FAIR Data in Medical Research

Incorporating the FAIR Principles in the Research Data Life Cycle

Kersloot, M.G.

Publication date
2022

Chapter 1

General introduction

Data are the foundation of modern medicine: they contribute to building evidence that flows into the published body of scientific knowledge [1]. In the last decades, it has been more common to combine this evidence with clinical expertise and patient preferences to provide the best care to patients (i.e., evidence-based medicine) [2, 3]. New evidence (e.g., for an unstudied medical condition or new treatment method) is obtained by collecting, analyzing, and contextualizing data in a research project and disseminating the resulting findings in a scientific article [4]. An essential part of such a research project is Research Data Management: the organization of data, from the start of a research project through to the dissemination and archiving of valuable results [5].

The Research Data Life Cycle

Research Data Management is a continuous process. The Research Data Life Cycle (Figure 1.1) describes this process by combining research processes and activities with concepts of data management (i.e., the curation, preservation, publishing, and sharing of data) [6].

The steps in the Research Data Life Cycle can be described as follows [6, 7]: first, one *plans and designs* their research project. This includes defining the study design, determining how data will be handled throughout the project, and setting up the data

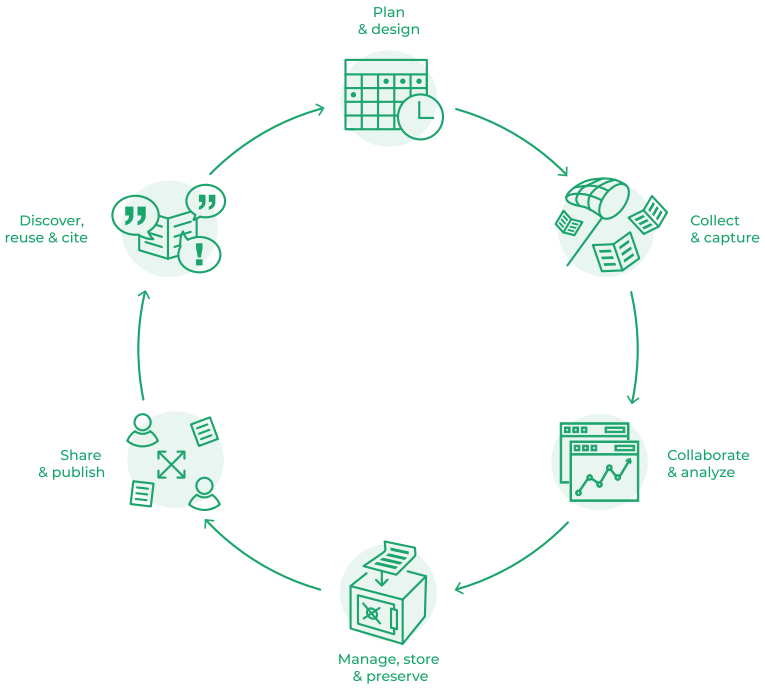

Figure 1.1: Research Data Life Cycle, based on [6–8]

collection process. Next, one *collects and captures* data. Originally, this phase included collecting patient data on paper-based Case Report Forms (CRFs) and entering (capturing) that data in an electronic format (database). Nowadays, data collection and data capture are commonly used interchangeably to describe the same procedure: capturing data of a patient in electronic format on electronic Case Report Forms (eCRFs). After the data are collected, a researcher *processes and analyses* the data. In this phase, the results of the analysis will be interpreted and new research output will be generated. Then, the researcher *manages, stores, and preserves* the data. This includes the migration of the data to a format and medium that can be used in the future, the creation of metadata (i.e., data about the data, such as contextual and contact information or a data dictionary describing the format of and variables in a dataset), and the archiving of the data. Subsequently, the researcher *shares and publishes* the findings of the research project in a scientific paper. At this point, the data should ideally be published, either open (i.e., data available as a supplementary file or via an online repository) or restricted (i.e., data not available for externals or available on request). All these steps contribute to other researchers being able to *discover, reuse, and cite* the collected data for their research.

The need for data sharing and reuse

The last two steps of the Life Cycle, *share and publish* and *discover, reuse, and cite*, are drivers of evidence-based medicine [9] and progress in science [10]. Data sharing and reuse can accelerate discoveries by avoiding duplicative trials and stimulating new research ideas [11,12]. It enables others to conduct additional analyses and replicate published findings, as well as combine data from various studies and aggregate it for meta-analysis [11, 12]. Researchers and knowledge consumers are increasingly aware of this need [13], and a recent survey from Springer Nature among researchers from a variety of fields shows that 61 percent of the medical researchers share data in some way [14]. However, it is estimated that 80 percent of the data that researchers collect and share are 're-useless' since most datasets are not machine-actionable (i.e., data that can be resolved on the web by web services [15]) [16]. Making data machine-actionable is becoming more urgent [17], since humans are unable to operate at the scope, scale, and speed required by today's growing body of knowledge and amount of available data [18]. Machines, on the other hand, can process these data in a more efficient and scalable manner, and allow for (re)use of the data in other systems and analysis workflows [17,19].

FAIR Data Principles

In 2016, a diverse group of stakeholders representing academia, industry, funding agencies, and scholarly publishers published the FAIR Data Principles, focusing on machine-actionable (meta)data [18]. These Guiding Principles state that research data and associated metadata should be made Findable, Accessible, Interoperable, and Reusable (Box 1.1), both for humans and machines [18]. Since their publication, the Principles have

Box 1.1: Description of the aspects of FAIR [18], based on [18–21]

The FAIR Guiding Principles refer to three types of entities: data, metadata (information about the data), and infrastructure.

Findable: Datasets should be described, identified and registered or indexed in a clear and unequivocal manner

→ The first step in (re)using data is to find them. It should be possible for other researchers to find (metadata about) datasets in repositories. Metadata provide more details on datasets (e.g., the study population, applied interventions, and location) and allow for searching and filtering. Globally unique persistent identifiers (i.e., identifiers that are never reused in another context, and continue to identify the same resource, even if that resource no longer exists, or moves) and machine-actionable metadata ensure that datasets and services can be found and resolved automatically.

Accessible: Datasets should be accessible through a clearly defined access procedure, ideally using automated means. Metadata should always remain accessible.

→ Once a researcher or machine finds (meta)data, they need to be able to retrieve them using a standardized protocol (e.g., the HTTP protocol used on the World-Wide Web). FAIR Data does not necessarily mean open data, therefore, researchers and machines need to know how and if data can be accessed. This includes data access restrictions and authentication (i.e., assuring that the access requester is indeed that requester) and authorization (i.e., assuring that the requester's profile and credentials match the access conditions of the resource) protocols.

Interoperable: Data and metadata are conceptualised, expressed and structured using common, published standards.

→ Data usually need to be integrated with other data to bring more value. In addition, the data need to interoperate (i.e., the ability of data or tools from non-cooperating resources to integrate or work together with minimal effort) with applications or workflows for analysis, storage, and processing. To do so, there should be a clear representation of the data to make sure that each data item that is the same in multiple resources is interpreted in exactly the same way (e.g., *blood pressure, measured while sitting* in a dataset from an American researcher should be interpreted the same as *bloeddruk in zittende positie* in a Dutch researcher's dataset). Ontologies (explicit formal specifications of the concepts in a domain and relations among them [22]) can provide these representations for data items.

Reusable: Characteristics of data and their provenance are described in detail according to domain-relevant standards, with clear and accessible conditions for use.

→ The ultimate goal of FAIR is to optimize the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings (e.g., one should add descriptions of how, why, and by whom the data was collected, who owns the data, and how the data should be cited). Moreover, humans and computers need to be able to assess if discovered data are appropriate for reuse, given their research question (e.g., a dataset with data from *0-18 year old patients* with *haemangiomas* is of interest to a researcher looking for datasets on *vascular anomalies* in *children*). A clear (meta)data usage license helps humans and machines assess under which conditions the (meta)data can be used (e.g., CC0 [23], where there are *no rights reserved*). The license may be different for the metadata and the data itself.

experienced a significant surge in acceptance and implementation by researchers and research institutes [24]. In addition, funding [25] and government bodies [26, 27] are raising awareness of the importance of making data FAIR. Most funders now require researchers to document their FAIRification methodology as part of a Data Management Plan (DMP) [25], since such a sound and elaborate plan is a necessary precursor to making data FAIR [28]. The European Commission also emphasized that "we have come to realize ourselves how important it is to have FAIR research data" and estimated that not having FAIR research data costs the European economy at least €10.2 billion per year [29].

When data are made FAIR (*FAIRified*), it will become much more straightforward to discover data over distributed sites and accurately integrate those data or analyze them by 'data visiting' (i.e., analyzing data across multiple data sources without individual patient data leaving the source [30]) [31–33]. Furthermore, data FAIRification is expected to speed up innovation, lower costs, and enable data sharing between research groups within and across institutions and companies [17].

Incorporating the FAIR Principles into the Research Data Life Cycle

FAIRification workflows, both generic [19] and specific (e.g., workflows for health research [34], observational COVID-19 [35], and nanosafety data [36]), have been developed over the years to make data and metadata FAIR step by step. The workflows emphasize that making data and metadata FAIR is an iterative process: every step of a FAIRification workflow attempts to enable the implementation of the FAIR Principles and aims to enhance the *FAIRness* (i.e., FAIR status) of the data and metadata [19, 34]. Currently, these FAIRification workflows are designed to be executed after research projects have been already conducted and data are collected, rather than throughout the life cycle of a research project. However, the FAIR Principles are designed to guide the implementation of good data management and stewardship [18, 37], practices that encompass the entire research process. Data management concerns all operational data-related activities throughout the Data Life Cycle and data stewardship refers to the assignment of responsibilities in, and planning of, data management [38]. We, therefore, believe that the application of the FAIR Principles should also be incorporated into the Research Data Life Cycle.

Aims and outline of this thesis

The aim of this thesis is to incorporate the FAIRification steps into the Research Data Life Cycle, ensuring that data are FAIRified throughout the research process rather than after project completion. To investigate this, we distinguish three strands of research, which are addressed in the three parts of this thesis. First, we aim to gain insight into researchers' knowledge and perspective on the implementation of the FAIR Principles in practice (Part I). Secondly, we aim to determine if Natural Language Processing (NLP) can be used to make data more FAIR (Part II). Lastly, we aim to develop a process for

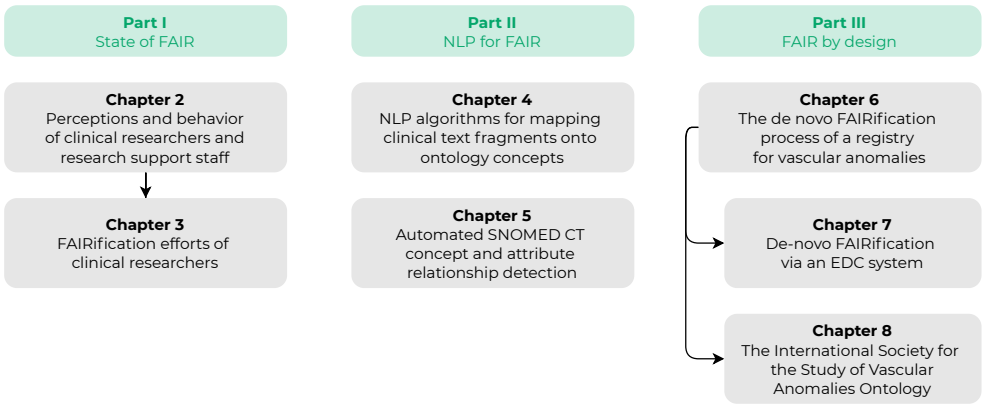

Figure 1.2: Outline of this thesis

FAIR: Findable, Accessible, Interoperable and Reusable, NLP: Natural Language Processing

making data FAIR from the beginning of the research project and at the source (Part III). Figure 1.2 illustrates an outline of this thesis and the relationships between the chapters.

Part I: State of FAIR

The FAIR Principles provide guidance for research data management, but are not a standard and do not offer explicit steps for achieving FAIRness [16]. The Principles must be interpreted for each use case, and technical implementation decisions must be made accordingly [39]. The recently developed FAIRification workflows can help researchers with that. However, researchers still require specific resources and expertise throughout the FAIRification process that they might not have themselves [19, 34, 39]. There is currently little information available about clinical researchers' familiarity with the FAIR Principles, as well as their experience with FAIRification workflows. In Part I (Chapters 2 and 3), we, therefore, gain insight into clinical researchers' and research support staff's knowledge of the FAIR Principles and their current FAIRification efforts using a questionnaire. The results will allow us to determine what additional resources and expertise researchers and support staff may need. Chapter 2 describes the questionnaire conducted among clinical researchers and research support staff and assesses their awareness and attitudes regarding data FAIRification. Chapter 3 uses the data collected in the previous chapter to describe the current FAIRification efforts of researchers and their understanding of the Principles.

Part II: NLP and FAIR

Clinicians currently document clinical findings in the Electronic Health Record (EHR) primarily in free-text notes, because they are unable to fully express complex findings and nuances of each patient in a structured format [40, 41]. However, these notes currently have limited value for clinical research, since free-text data cannot be easily processed

by a machine (Interoperable), nor easily reused by others. Natural Language Processing (NLP) might aid in making these notes more Interoperable, for NLP algorithms can (semi-)automatically process text and these algorithms can perform entity linking (i.e., mapping free-text phrases to ontology concepts: machine-readable definitions of concepts in a particular domain). For example, the phrase *systolic blood pressure (while sitting)* can be mapped to SNOMED CT concept *407554009*, a concept that represents a *observable entity* that measures the *pressure* in a *structure of cardiovascular system* in the *systolic phase* in a *sitting position*. By performing entity linking, one makes parts of the notes machine-readable, therefore, easier to analyze. This fosters the reuse of data available in the notes, which is a step toward more FAIR data. In Part II (Chapters 4 and 5) we explore the possibilities of using NLP algorithms to make free text more Interoperable (the I in FAIR) by linking machine-readable definitions to phrases in the text. Chapter 4 reviews the current state of the development and evaluation of NLP algorithms for entity linking and proposes a structured list of recommendations for future studies. Chapter 5 describes the development and evaluation of an NLP application for the detection of concepts and relationships between concepts in free text.

Part III: FAIR by design

As mentioned earlier, existing FAIRification workflows are usually carried out *post hoc*: after the research project is conducted and data are collected [19,34], rather than throughout the life cycle of a research project. In Part III (Chapters 6, 7, and 8) we describe the process of de-novo FAIRification, in which the FAIRification steps are incorporated into the Research Data Life Cycle of a research project. This ensures that data are made FAIR automatically and in real-time, upon collection without any intervention from data management and data entry personnel. By doing so, the reusability and scalability of FAIRification across research projects can be greatly improved. The Registry of Vascular Anomalies (VASCA), a rare disease (RD) registry that is part of the European Reference Network (ERN) on Rare Multisystemic Vascular Diseases (VASCERN), is used as a case study. Chapter 6 presents a workflow for *de-novo* FAIRification: a workflow where FAIRification steps are incorporated in the process of setting up and collecting data for a registry or research project. Chapter 7 describes the technical implementation of Chapter 6's workflow in an Electronic Data Capture (EDC) system, the place where medical research data are often collected and stored via electronic Case Report Forms (eCRFs). Chapter 8 presents the International Society for the Study of Vascular Anomalies (ISSVA) ontology, introduced in Chapter 6: a machine-readable representation of a classification for vascular anomalies that enables medical specialists to register diagnoses in a FAIR manner.

Finally, Chapter 9 summarizes the main findings and provides an overall discussion of the work presented in this thesis.