FAIR Data in Medical Research

Incorporating the FAIR Principles in the Research Data Life Cycle

Kersloot, M.G.

Publication date
2022

Chapter 9

General discussion

This thesis set out to investigate how the FAIRification steps (i.e., steps to make data and metadata more Findable, Accessible, Interoperable, and Reusable (FAIR)) can be incorporated into the Research Data Life Cycle, ensuring that data are FAIRified throughout the research process rather than after project completion. First, we investigated clinical researchers' and research support staff's knowledge of the FAIR Principles and their current FAIRification efforts (Part I). Secondly, we explored the possibilities of Natural Language Processing (NLP) algorithms to make free text more Interoperable by linking machine-readable definitions to phrases in the text (Part II). Lastly, we developed, implemented, and evaluated a new FAIRification process to make data FAIR automatically upon collection, with a rare disease (RD) registry as a use case (Part III).

Main findings

Figure 9.1 summarizes our findings and conclusions, mapped to the Research Data Life Cycle.

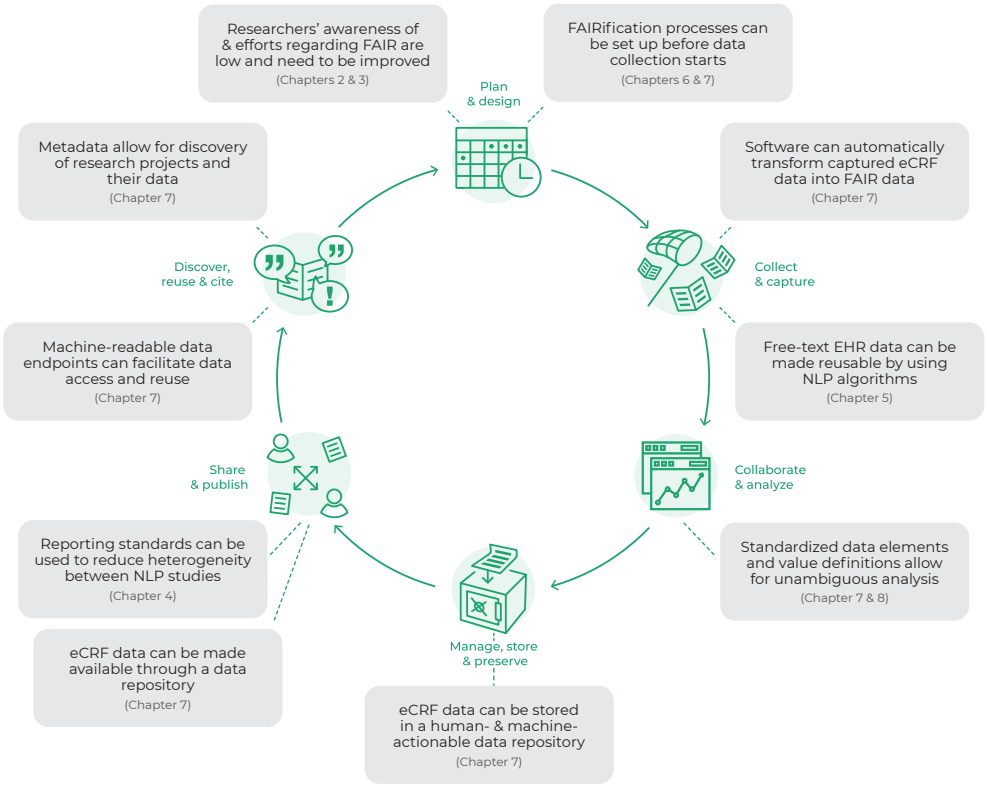

Figure 9.1: Summary of findings, mapped to the Research Data Life Cycle

FAIR: Findable, Accessible, Interoperable and Reusable, NLP: Natural Language Processing, eCRF: electronic Case Report Form, EHR: Electronic Health Record

Part I: State of FAIR

The FAIR Data Principles are being rapidly adopted by many research institutes and funders worldwide. However, little is known about the knowledge, perceptions, and efforts of individual clinical researchers and research support staff regarding data FAIRification. In Chapter 2 and Chapter 3 we developed an online questionnaire and distributed it to researchers and support staff in six Dutch University Medical Centers and Electronic Data Capture platform users. Our studies showed that a large number of researchers and staff are currently unaware of the definition of the FAIR Principles and their emphasis on both human and machine readability. We found that 60.5% of the respondents had heard of the FAIR Data Principles and, after explaining the FAIR Data Principles to the respondents, 11.0% of the researchers and 23.8% of the support staff stated that they have spent at least some effort achieving all aspects of FAIR for their data. Of all the FAIRification efforts mentioned by researchers, 93.9% focused on making data readable for humans and 31.2% focused on making data readable for machines. Nearly all researchers (94.5%) found the FAIRification of their data useful for others, and 89.3% would like to make their data FAIR when they have the opportunity and resources to do so. However, the majority of the researchers (81.6%) indicated that they require assistance throughout the process. Researchers need proper training, support, and tools to help them understand the importance of data FAIRification and guide them through the FAIRification process. Only then will machine-readable, FAIR data become a reality.

Part II: NLP and FAIR

Free-text descriptions in electronic health records (EHRs) can be of interest for clinical research, but cannot be readily interpreted by a computer. Natural Language Processing (NLP) algorithms can make free text machine-interpretable by attaching ontology concepts to it. In Chapter 4 we reviewed the current methods used for developing and evaluating NLP algorithms that map clinical text fragments onto ontology concepts. We found many heterogeneous approaches to the reporting on the development and evaluation of NLP algorithms. Of all the identified publications, 26.6% did not perform any evaluation. In addition, 28.6% of the included studies did not perform any validation, and 88.3% did not perform external validation. In order to standardize the evaluation of algorithms and reduce heterogeneity between studies, we developed a list of sixteen recommendations for future studies that can be used along with adherence to a generic reporting standard, such as TRIPOD [84], STROBE [85], RECORD [86], or STARD [87]. In Chapter 5 we developed and evaluated an application based on an existing NLP system (cTAKES). Our application includes generic algorithms for the detection of (misspelled) names of concepts and relationships between them. The application was evaluated by encoding free-text oncology charts and the evaluation results show that our application can detect oncology concepts and relationships with high precision and can detect recurrence with a significant increase in F1-score, compared to the original implementation of cTAKES. These concepts and relationships can be used to encode clinical narratives, and can thus substantially reduce manual chart abstraction efforts, saving time for clini-

cians and researchers.

Part III: FAIR by design

As concluded in Chapter 3, research software vendors should develop systems that help researchers and research support staff with the FAIRification process, and integrate them with systems that are currently used. In Chapter 6 we described the process of setting up a research project and making the project FAIR from its conception (de-novo FAIRification), and applied this process to a rare disease registry. An implementation of the process in an Electronic Data Capture (EDC) system, the place where data are often collected and stored in medical research, is described in Chapter 7. The International Society for the Study of Vascular Anomalies (ISSVA) ontology, developed in Chapter 8, provides machine-readable definitions for vascular anomaly diagnoses and is used in combination with the EDC system. Our implementation ensures that data entered on electronic Case Report Forms (eCRFs) in the EDC system are automatically transformed into machine-readable, FAIR data, without any intervention from data management and data entry personnel. The data are transformed into machine-readable data using a semantic data model (i.e., a canonical representation of the data, based on ontology concepts and semantic web standards) and mappings from the model to questions on the eCRF. The FAIRified data and metadata can be accessed through a FAIR Data Point (i.e., a metadata repository that provides access to metadata of digital objects in a FAIR way [210]) and queried through a SPARQL endpoint, which allows for the combination of data from various sources. The reusability and scalability of FAIRification across research projects can be considerably increased by implementing a de-novo FAIRification process in an EDC system.

Implications for science and practice

Raising awareness for FAIR

This thesis shows that the FAIR awareness of clinical researchers still leaves much to be desired. Despite funders' and institutes' mandates to make data (more) FAIR [25], we found that the majority of the researchers did not know what the Principles entailed and that only a very small number of the researchers actually make their data FAIR for machines. That is a major concern, because the European economy wastes more than €10.2 billion every year on 're-useless' data, with one of the primary causes being that researchers waste time on collecting, cleaning, integrating, and analyzing data, while data might already be available, but difficult to find or interpret [29]. The FAIR Principles could be a part of the solution to the problem of re-useless data [16], but achieving FAIRness of data starts with being aware of the meaning and practical implications of the Principles [269]. The research presented in this thesis quantifies this lack of awareness, and since the adoption and impact of the FAIR principles is an ongoing process [58], we believe that our work establishes a baseline of researchers' FAIR awareness upon which can be improved. To do so, institutions should establish data stewardship programs providing

simple and intuitive training for researchers [270]. In addition, the research community should translate the tacit knowledge that comes with the principles (e.g., rich metadata and persistent identifiers) into clear steps that can be verified using FAIR assessment indicators [271]. Even though educating researchers and research support staff on how to practice FAIR Research Data Management is a challenging task [25], it is one of the key elements to "turn FAIR into reality" [27].

Setting up clinical research projects

In this thesis, we found that researchers feel more confident in making their data FAIR if they are assisted in the process and are more willing to make their data FAIR if they do not have to spend any additional time on FAIRification tasks. By introducing our method to make research data collection FAIR by design (de-novo FAIRification), we strive to realize that researchers are assisted by FAIR data experts and relieved of technical data conversion steps. Where other work has focused on solely increasing interoperability [227, 238] or performing FAIRification in a post-hoc manner [239], our work automates data conversion steps and facilitates FAIRification and storing of FAIR data at the source. By doing so, collected research data will have more impact by design, since FAIRified data enables reduction of internal data silos and efficient integration with external data [66,272]. Domain and use case specific semantic data models are critical for this [20]: they provide machine-readable definitions of the collected data, allowing for unambiguous interpretation and analysis. The developed tooling described in this thesis opens up the possibility of FAIRifying related research efforts (e.g., studies in a specific disease domain) using a community-supported data model. Ultimately, this holds the potential to work toward reusability and availability for the 80% of collected research data that is today deemed re-useless [16] and the 17% of collected research data that has the chance to get lost every year [273]. Furthermore, it offers the perspective of optimal reuse of collected data and prevents other researchers from 'reinventing the wheel' because they were not aware of a similar project that was already carried out. Eventually, this all increases the value of the collected data and reduces waste and inefficiencies in biomedical research [274] and by practicing evidence-based medicine, the use of evidence that can be derived from these data can ultimately lead to better patient care [3].

Reusing free-text EHR data

An often-repeated, although never confirmed, speculation states that 80% of the data in EHRs is unstructured [275]. In this thesis, we have demonstrated that there is a large number of use cases for the application of NLP algorithms on these unstructured data and that NLP algorithms can be used to make free-text EHR data more interoperable via annotation with machine-readable ontology concepts. However, determining whether existing NLP implementations can be reused in different settings and for different use cases is difficult due to insufficient reporting on the development, evaluation, and validation of these algorithms [82,83]. Others will be able to determine whether the implementation can be reused for their context and use case once future research reports on it in

a standardized manner. This will, ideally, stimulate the reuse of existing NLP algorithms and implementations and lower the threshold for other researchers to start reusing free-text EHR data. When EHR data are annotated and structured with ontology concepts, one can leverage the relations between concepts for a variety of use cases. Examples for clinical research include finding patients who can be included in a clinical trial [73] and reusing EHR data from patients from multiple data sources [77]. Since making use of routinely-collected EHR data in research reduces research costs and burden on participants [276] and these data contain clinically relevant historical patient data and outcome measures, an accessible way to implement and reuse algorithms can be of great value for clinical research.

Sharing and reusing clinical research data

Clinical researchers have an ethical obligation to research participants to responsibly share data collected for their research project, for patients have voluntarily put themselves at risk to help generate information about the safety and efficacy of interventions [277,278]. However, there are many factors that frustrate these data-sharing efforts at individual, institutional, and international levels, such as the lack of time of researchers, lack of available data infrastructures, legal difficulties, and interoperability issues [279, 280]. The work presented in this thesis aimed to overcome these issues by offering a way to generate, store, expose, and query research data and their metadata in a FAIR manner, with limited intervention from researchers. Since clinical researchers often lack the time, knowledge, technical infrastructure, or financial resources to prepare and share their data [281], this semi-automated FAIRification process has great added value for the research community. After data are made FAIR, one can (programmatically) browse the metadata of research projects to find projects of interest, contact researchers working on them, and query the associated data directly from the source. Because the data are available in a machine-readable format, this querying can be automated, allowing for the real-time integration of research data with data from various sources, where the possibilities are endless [31,33]. All of this aligns with the vision of a global research community in which sharing de-identified data becomes the norm and where the knowledge gained from the efforts and sacrifices of clinical trial participants is maximized [282].

Strengths and limitations

The notion of making data FAIR for humans and machines is relatively new. A strength of our work is, therefore, our attempt to quantify the current FAIR awareness and FAIRification efforts of researchers using a questionnaire, something that – to our knowledge – was not done before. Since we recruited a large number of clinical researchers who have recently set up data collections for research projects, we got a good overview of the current 'gap' between the expectations of funders and institutes and the actual knowledge and skills of researchers. We, then, aimed to bridge this gap using the developed FAIRification workflow and tooling.

An additional strength is our approach to map out how NLP algorithms are developed and evaluated and how they are subsequently reported on. By using a large number of databases for our search, we made sure we included publications from many different sources (e.g., medical journals and computer science conferences) and domains (e.g., computer science or specific medical domains). Moreover, the proposed list of recommendations is grounded in the literature by harmonizing existing statements from domain-accepted reporting standards.

Another important strength is our methodology for developing the de-novo FAIRification workflow. By using a patient registry as a use case, we were able to develop, evaluate, and improve the workflow using real-world examples and issues. Furthermore, the workflow was created by an interdisciplinary team (consisting of local and FAIR data stewards, an EDC system specialist, clinicians, a patient advocate, a software developer, a healthcare interoperability expert, and an access protocol and privacy preservation expert), making sure the steps in the workflow relate to the entire process of and all disciplines required when setting up a research project.

A possible limitation of the studies focused on NLP that are described in this thesis is the lack of use of state-of-the-art methods, such as graph and word embeddings. These methods could have provided better outcomes than a more classical NLP algorithm. However, the results of our systematic review show that these methods are not broadly applied yet and that future research into these methods is needed.

Another potential limitation of our work is that the developed FAIRification workflow still requires some technical expertise at the start. Despite the fact that this task is expected to be done by research support staff with stronger technical skills, we recognize that this might still be a barrier to implementing de-novo FAIRification in practice. Future work should, therefore, concentrate on making the created technology in a more "plug and play" manner.

Lastly, all our studies were performed in the Netherlands, albeit embedded in an international context, e.g., by focusing on NLP on English clinical notes, on international literature, and on international registries and vocabularies. Although we believe our results are generalizable to other countries, there may be differences between the Dutch and other research communities. Because the FAIR Principles had their origins in the Netherlands, it is reasonable to assume that the Netherlands has a higher rate of acceptance and awareness of the Principles than other countries. Furthermore, FAIR data stewardship is increasingly being embedded into the organizational structures of Dutch research institutes, but this may not be the case in other countries.

Future perspectives

In the data-driven domain of healthcare, it should be common practice to make sure that collected data and their associated metadata can be found, accessed, interoperated, and reused, by others. Ideally, the research community should work toward an Internet of FAIR Data and Services (IFDS), a virtual space where machines and people can do exactly that in a trusted, affordable and sustainable way [283]. Figure 9.2 shows a proposed

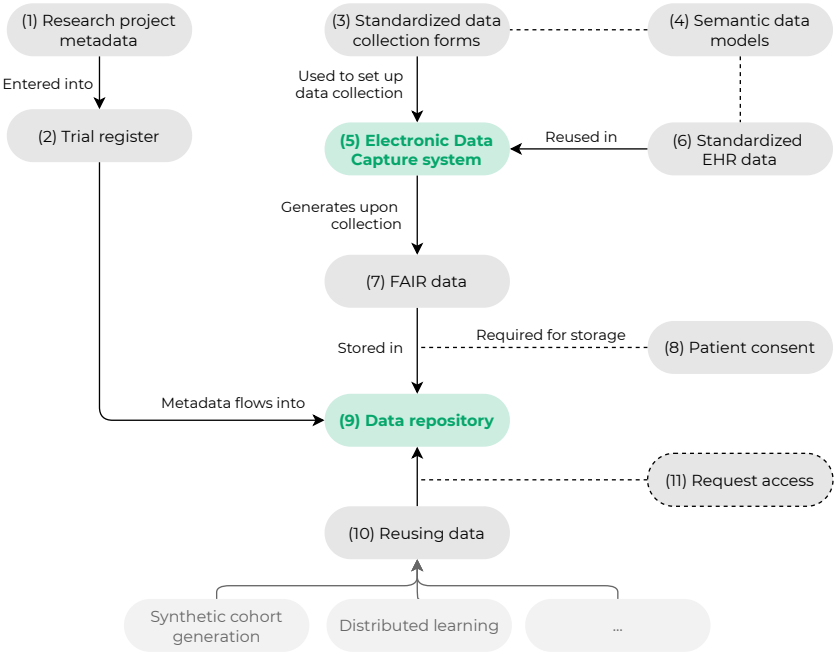

Figure 9.2: Future perspectives

FAIR: Findable, Accessible, Interoperable and Reusable, EHR: Electronic Health Record

future state of medical research, where FAIRified, machine-readable data and metadata are the norm, rather than the exception.

Standardized research project metadata (Figure 9.2.1, e.g., study design, objective, and contact information) are essential to making biomedical data findable and reusable for downstream analyses [284]. If these metadata are made available in a human- and machine-readable manner (Figure 9.2.2), one can easily reuse them in a data repository and link them to data collected in the project (Figure 9.2.9). Researchers should let go of one-off, non-standardized data collection forms and use versions that are standardized for their institute or research domain (Figure 9.2.3). Examples of standardized form elements that are currently available are Common Data Elements (i.e., data collection units comprising one or more questions together with a set of valid values [285]) and examples of standardized forms include CDISC's CDASH standard [286] and forms available on the Portal of Medical Data Models [287].

To ensure interoperability, these forms should contain a machine-readable definition in a semantic data model (Figure 9.2.4). Some standardized data elements and forms already have an underlying data model that is linked to ontology concepts, such as a part of the CDEs available in a repository from the National Institutes of Health. However, this is not the case for the majority of these data elements, and for less-common data elements, such as for disease-specific information, there are no standardized forms, nor data models available. The research community should work together with seman-

tic data experts to develop these standardized data elements, forms, and semantic data models. By doing so, the time needed for setting up the data collection into an EDC system (Figure 9.2.5) can be significantly reduced. The machine-readable definitions in the interoperable data models ensure that standardized EHR data (either at the source or using NLP methods) can be imported in the EDC system. Moreover, the semantic data model contributes to making data collected in the EDC system FAIR, and thus machine-readable, upon collection (Figure 9.2.7), enabling data FAIRification at scale. Patients should, throughout and after finalization of the research project, be able to decide if data pertaining to them may be reused for other purposes (Figure 9.2.8). Their consent should also be made machine-readable, such that data can automatically flow to the data repository upon consent, or automatically be removed from the repository when the consent is withdrawn. Dynamic consent, in which individuals can revisit and review consent decisions and preferences over time via a secure digital portal, could be a potential solution for this [288]. Once the data and associated metadata are available in the data repository (Figure 9.2.9), others are able to find the metadata and potentially reuse the data (Figure 9.2.10). There should be a workflow for cases where requesting access and signing a data use agreement is required (Figure 9.2.11). Again, the workflow's output should also be machine-readable, so that this process can be partially automated and integrated in the software that researchers use. The possibilities for reuse are endless, but may include the generation of synthetic datasets, enabling pool-and-reuse studies by preserving the essential properties of the data [289] (e.g., using synthetic control arms in randomized controlled trials [290]), or federated learning, where data remain in their original location, and analytical tasks visit data sources and execute the tasks [64] (e.g., the Personal Health Train, that was demonstrated to be capable of analyzing data from various healthcare institutions in different countries [291]).

Standards and standardized processes are crucial in the envisioned future of medical research outlined above. They support the optimal use and reuse of research data. Instead of 'reinventing the wheel' (e.g., developing new standards, forms, and software), the first step of the medical research community should be to join forces to create, implement, and use these standards and processes in practice.

Conclusion

The FAIR Principles are getting more and more traction in the medical domain, despite the fact that many researchers are currently unaware of the (definition of the) Principles or how to apply them to their projects. Incorporating the FAIRification steps into the current way in which researchers work, the Research Data Life Cycle, guides them through the FAIRification process and helps them to contribute to good data management and stewardship. User-friendly tools that are integrated with the software used by researchers play a crucial role in this. These tools facilitate automated data FAIRification throughout the course of a research project and take the majority of the work off researchers' shoulders. The work presented in this thesis quantifies the lack of FAIR awareness of clinical researchers and research support staff and provides a FAIRification

workflow and tooling that can assist them in FAIRifying their data. To accelerate FAIR-ification at scale, policy makers, funders and research institutes need to work together to provide standards, methods, support, tools, and funding to the research community, effectively making FAIRification of research data a joint mission. Ultimately, this paves the way for a future where collected data bring maximal value to patient care.