperability of datasets. Therefore, one should think of how the community should deal with the use of different models (versions). Researchers should be able to use different versions of the model. Therefore, mapping between versions is essential. We foresee different approaches to deal with this. One would be that the 'owner' of the registry adjusts to a new model or new version. Another approach would be that newly developed models or versions are made mappable to earlier versions, meaning that the community should be provided with either mapping tools or mappable models when the CDE-based semantic data model is further optimised. We would argue that the latter approach would be preferred as it requires less effort of the end users. Particularly if many researchers (end users) make use of the same model, this second approach is beneficial, as the modelling work only needs to be done once. In contrast, in the first approach all users would need to adjust to the model individually. Further optimisation of the model also leads to further complexities such as different versions of semantic models needing to be mapped to different versions of the eCRF. In both approaches, our de novo FAIRification framework implies less extra work when a model is changed compared to post-hoc FAIRification. The conversion into a machine-readable format is more or less automatic and would only require implementing the updated model in the eCRF (Methods step 6). In contrast, post-hoc FAIRification would require additional redoing the semi-manual conversion into a machine-readable format.

FAIR implementation in the EDC system

Enabling de novo FAIRification in the Castor EDC system required developing the necessary technology from scratch. We first piloted the generation of machine-readable data to test the integration between the data transformation application and the EDC system. We prioritised developing a generic tool, rather than a smaller registry-specific tool, as it can be used by a large number of registries and clinical studies. The scalability of our approach contributes to making more FAIR data available for the community.

In addition, we decided to implement authentication and authorisation layers in the FAIR Data Point by reusing the authentication and authorisation of the EDC system. This means that at the moment, researchers that do not have access to the database in the EDC system are not able to access the data through the FAIR Data Point.

Informed consent

Informed consent is usually required for collecting prospective patient data for scientific purposes. The European Commission has provided the ERNs with a standard patient information folder (PIF) and broad informed consent form (ICF). Our Institutional Ethical Review Board did not approve the PIF and ICF for scientific registries. Main reasons were that the information provided on data handling was too limited. Therefore, our Institutional Ethical Review Board requested us to redraw the PIFs and ICFs. This has several possible consequences. Not only do the different centres need to follow local guidelines, one also needs to make sure data exchange is facilitated in an easy way. Future collaborations including data sharing with other parties and the own ERN working group should explicitly be part of the PIFs and ICFs.

Preconditions for an effective (FAIR) registry

Previous research has investigated the preconditions for the establishment of a RD registry. Using focus group sessions, Stanimirovic et al. [194] identified that "the effective development of a national RD registry, followed by the establishment of a RD ecosystem, requires a broad approach that entails a whole series of systemic changes and considerations. Moreover, well-orchestrated and well-funded efforts to achieve this goal should involve coordinated action of all stakeholders, including a regulatory framework, quality design, and enactment of a general RD policy, as well as the alignment of medical, organizational, and technological aspects in accordance with the long-term public healthcare objectives". Most of these aspects are also identified by Kodra et al. [193]. All these prerequisites are also essential for setting up effective FAIR registries. Adding the FAIR aspects to a registry, puts extra 'pressure' on several of these preconditions. First, additional demands are made on the IT infrastructure, as it should also facilitate the conversion of clinical data into ontological (meta)data and federated querying via FAIR Data Points. In case of the latter, the FAIR Data Points should be able to connect different (types of) registries. These additional demands on IT infrastructure apply to both development or setting up of the registry and long-term maintenance of the registry. Secondly, the legal basis might be more complex, as there should not only be a legal basis for collecting

data, but also for (automated) sharing and re-using data (by others). In case others aim to re-use the data via SPARQL queries in the FAIR Data Point, one should determine if the nature of the query and purpose for which the query results will be used, match the original legal basis of the registry. Ideally, these aspects are checked automatically in the FAIR Data Point. This technology is yet to be developed. Furthermore, FAIR data stewards, semantic modelling specialists, interoperability experts, and experts on standards for automated access protocols and privacy preservation should be added to the already highly interdisciplinary group of professionals tasked with setting up the registry.

Future developments

The rapid development of FAIR technologies and possibilities requires us to continuously improve our FAIRification workflow. We are currently working on several aspects, discussed below. The European Patient Identity Management (EUPID) pseudonymization tool [222] is recommended by the European Commission [223] and aims to ensure that different registries can be mapped on a patient-to-patient level. However, at the time of setting up the VASCA registry, EUPID was not up and running yet and, therefore, not implemented in the VASCA registry. We are currently exploring the technical options to integrate EUPID into the registry, taking aspects related to automation, security, privacy and efficacy into account.

As described in the Methods section, we mapped the International Society for the Study of Vascular Anomalies (ISSVA) terms to the ORPHAcodes. However, the ISSVA terms not present in ORDO lacked a unique identifier. To comply with the interoperability principles, we are currently transforming the ISSVA classification into an ontology (OWL format), keeping the structure and adding all possible concepts and terms mappings to HPO, ICD, SNOMED CT, ORDO and NCIT. This way, in case an ISSVA term is not present in other existing ontologies, it has a unique identifier.

Setting up a registry requires a good balance between the amount of information one would like to collect, and the amount clinicians are able to provide given the limited time they can spend for each patient. In the current registry, clinicians provide all information. We are currently looking into the possibilities for a patient-driven registry. In patient-driven registries, patients fill in (part of the) data themselves rather than the clinician. This way, we would be able to collect more data with less effort. This would additionally enhance the options for collecting longitudinal data (which is not covered by CDEs), for example on quality of life, medication intake or treatments, thereby allowing additional research questions to be answered.

In addition, reaching interoperability, and thereby facilitating secondary use of data from the EHR, requires the use of ontologies during data collection. Currently, this means that the data from the EHR (both structured fields and notes made by clinicians) should be “converted” into terms used in the ontologies. This is currently mostly manual work and is heavily dependent on interpretation by the person carrying out the data entry in the EDC system. To further optimise and automate this process, we are currently exploring whether software tools that automatically map free text to ontologies can aid in this. An example implementation would be the facilitated mapping of diagnoses extracted

from the EHR to HPO or other ontology terms, using software, such as Phenotips [224], Zooma [225], and SORTA [226, 227]. Alternatively, it would be interesting to work on a tool for mapping eCRF data with ontology terms.

The web-based query method in the EDC system can currently only be used to query data in one registry, but work is being done to support querying over multiple registries. This would allow for easier retrieval of relevant information from multiple registries. For further interoperability, we would require an interface that facilitates queries over multiple registries, independent of the EDC system used for construction of the registries.

Next steps will also include the development of human and machine-readable access conditions to the data and, subsequently, the implementation of a mechanism for requesting and granting access to the data.

Conclusion

In conclusion, we successfully set up a workflow for de novo FAIRification of the CDEs for the registry of vascular anomalies. The methods and lessons learned in the different phases of the FAIRification process are described in detail. This may help other ERNs in setting up their FAIR registries.

Next steps are to extend the VASCA registry with disease-specific data elements and to set up this registry in the other VASCA institutes and VASCERN working groups. This will allow us to analyse data across multiple registries using federated queries and, thereby, to demonstrate the added value of making them FAIR.