FAIR Data in Medical Research

Incorporating the FAIR Principles in the Research Data Life Cycle

Martijn G. Kersloot

Obtaining a PhD?

For the past four and a half years, I was a *PhD candidate* doing research on how we can help other researchers make the data they collect in their research reusable. In this *layman's booklet*, I explain more about what doing a PhD entails and what my research was about.

What is a PhD trajectory?

During a PhD trajectory, you perform **scientific research.** The results of the various research projects you conduct are written down in articles that are then published in scientific journals.

These **publications** are then bundled into a **dissertation/thesis**, along with a general introduction and discussion about the research.

This dissertation will then be defended in front of a **committee**. After a successful defense, a person may assume the title of '**doctor**' (Dr. or PhD).

Who are the key stakeholders in a PhD trajectory?

The **PhD candidate** is supervised by one or two **supervisor(s)**. These are professors or associate professors who are responsible for a specific line of research at the university. The PhD candidate is usually supervised on a daily basis by one or two **co-supervisor(s)**. They also hold a PhD themselves and are usually very closely involved in the research project.

Prior to the defense, a **doctoral committee** assesses whether the dissertation is of sufficient quality. The committee consists of a number of professors or experts who are knowledgeable about the subject for which the doctorate will be granted. The defense may not take place unless they have given their approval. During the defense, they ask the PhD candidate a series of questions about the dissertation's contents.

Paranymphs, who are generally friends or colleagues, support the PhD candidate throughout the 'big day'. They used to help the candidate answer tough questions or even stand in for them if they were ill, but currently they mainly fulfill a ceremonial role. They help with the preparations and sit next to the PhD candidate during the defense to read one of the propositions or a passage from the thesis.

Finally, there is the **beadle**, who makes sure that everyone adheres to protocol during the defense.

What is the defense ceremony like?

PhD defenses always start **on the hour**. Just before the ceremony starts, the paranymphs make a number of announcements (including the fact that clapping is not permitted during the ceremony!). On the hour, the PhD candidate starts with their **layman's talk**, a short presentation explaining what the thesis is about. It is basically a spoken version of this layman's book!

The defense formally begins at **a quarter past the hour**, with the committee (professors in gowns) entering and the chair opening the meeting. The PhD candidate reads a formal text, after which the discussion is opened. During the defense, members of the committee are addressed as **highly esteemed** (professor) or **highly learned** (doctor) opponent. After three-quarters of an hour, the beadle comes in, exactly on the hour, saying **"hora est"** (it's time). The PhD candidate will then read out a formal text and the committee will then withdraw for **deliberation**.

After 5 to 10 minutes, the committee, led by the beadle, returns with the verdict and (hopefully!) the **diploma**. The diploma is handed over and this is followed by the **laudatio**, a personal message from the supervisor. After this, the session is ended (you can applaud now!) and it is time for the reception.

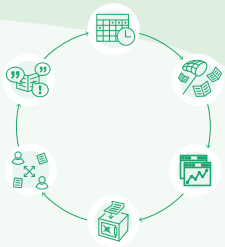
Background

Evidence-based medicine

Modern medicine is largely based on the **results of clinical research**. An example of a clinical research project is determining which medicine works best against a certain disease.

When you see a doctor with a complaint, he uses his clinical knowledge, research findings, and your personal preferences to provide you with the best possible care. This is called **evidence-based medicine**.

Research Data Life Cycle



A clinical research project can be divided into different phases. In all phases of the project, **Research Data Management** is crucial. It focuses on how do should handle the data you collect in your research project. The **Research Data Life Cycle** describes the phases of a research process and the associated data management tasks.

First of all, you **plan and design** your project: how will you conduct your research and what data will you collect? You then **collect data** from the patients who participate in the project (for example, the blood pressure or the weight and height of the patient). After these data have been collected, you **analyze** them (for example, in a drug study: is there a difference in the group of patients who did receive a drug and the group that did not?). You **store** these data and results so that they can be used later. Finally, you ***share and publish*** your findings in a scientific article, ideally also including the data you have collected. Others can then **find and reuse** that data for their research.

Data are merely archived, never to be used again

The last step of the Research Data Life Cycle, **finding and reusing data** is very important: it advances science and, above all, ensures that other researchers do not "reinvent the wheel" by starting a research project on something that has previously been studied. However, in the majority of the cases the acquired data for completed clinical research studies are merely archived and never reused.

Because the majority of research data is poorly documented (for example, what data was collected under what circumstances), it is estimated that **80%** of the collected data in clinical research cannot be reused at all.

The FAIR Principles

The FAIR Principles were created in 2016 to address this reusability issue. They state that research data, research information, and associated data (metadata) must be FAIR for **other researchers** (people) and for **software** that analyzes the data (computers).

Findable
Well described what the (research) data is about

Accessible
Clear whether and how you can access the data

Interoperable
Ready to be combined with other datasets

Reusable
Clearly described how the data can be used

Aim: Include the steps to make research data FAIR in the Research Data Life Cycle

Many of the tasks a researcher performs on a daily basis interface with the steps required to make research data more FAIR (FAIRification).

In my thesis, we, therefore, investigated how we can **include these steps into the way researchers work**: the Research Data Life Cycle.

Contents

My thesis consists of three different parts.

In **Part I** *(State of FAIR)*, we examined what researchers know about the FAIR Principles and think about making their own research data FAIR.

In **Part II** *(NLP and FAIR)*, we explored how Natural Language Processing (a way of letting a computer analyze 'free' text) can help in making data FAIR.

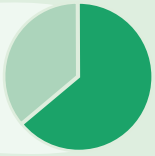
In **Part III** *(FAIR by design)*, we developed a process to make data FAIR, immediately upon collection.

Part I
State of FAIR

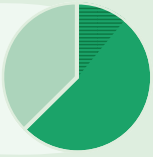
Chapters 2 and 3

The chapters in this part describe the results of a questionnaire sent to clinical researchers. In the questionnaire we asked what the researchers knew about the FAIR Principles and what steps they were taking to make their data more FAIR.

64.0% of the surveyed researchers had **heard of the FAIR Principles before**.



62.8% spent at least some effort to achieve **any aspect** of FAIR.
11.0% spent effort to achieve **all aspects**.
(findable, accessible, interoperable, and reusable)



164 researchers completed the questionnaire

93.9% focuses on FAIRification **for humans**, and just 32.1% on FAIRification **for computers**.

Training, (financial) support, and tools

are needed to help researchers make their data more FAIR.

35,1% indicate indicate that they can **make data FAIR themselves** and 81,6% indicate they **need help** with this.

Part II
NLP and FAIR

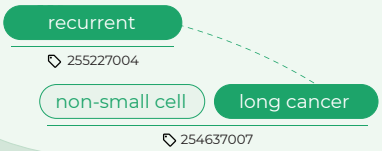
Chapters 4 and 5

This part contains a chapter with a literature review on the use of Natural Language Processing (NLP: the 'understanding' of text by a computer) in healthcare and a chapter describing the development and testing of such an NLP application.

77 articles reviewed

Big difference in the testing of algorithms and description of them in articles

therefore, we made **16** recommendations for future research



The developed application detects medical terms and their --- interrelations in text, which can reduce the need for manual reading of patient records for research purposes.

Part III
FAIR by design

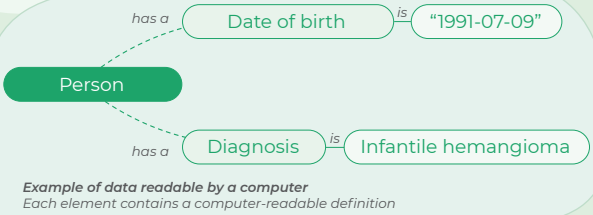
Chapters 6, 7 and 8

This section describes a new process that enables automated FAIRification of data. The process is integrated into the process of setting up a research project, so that researchers ultimately spend less time on making their data FAIR.

Registry of Vascular Anomalies

The **FAIRification process** developed in this part has been applied to a **registry** *(a research project that collects data from a specific group of patients over a long period of time)* **for vascular anomalies**. Because these types of abnormalities are **rare**, it is very important to collect more data and thus information about them. Hopefully, better treatments can be developed using this knowledge and data.

After patient data is entered in the registry, these data is automatically converted into a format that is FAIR and **readable by computers**. The format looks like a mind map in which the data of the patient and descriptions of how that data are related to the patient are recorded.



Example of data readable by a computer
Each element contains a computer-readable definition

Because computers can read and 'understand' this particular format, it is possible to **analyze and combine data from multiple sources** on a big scale.

The process is integrated into the software that researchers use to collect data, so data is made FAIR 'at the source'.

Chapter 9

Discussion & Conclusion

The seven chapters of the thesis described how researchers can make the data they collect more FAIR. A number of conclusions drawn in the thesis are listed below.

Many researchers are currently **unaware** of the (meaning of) the FAIR Principles.

Processes to make research data FAIR can be set up **before the research project starts.**

Software can **automatically make data FAIR** upon collection.

Text from patient records **can be made reusable** through Natural Language Processing.

FAIRified research data **can be made available automatically** so others can reuse them.

To accelerate FAIRification 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.

Cover: Nadine Verhoek

Images: Freepik
starline, macrovector, vectorjuice, jcomp, pikisuperstar, pch.vector

Propositions

Propositions accompanying the thesis are a number of claims made by the PhD candidate which they want to defend against the doctoral committee. The majority is related to the research that has been conducted. In addition to the "serious propositions", there are also a number of funny ones.

1. Rather than requiring researchers to make their data FAIR, funders and institutions should first focus on raising FAIR awareness and providing researchers with the right support and tools.
   This thesis, chapter 2 and 3

2. It is important to demonstrate the usefulness of FAIR data to researchers since this will positively influence their attitude and intention to make their own data FAIR.
   This thesis, chapter 2

3. Researchers should prioritize making their data machine-readable over making it solely readable for humans.
   This thesis, chapter 3

4. The adoption and reuse of NLP algorithms in healthcare can be greatly improved if these algorithms were evaluated and reported on in a uniform manner.
   This thesis, chapter 4

5. To support FAIRification at scale, de-novo FAIRification workflows must become the standard.
   This thesis, chapter 6 and 7

6. Integration of FAIRification workflows and tools into the Research Data Life Cycle is essential to make data FAIRification more accessible and understandable for researchers.
   This thesis, chapter 9

7. One could devote an entire PhD trajectory to the puns that can be made with the acronym "FAIR".

8. For many researchers, including those who focus their entire PhD trajectory on this subject, FAIR data feels like something that is quite FAIR-fetched.

9. Reducing waste in research: I want it (meta)data way.