FAIR Data Advanced Use Cases:

from principles to practice in the Netherlands

(Preliminary) Report dated April 2018

Colophon

Contents

Executive Summary

The idea that data needs to be **F**indable, **A**ccessible, **I**nteroperable and **R**eusable is a simple message which appeals to many. The 15 international FAIR principles were published in 2016. They serve as a guideline for preparing research data for reuse under clearly described conditions by both people and machines. They are intentionally principles and not standards. Various organisations and disciplines have since developed standards, tools and training based on their own interpretation of the FAIR principles.

The six use cases included in this report describe FAIR data developments and different approaches taken within different domains. For SURF, it is important to gain a better picture of the best way to support researchers who want to make their data FAIR. This can serve as a starting point to develop the infrastructure and services needed for FAIR data.

The use cases show that implementing the FAIR principles is seen as a series of improvements. There are always steps ahead that can improve reuse even further. Machine readability is sometimes one of those next steps; there is a tendency to focus on human interoperability first. Other preliminary conclusions that can be drawn from this report are:

1. FAIR is seen as part of a larger culture change
FAIR is seen as part of a larger culture change towards more openness in research and interdisciplinary cooperation. Together with developments such as new national and international privacy regulations and policies, FAIR highlights the need to update policies and to invest in support and awareness activities, new infrastructures, software and tools.

2. There is a tension between domain specific needs and maximum interoperability
No matter the FAIR maturity of the community, there is a tension felt between trying to build on existing domain-specific principles and workflows on the one hand, while trying to get to a maximum level of cross domain interoperability on the other. Trying to get consensus on minimal cross domain standards, and sharing FAIR examples from different domains to get an understanding of the potential to align standards and workflows are seen as ways to overcome this tension.

3. Policies can't be about FAIR compliance alone
It is not easy to derive a set of metrics from the principles, especially in the non-life sciences domains. Any policy in which FAIR is mentioned should be open to discipline specific solutions which at least satisfy the overarching requirements of **F**indable, **A**ccessible, **I**nteroperable and **R**eusable.

4. A way forward: integrated approaches with domain specific guidance
To make it as easy as possible for researchers, it is said that the FAIR data principles need to be translated into more practical guidance. There is a tendency to take an integrated approach when doing so, in which domain-specific needs are leading.

5. FAIR takes effort, but it is worth it
It takes effort to get to a certain level of FAIRness, but some use cases show that once you have reached that level of FAIRness, a whole world of possibilities opens.

6. The Future: recommendations for further exploration
Open questions are often linked to interdisciplinary interoperability, and the long-term financial business case for the implementation of FAIR. How much effort should go into preparing data for re-use and long-term preservation of datasets? These issues need further exploration**.**

Introduction

The FAIR principles
The idea that data needs to be **F**indable, **A**ccessible, **I**nteroperable and **R**eusable is a simple message which appeals to many. It is also clear that FAIR helps to explain the importance of machine-readable data and good data curation to a wider audience and clarifies the great potential, the possibilities that these create for all researchers.

The international FAIR principles were formulated in 2014 during a workshop at the Lorentz Center in Leiden. Two years later, after a round of open consultation via the FORCE11 platform, the 15 FAIR guiding principles were published. These principles have now been broadly acknowledged across the international research data management community.

Various organisations and disciplines have since developed standards, tools and training based on their own interpretation of the FAIR principles. Some domains have already done a great deal of work on this, although not always under the FAIR banner. Other domains do not traditionally use large quantities of research data and are at an earlier stage. This wide diversity makes it a challenge to present a clear overview of the current implementation of FAIR principles in the Netherlands.

The principles serve as a guideline for preparing research data for reuse under clearly described conditions by both people and machines. They are intentionally principles and not standards. This is because research data and the way in which these data are processed are different in each research domain. Different domains can use the FAIR principles as a basis to develop their own standards and ways of processing and publishing data.

The FAIR principles are already becoming more broadly known, including among policy makers. This raises one concern: there is a tendency toward policy and regulations requiring data to be made FAIR. Given the nature of the "FAIR guiding principles" – to help implementers of FAIR data check whether their particular implementation choices are indeed rendering the resulting data FAIR – they cannot simply be applied directly, this point is illustrated several times in these use cases.

What is the purpose of this report?
This report was drawn up as part of SURF's Open Science Programme. The purpose of this report is to build and share expertise on the implementation of FAIR data policy in the Netherlands. For SURF, it is important to gain a better picture of the best way to support researchers who want to make their data FAIR. A step in this direction is to map out examples of good practice that are already available in the Netherlands and what is still needed. This can serve as a starting point to develop the infrastructure and services needed for FAIR data.

How is the report built up?
The six use cases included in this report describe developments in FAIR data and different approaches taken within different domains and within a number of projects, institutes and university libraries. They illustrate the move from principles to policy and the development of standards for creating, processing, saving and using FAIR data.

These use cases are examples, but they are not instructions on how to make domain-specific data FAIR. SURF realises that the choice of these six use cases means that a lot has been left out. The number of use cases may be expanded in the future.

What do we mean by the FAIR principles?
For the sake of readability, this report uses the terminology of 15 FAIR principles, following the publication *The FAIR Guiding Principles for scientific data management and stewardship*. It is important to note that some interviewees take a slightly different approach to how the "FAIR principles" should be understood. They see the starting point that data should be **F**indable, **A**ccessible, **I**nteroperable and **R**eusable as the principles. The subsequent 15 points are then seen as 15 'Facets' or 'FORCE11 interpretations'. Where people in the use cases only refer to the overarching criteria that data must be **F**indable, **A**ccessible, **I**nteroperable and **R**eusable, that is stated explicitly.

What work method was followed?
Five of the six use cases are based on interviews with people involved. The sixth use case was delivered via the Dutch GO FAIR office as an example of their work in the biomedical and healthcare field. This use case is an abridged version of an interim report of a practical test conducted by the Dutch National Health Care Institute in cooperation with GO FAIR. This use case therefore has a somewhat different structure.

The interviews discuss the extent to which people had already been working with research data management before the term FAIR was introduced, what they see as the advantages of FAIR, what is currently being done to stimulate the FAIR work method within their domains, what difficulties they face and what is happening in terms of interdisciplinary developments. The future of FAIR is also discussed, how people look at the resources needed, among other things in terms of long-term data storage, and what further details and support are needed.

Openness
The use cases show that implementing the FAIR principles is a process. Certain developments, like complete interdisciplinary interoperability, are often still something for the future. The idea behind these use cases is to gain an idea of where people are on the road to FAIR data. It has become evident that a lot still needs to be done before we can talk about full implementation.

SURF thanks everyone who worked on these use cases, made time available to give a detailed vision of the implementation of the FAIR principles and shared their own experiences very openly. This gives a very interesting look behind the scenes at work in progress. This openness is greatly appreciated.

We hope that this report will help and inspire not only us but also others who are working out their own route toward the FAIR processing, storage and provision of research data; and that it will also increase understanding about the further developments that are necessary in this area.

Links:
https://www.nature.com/articles/sdata201618
https://www.force11.org/group/fairgroup/fairprinciples
Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud: https://content.iospress.com/articles/information-services-and-use/isu824

The FAIR data guiding principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

The Research Data Services at TU Delft and the 4TU.Centre for Research Data
The 4TU.Centre for Research Data is hosted by the Research Data Services team of the Library of the Delft University of Technology. It supports research data management and long-term archiving of research data output for the technical sciences in the Netherlands and internationally. Its work within TU Delft includes providing advice and help on Data Management Plans, for general project planning as well as for funding requests.

http://researchdata.4tu.nl/en/home/

TU Delft: 'We should try to get as much community-focused support in place as possible'

Jasmin K. Böhmer MSc
Research Data Officer
TU Delft

'Regarding FAIR data and research data management in general, we are in a highly fortunate position here at TU Delft. Three important elements are in place:
- the board of the university is advocating Open Science;
- we have sophisticated ICT and research support services including trusted data archives in place;
- we have more and more data stewards on campus.

These elements underpin the new TU Delft strategic framework for 2018-2024 which is '*Impact for a better society*'. FAIR data is part of enabling the best possible research data support for a range of technical and science subjects.

Our Research Data Services team has explored some of the implications of FAIR data. How to interpret FAIR data for technical and science data and subsequently how to adapt and improve our current services to support these needs? As one of the Front Office staff for research data support, I embrace the new concept of "FAIR data". It appears to be less intimidating than the notion of "open data" and therefore makes it easier to start a conversation about data curation.'

FAIR for researchers
'The research data management plan is the vehicle to make researchers work in a certain way. 'FAIR data' is not a buzz word here amongst researchers. When I talk to them, the concept of data management planning alone is already too abstract to really integrate the FAIR principles into a support session about research data management. At the beginning the recipients are overwhelmed with the data management plan concept and the related topics. They are not used to thinking within this framework and the feedback we get from researchers is that good data management is very time consuming.

I always try to get researchers to understand that if your data management is solid from the start of your research project you help yourself in the long run. But there are not yet enough tools or workflows in place to make it easy for them. That is what we will be working on in the coming years.'

FAIR compliance is trickier than it seems
'The first project I worked on as a Research Data Officer in 2016 was a paper scoping the FAIR readiness of repositories and data archives, predominantly in the Netherlands: *Are the FAIR Data*

Principles fair? The term FAIR data was not used or applied at TU Delft before this project. Open data on the other hand was already a familiar term here, due to the sympathy and support towards Open Science.

We focussed on FAIRness at the repository level. We picked 37 repositories, including DANS, 4TU.Reseach Data but also heavily used foreign repositories like Zenodo. We chose a traffic light system to give an idea of the status of compliance with all the principles. For this study we crafted our own interpretation of compliance: for instance, that you must use the HTTP(s) protocol and you must have clear licences. This was not always easy. As we stated in the paper, the brevity of the 15 FAIR principles gives the impression that they are all items that can be checked off, but some appear to overlap, some are vague, others are open ended, while others require interpretation from external parties. Also, some appear to be technical in scope, whereas others are more policy driven. When it comes to working with the guidelines, we must acknowledge this variation.

To make FAIR data mandatory is tricky. Compliance should not be a stick, but rather a desirable goal. Funders, including the European Commission, should take heed of this, to avoid the principles being misused and sanctions prematurely applied. Can you require FAIR data if support and agreement on interpretations are not yet in place? In our analysis it turned out that larger, certified and standardised archives like 4TU.Research Data and DANS are doing relatively well on compliance, whereas the subject based repositories were scoring lower. But the communities prefer to use the subject based repositories because they fulfil their specific needs. How can these subject based repositories be encouraged to improve FAIR compliance?'

Focus on communities

'One of the principal conclusions of this FAIR study was that we must focus more on research communities in the coming years. The priority is to ensure that as many datasets as possible are Findable and Accessible. To achieve this, we will focus on human readability first; machine readability is still miles away for some communities. Not that machine readable metadata isn't important. Currently most of the Dublin Core metadata applied by 4TU.Research Data is fully machine readable. But Dublin Core does not serve the subject specific needs of all scientific and technical disciplines. So it is common that the relevant information for making decisions on reuse and interoperability are "hidden" in the readme file, which currently is not machine readable and not searchable via the search engine of 4TU.Centre for Research Data.

A potential future development here at Delft could be working with different levels of FAIR data compliance for each community. With some communities which already score high on the I and R, we can try to progress this even further, whereas for most communities achieving the principles under F and A would be a significant step forward. By helping them and by promoting the **F**indability and **A**ccessibility of data in the archive, it should be possible to attract more researchers to use 4TU.Research Data. Subsequently the **I**nteroperability and **R**eusability of their data will be improved by adhering to the preservation and curation standards that the archive is applying to its archived data.'

Domain specific examples for each principle

'In collaboration with the Task Force for Open Science within the CESAER consortium of engineering institutions, we are working on examples for each of the FAIR data principles. Our goal is to write subject specific guidance for each of the principles for technical and science data. We will do this one discipline at a time, based on the work of Dutch Techcentre for the Lifesciences (DTL). The life sciences have a different approach to FAIR from the researchers here, so we are shaping different guidance for our domains. Interviewing researchers from different domains here

at TU Delft was very insightful. How do they work, which standards do they use and how do they see reuse? Some domains want to share raw data, where other domains don't want to publish any of their raw datasets.

Based on this CESAER work, I get a bit philosophical: before we can really look into implementing the FAIR principles, we need to define what reuse is. Is it cataloguing and indexing of metadata; reviewing data-sets in the same manner as literature; or actual application and integration in new research? That is still unclear for most people who want to make sure their data management adheres to the FAIR principles. Also, are the demands that data creators have the same as the data reusers? We don't know yet, but it is very interesting to work on issues like that.'

The Future of FAIR
'FAIR works as a driver for innovation. We are working on a new infrastructure with new features. Our dream is to have automated workflows available for different disciplines. It would be so helpful for everyone to use an environment where all the change tracking is done for you instead of having to track everything manually, but we don't have this available for every discipline yet.

For a real interdisciplinary approach, we need even more. A national data search platform that crawls DANS and 4TU.Research Data et cetera would be helpful: one simple interface where you can search for all kind of FAIR datasets. If we can attach that to existing national initiatives, that would be even better. In the Netherlands we should try to get community focused support for good RDM in place to overcome some of the barriers and to ensure it is easier for the individual researchers to find information.

There is still a lot of work to do. To get to real reuse of research data we need consensus at the international level about how to interpret the principles in a domain specific context. At TU Delft, we have data stewards who can focus on practical, subject-specific help, and therefore ease the implementation of FAIR data. This work is highly engaging and stimulating and FAIR is a real inspiration for me in my research data work. The FAIR principles and research data management work well together, FAIR data and data stewards work well together, so I see a bright future with a strong network of services under the FAIR data umbrella.'

////text box////

The following is an excerpt from the conclusion of the paper: *Are the FAIR Data Principles fair?*

Implementing the Fair Data Principles
Our analysis reflects the difficulties in interpreting the FAIR guidelines, and also putting them into practice. For many facets, less than half the sampled repositories were compliant. The Interoperable and Re-usable facets were, in particular, the most difficult to adhere to. But for many of the repositories sampled, implementing basic policies can help achieve compliance.

If a repository implements policy and practice in the four following areas...
• creating a lasting policy for deploying PIDs
• insisting on a minimum set of metadata, ideally coupled with the preferred used of semantic terms
• having a clear licence
• using HTTPS
...then are well on the road to achieve working in accordance with the FAIR principles. The principles also demand that repositories are transparent about the implication of such policies.

(…)

Our analysis leads us to the three following conclusions:
1) The FAIR principles are not just about compliance. Some of their facets need to be seen as being open-ended guidelines that can be interpreted in different ways; and varying interpretations can all be within the spirit of the original guidelines.
2) Implementing some basic policies (and publishing details of these policies) on identifiers, metadata, licensing and protocol will help all repositories align with the FAIR principles.
3) And finally closer alliances between data archives and researchers building subject-based repositories should be sought.

Archives can bring the policy and long-term expertise, whereas researchers understand tools and their domains. Satisfying the FAIR principles requires both sets of skills to be brought together.

Are the FAIR Data Principles fair? https://doi.org/10.2218/ijdc.v12i2.567

Links:
Survey dataset: https://data.4tu.nl/repository/uuid:5146dd06-98e4-426c-9ae5-dc8fa65c549f
https://www.dtls.nl/fair-data/fair-data/

The Royal Netherlands Meteorological Institute (KNMI)
The climate scientists at the KNMI make climate risks visible, both for the Netherlands and globally. As the national knowledge and data centre for weather, climate and seismology, KNMI researchers show how the climate might develop in the coming decades and they map out the consequences. The KNMI provides reliable and consistent measurements, data and forecasts that form the basis of important decisions to keep the Netherlands safe – from a code red for road traffic to the climate scenarios for the Delta programme which involves billions of euros. The KNMI works for a safe Netherlands that is prepared for the effects of weather, climate and earthquakes.

http://www.knmi.nl

KNMI: 'The climate is an international matter'

Andreas Sterl
senior scientist, climate modeller
Royal Netherlands Meteorological Institute (KNMI)

Wim Som de Cerff,
scientific advisor
Royal Netherlands Meteorological Institute (KNMI)

'Climate models, such as those that are generated and stored at the KNMI, produce vast quantities of data, hundreds of terabytes or even petabytes. This data is used to estimate, for example, the sea level increase, changes to wind speeds or the number of heatwaves. The large amounts of data involved mean that it is impossible to do the analyses on these models by yourself. There are a number of groups around the world that work on this data analysis collectively. That means that petabytes of data need to be shared with people who are sometimes on the other side of the world. To make that possible, we are working together on international projects such as the C3S-MAGIC project.'

Our data has to be findable
'When we set up new projects such as the C3S-MAGIC project, FAIR is not explicitly mentioned as a starting point. However, the FAIR principles are used implicitly. We need standards, our data must be findable, it must be in a certain format, there has to be good metadata, et cetera. For a long time we have been sharing data, using fixed data formats and standards and harmonising certain parameters and calendars. This understanding has not been introduced with FAIR, but our practice is very much in keeping with FAIR.

At the KNMI, we want to be FAIR compliant because it makes it possible to reuse our data. We are a government organisation, we believe that the more available our data is, the more people use our data, the better it is for the Royal Netherlands Meteorological Institute, for science and for the Dutch economy. We learn much more when we share and there is more to sharing than saying "here's my dataset". It means working together on improvements and increasing knowledge, nationally and internationally, within and outside our field. FAIR will help with this.'

Sharing and using standard is integral to our work
'The international character of our research means that sharing data is ingrained in our organisation. We work in the field of meteorology, and neither weather nor climate respect

borders. Climate is international. So there has long been a very clear need to work in a standardised way. Around 2012, certain general global standards were established that ensure that when we receive a model, we can understand and use it immediately. We have an exchange platform of data centres within our community, which is called the Earth System Grid Federation (ESGF). It works as a central point for this type of data.

The FAIR principles do not require much adjustment to our methodology. Our work methods were once based on telex, but of course that is no longer the case. Our history of international exchange and cooperation and the infrastructure and trust it has built mean it is now much easier to adapt the work methods of the whole community as new technological possibilities become available. For us, adaptation and renewal is a continuous process.'

Changes due to interest from other domains
'The greatest change we've made was the transition from GRIB, a standard for exchanging weather model data that is not used outside weather forecasts, to netCDF, because we wanted to exchange more with other communities. Our weather models get a lot of attention from other domains. We saw that our internal community standard was not a good fit with what was common in other communities, so we switched to netCDF. That was the first step; now we're also developing APIs that connect with this. When we look at the various FAIR principles, we satisfy the principles under the F and the A quite well. However, it is not entirely clear to us how to interpret and apply some of the I and R principles. For example, we have a standardised API you can consult, but how far do you take the idea that everything should be able to be done by machines? We often need people to put things into motion; then a lot can be done by machines. In some cases, provenance can be somewhat difficult to trace. They're better at that in seismology so we currently use some of their libraries for this.

Machines can also only go so far to check the quality of the data. Whether the correct terms have been used: yes, but whether the figures are actually correct: no. You still need real people for this. Further, if you suppose that the methodology must follow community standards, how do you handle collaborations with other domains? This still needs to be looked at carefully. We will have solutions for all principles but it still needs to be clarified which solutions or standards are needed to call your data or methodology FAIR. However, the overarching idea definitely fits with how we already work.'

C3S-MAGIC tools help to find the correct methodology
'There are many things that influence how you approach the work in your research. For example: what model will I use? What is my research focus? What is my area of application? The context of the data is then very important. I need knowledge of the various models, and how robust they are in my field. A machine cannot extract that from the metadata reliably. What we are doing now in C3S-MAGIC is facilitating and supporting the interoperability of the datasets. For instance, we are developing tools to make quality checking of datasets as automated and easy as possible. The final interpretation, of course, remains something for researchers themselves.'

How much effort are you making?
'Although we must always consider the amount of effort we put in relative to the added value it creates, we take a long-term view. For years, we have put a lot of effort into the **F**indability and **A**vailability of our climate data. If we had not done this as a global community, we would never have been able to start a project like MAGIC, which is focused on increasing **I**nteroperability and facilitating **R**euse. Previously, no one would even have thought of working on this. That is also something for other domains to be aware of: it takes a lot of time and effort to achieve a high standard of interoperability. For example, we went to a great deal of trouble to harmonise ontologies. Once the community has achieved this, it creates all kinds of possibilities for collaboration and expanding knowledge.'

The future: Long-term archiving and further harmonisation
'There are examples of past climate data projects where people were hesitant to make data freely available. But the first few projects where this was done suddenly had many, many more citations and reached a much larger audience. Everyone in the world could see it. That hugely increased the impact. The more open the better in terms of reuse. As few forms as possible, as easy as possible to analyse; that is the future.

We are currently working on drawing up KNMI policy for long-term data storage. How long do we want to save certain models? What will we do with old software that is used for analysis when new versions become available? There is of course a cost aspect to these questions. We are talking about a very large amount of data. It costs a lot either to save it or to dispose it off properly. A lot of money was also spent on creating the datasets. So what do you do in the longer term? This discussion is already becoming more relevant.

Beyond that, there needs to be some coordination to monitor and approve standards and encourage people to agree to use them. More clarity in harmonising guidelines is also desirable. For example, the KNMI must satisfy the open data guidelines for government data. To what extent do they correspond to the FAIR principles? That is very important to us. SURF could play a facilitating role in this.'

////text box////

The C3S-MAGIC project: Developing software for data from climate models
The idea behind the C3S-MAGIC project is that someone who is interested in climate data no longer has to copy all the data to local storage to be able to perform calculations. With the software being developed within C3S-MAGIC, analyses can be performed on these datasets remotely without the data needing to be copied or moved. The data stays in the place where it was generated and saved. The researcher can indicate what calculations should be done on what data via a web interface. The website climate4impact.eu largely developed by the KNMI, will form the basis of the C3S-MAGIC software. Climate4impact.eu already makes it possible to combine information from different climate models with each other and to visualise this information, but further processing is not yet possible. This will be solved by the new software from C3S-MAGIC.

https://www.knmi.nl/over-het-knmi/nieuws/ontwikkeling-software-voor-data-uit-klimaatmodellen

Links:
Earth System Grid Federation (ESGF)
https://data.overheid.nl/over-open-data-0
https://www.unidata.ucar.edu/software/netcdf/
https://github.com/c3s-magic/

ODISSEI: A new national data infrastructure for the Social and Economic Sciences
The Open Data Infrastructure for Social Science and Economic Innovations (ODISSEI) is a collaborative national initiative that started in 2016. Its goal is to create a national data infrastructure for the social and economic sciences in the Netherlands. ODISSEI will bring together a vast number of datasets in a single data infrastructure and provide scientists with the tools necessary to analyse it. It focuses on the data collection, sharing, processing and archiving phases of the research life cycle. One of the main challenges facing this new data infrastructure will be interoperability. From the outset, the FAIR principles have been explicitly incorporated in the description of ODISSEI and its goals.

http://www.odissei-data.nl/

ODISSEI: 'We are on the brink of something revolutionary in our field'

Tom Emery
Executive Director
ODISSEI

A longstanding tradition of reuse
'The social sciences have a longstanding tradition of reuse. Large scale surveys are not designed for just one purpose. Different researchers have always used data from Statistics Netherlands (CBS) in different ways and we are used to reusing these datasets. Overarching principles along the lines of the FAIR principles were already acknowledged in our community before FAIR came along. However, the way we are working has changed a lot in recent years. Queries that used to take months can now be processed in two days. This opens up new possibilities. To take advantage of these new possibilities we need to change our way of working and work together with non-social scientists. In ODISSEI we work together with SURFsara and the eScience Centre. They are not used to working with the standards we use in the social sciences, so we have to find and define common ground. FAIR is very helpful to ensure we all adhere to the same principles.'

Digging deeper into the FAIR Principles
'The increasing scale of social science research queries make it very important that the FAIR principles are applied. Even though it is obvious that they were written with data from other domains in mind, the FAIR principles are useful and we must try to make our outputs as machine readable as possible. The increasing acceptance of FAIR amongst funders means that there will be political pressure to adopt the FAIR principles. Superficially, it seems that current and recent practice in social science complies with this desire to make data FAIR. This means that social scientists may not feel the need to change the way we work as strongly as those in some other disciplines. This could hinder progress. Because if we dig a bit deeper and assess the detail in the 15 different FAIR principles then it is clear that, as a community, there is still much room for improvement.'

Drive practices towards more interoperability
'Part of the mission of ODISSEI is to encourage more interoperability in the social science data collection field. This requires more standardisation, which includes new standard forms which structure data with potential reuse in mind. We do use survey standards, like the international DDI standard, but in practice this standard is not FAIR enough. DDI should, in theory, make all the

data objects Findable because you can query the surveys. But if you want to do a cross query you must use a keyword search to do so and most data are machine readable in a crude way, but not interlinked.

The data has not been structured with discovery in mind, so it is not findable in the way that it should be. Datasets are citable and have a DOI, but the specific data within the datasets are not. While we are accustomed to reusing large-scale surveys, other data is traditionally not set up with reuse or citability in mind. This restricts current and future integration. These are some of the challenges we are taking up in the ODISSEI initiative, helping move the community towards a FAIR future.'

Going FAIR takes a bit of courage
'The ODISSEI infrastructure will help take these steps towards FAIR as part of a larger culture change moving towards openness and interdisciplinary cooperation. For example, it is common practice in some disciplines to charge to allow others to use your dataset, this is not permitted in ODISSEI. For us, that is part of Accessibility.

With the high volumes of personal data we process, 'Going FAIR' will take a bit of courage from the different players in our community. FAIR data does not have to be open data, but FAIR data is secure data. Making your data FAIR should heavily influence the way you interact with other scientists without necessarily opening it up to everyone. FAIR practice will make it evident to our community that we need to reassess our ways of working and that should be considered one of many good outcomes.'

The future of FAIR
'We are on the brink of something revolutionary in our field. Five years from now, if things progress well, our surveys will be almost unrecognisable. Existing standards will be adjusted. This will not always be easy. Much of the effort you put into making your data FAIR is not necessarily for your own benefit; it is more for the benefit of the community as a whole. It is not always easy to explain why this must be commonly seen as just good practice in science. A lot of the younger researchers see the advantages. For managers, to locate large amounts of money and resources to changing the way we work is still challenging. The return on investment may not be obvious.

At the moment, it seems that you have to have faith in the overall move towards openness and cooperation. If we can make sure that in the near future there are more tools available that you can use if you adjust your way of working, then the benefits will be more apparent. Having a new infrastructure like ODISSEI in place is a big step forward. The FAIR terminology helps us to talk consistently about all the different aspects of this change. It also ensures that scientists from other fields understand that we might work differently, but we do use the same processes. FAIR both enables a conversation and pushes us forward.'

Ruurd Schoonhoven
Senior Account Manager for Microdata Services
Statistics Netherlands (CBS)

Statistics Netherlands in ODISSEI
'The idea behind ODISSEI is to bring a large number of datasets in the social sciences together. Statistics Netherlands (CBS) data plays a special part in this concept. The information published by CBS is a very rich source for our community, especially their 'register data'. In recent years, CBS has increasingly been collecting register data: the municipal population register database,

the trade register, taxes, wages, benefits, education, criminal records, health, et cetera. This data is received automatically. The other datasets in ODISSEI are random samples but this CBS register data is integral: it includes data on everyone in the Netherlands. This means that if you link a certain random sample to our register data, you will always find your random sample respondents. This enables you to enrich your dataset with supplementary data to give you a much clearer picture of the group you are researching. Analysing your own data linked to CBS register data is possible only under very strict rules for statistical and scientific research. CBS must comply with the law that governs it and a crucial part of this is that no data that can be traced back to individual people or companies may leave CBS. We are extremely meticulous about this.

Need to find computing power
'Since the CBS microdata facility was established, more and more researchers come to us with their random sample data. This is then linked to the CBS data internally. In ODISSEI, the possibilities are now being expanded greatly. Among other things, because our data files and those from researchers themselves keep getting bigger, we needed to go to SURFsara for its computing power and storage capacity. Taking into account all the security and privacy rules and the Statistics Netherlands act, an environment has now been created within SURFsara that can legally be regarded as a part of Statistics Netherlands (CBS). There, we can do the calculations that an increasing number of researchers in the social and economic sciences request, and link different datasets. That's how you get 1+1=3. This is a major step forward for the research in this domain.'

Progress and expanding knowledge
'ODISSEI is still in an early phase of the partnership, and therefore also of working according to the FAIR principles. CBS is working first internally, and next externally, on updating the search functionality in our metadata, which will also make external use easier. However, it will take a while before this will be operational. Separate from the infrastructural progress in terms of technology and security, ODISSEI also offers an infrastructure for progress in collaboration, exchange and increasing knowledge. More and more datasets are available in linkable form through CBS, although there are also parties that have made a very large investment in their data and are therefore hesitant to share it. The idea is that, through ODISSEI, new grants can be awarded for data collection with the condition that the data is made available via ODISSEI in the most accessible form possible. That's the direction it will go in the coming years.'

The future is FAIR, but it will take time
'FAIR is a set of principles. It is important to harmonise data sets and make them available, but for CBS the guidelines for making official statistics take precedence when collecting data. With regard to access to this data for research, I would say: FAIR please, but it cannot be made compulsory and it won't be free. If you just say that everyone must work according to FAIR principles without any further explanation, you don't understand the complexities. For instance, when it comes to personal data, privacy protection comes first. That is why at CBS, we want to make data as accessible as possible, but at the same time we do not make any concessions on its confidentiality. It should also not be forgotten that harmonising metadata, even internally at CBS, takes a great deal of time and effort. If you want to do that throughout all the social sciences, this will require even more time and budget.

The I in ODISSEI stands for innovation
'What still needs to be done varies widely by dataset, but in five years there will be much more cohesion between the data collections. We will need to do less duplicate work and focus more on technical coordination. However, innovation is already leading to forms of data and data collection that didn't even exist in the past: consider the use of internet robots, biometric sensors or other

'wearables', and the many other types of Big Data. The interpretation of the FAIR principles for these kinds of data still requires further development. The intention is of course that ODISSEI will improve the quantity and quality of FAIR exchange and collaboration. To do this, we need to get many research groups and institutions on board and convince them of the utility and added value of this. But we are increasingly moving toward other forms of collecting data and we are already working together more. That is clear. Everyone for themselves just doesn't work anymore.'

////text box////

Some of the surveys and panels that are being included in ODISSEI:
- Netherlands Twin Register (NTR)
- Longitudinal Internet Studies for the Social sciences (LISS)
- European Social Survey (ESS)
- Dutch Parliamentary Election Studies
- International Social Survey Programme (ISSP)
- Generations & Gender Programme (GGP)
- European Values Study (EVS)
- Longitudinal Aging Study Amsterdam (LASA)
- Children of Immigrants Longitudinal Survey in the Netherlands (CILSNL)
- Dynamics of Youth (YOUth)
- Netherlands' Life Course Survey
- Socio-Cultural Developments (SOCON)
- Family Survey Dutch Population
- CONflicts And Management Of RElationship (CONAMORE)

Links:
Statistics Netherlands (CBS) general info https://www.cbs.nl/en-gb
Statistics Netherlands (CBS) for research data: www.cbs.nl/en-gb/microdata
Full list of ODISSEI surveys & panels http://www.odissei-data.nl/en/content/surveys+panels

The National Institute for Subatomic Physics: Nikhef
Nikhef is the Dutch National Institute for Subatomic Physics. The institute does research on the elementary building blocks of our universe, the forces between them, and the structure of space and time. On one hand, the research looks at elementary particles that collide with each other in controlled experiments with high energy and intensity, and on the other hand on observing particles that come to the earth from the universe. The scope of the data that is produced in the different experiments that Nikhef works and collaborates on varies enormously, from several tens of terabytes per experiment to more than a hundred terabytes per second in the experiments with the Large Hadron Collider (LHC), the underground particle accelerator near Geneva with a circumference of 27 km. Nikhef itself has about five petabytes of disk archive, at SURFsara there is about that again in disk archive and almost 13 petabytes of LHC data in tape archive so far. That volume will only increase.

https://www.nikhef.nl/en/

Nikhef: 'Look at the content, not the label'

Dr. David Groep
Researcher 'Physics Data Processing' group
Nikhef

Our policy does not correspond with others' policy
'Recently Nikhef adopted a new research data management (RDM) policy. But already before that time we contributed to drawing up the general data policy framework for the Institutes organisation of Netherlands Organisation for Scientific Research of which we are a member. In this process we came up against the domain-specific aspects of Nikhef's and other institutes' work. For example, in early discussions about this general data policy it was mentioned that the hardware and tooling that are necessary to understand the experiments would need to be saved in order that others will be able to reproduce the experiments. This often simply does not work in our domain. What kind of hardware and tooling do you want to save? The LHC itself? The filter hardware to get from terabytes per second to tens of petabytes per year? Or only the 'firmware' source code for this? There are other domain-specific examples we could give to show that our RDM policy cannot be exactly the same as others' policies.'

Our own policy
'Our own new policy does not mention FAIR by name. However, we did look closely at the FAIR principles when we drew it up. The new policy gives us a structure to improve our internal processes. In our many international projects, the data policy has often been well-established; in our smaller internal projects, less so. Internal consultation has shown that the heterogeneity in terms of handling data of locally generated is greater than expected. Not all PhD students were educated in the Netherlands, for example. How they handle logbooks and notations and the like is by no means always uniform. We plan to implement this policy in a few new, smaller projects in the range of ten terabytes of data. From there, we will be able to expand this policy in phases to larger projects involving datasets running into the petabytes.'

FAIR is good for awareness
'Without referring directly to FAIR, making data findable and accessible, interoperable and reusable can identified clearly in our policy. FAIR can also be used very effectively to raise awareness of good data management. It works to use FAIR, especially the four main criteria, in training programs in order to explain what it means to do 'good science'. And also why we want certain tools to be used in a certain way; for example because the correct metadata is then generated automatically at specific moments during your research.'

FAIR: actually not as general as it seems
'When you look at the 15 FAIR principles, they contain an assumption about the research methods used. This seems to be primarily inspired by the life sciences. When you read the principles literally, it is as if you could make the scientist redundant in the area of data analysis. After all, the data is written so machines can analyse all the data. That is impossible for the datasets in our field. For example, I have a fundamental problem with how principle I.1 is formulated:

(meta)<u>data use a formal</u>, accessible, shared, and <u>broadly applicable language for knowledge representation</u>.

You would then expect a complete ontology, a semantic description in OWL – an ontology language for the Semantic Web – or something of that sort. The ambition behind this is that you could then generate new knowledge with machine learning technologies. I think this principle could be applied well if you work with datasets that are common in the life sciences, for example gene sets. But there is no way that will work with our complex datasets – the risk of false positive results alone would be enormously high. You need people with years of postdoc experience – and often specific knowledge within a collaboration - to be able to interpret them. A 'blind' analysis – first performing the research methods on simulated data and only then looking at whether the measured data deviates from the simulation models used – is the norm to be able to make claims. That cannot simply be taken over by machines.'

70 years of international research tradition
'We have a 70-year long tradition of international research with inherent data exchange. We do not use ontologies but have other ways of standardising our methodology. For example, by using a specific type of software globally, or code that makes reuse and replication possible.

Within our international collaborations, data management is well organised from the beginning, especially for the retention and accessibility of the measurement data. We promote good science, focussing on innovation; our goal is to gather new knowledge about matter at the smallest and the largest scales. The fact that you need to save and process your data properly in order to do this is obvious. It is simply necessary for our own analyses. People started years ago on things like recording the metadata structures properly. Models are very thoroughly tested, for example with large quantities of generated 'simulation' data. For each large project, this is specifically adapted and tested again. So this is all robust now, although it did take time to develop the methodology.

We also have a chain of responsibility in our domain that includes international accreditation. For example, the DPHEP working group (Data Preservation in High Energy Physics) looks at the persistence of our data, which means keeping it accessible and reusable in the longer term. That can involve the specific software that is used for an experiment and therefore needs to remain available. That was already there before FAIR.'

FAIR as a normative standard is dangerous

'As long as the FAIR principles are seen as principles, with a clear reference to domain-specific situations or implementation, they can be applied quite well. The overarching idea of findability and reuse is very useful. But given that FAIR, as it is now defined including principles such as the 'formal language', does not really fit with our way of collaborating, there is not much point in trying to follow this exactly. However, satisfying the four overarching criteria of 'FAIR' is trivial for our datasets.

So if compulsory national or European policy is drawn up, instead of 'your data must satisfy FAIR principles' it should say 'FAIR principles or a domain-specific equivalent'. We think this is very important. The danger lies in making the principles as they are formulated now into policy setting standards. FAIR doesn't seem to be meant for this, but in spite of this you do see this happening among policy makers. For instance, a draft version of the Dutch 'code of conduct for scientific integrity' said that you had to work on making relevant data compliant with the FAIR criteria: findable, accessible, interoperable and reusable, without further explanation about the definition of the specific elements or how the criteria are applied. If a specific reference to this standard is then made with regard to integrity weighting, and possible consequences if you do not satisfy this requirement, that goes too far. We do not have a 'formal, shared, broadly applicable language' on hand for our data. How could we then satisfy the FAIR principles?'

Specialising by research domain

'The principles would be much more palatable if you could specialise them according to your research domain. You would have to put a disclaimer at the top like 'The following principles must be interpreted within the context of the specific discipline on the basis of international consensus'. For example, in high-energy physics everyone around the world, in all experiments, has been using the same technically interoperable data format for the last 15 years: ROOT files. Further, there is even an emerging standard for making, storing and sharing the statistical models that lead from data to results ("ROOFIT" models). This means that, for example, the results of the ATLAS and CMS experiments at the particle accelerator in Geneva can be correlated and combined and data is in the same format at Fermilab in the US. Results from many different experiments (for over 60 years now!) by the Particle Data Group can routinely be combined into the 'best current result' again each year. Things like this are more important for our domain than a 'formal, broadly applicable language'. I can imagine that this also applies to other domains: for example in the humanities you certainly cannot satisfy all the principles as they are now formulated either.'

Cooperation with other domains

'If interaction with other disciplines is possible in a certain area, it often pays off to make the effort to develop a common vocabulary, even where, as we have said, we do not use formal ontologies as standard in our own discipline. You see that now in gravitational wave research, where we are working with many other disciplines. About ten years of work was invested to make sure that, for example, all kinds of telescopes are notified and able to register the same thing at the same time when a special event occurs. The added value of this is clear to see, which is why there was also major investment in it.'

The future of FAIR: Investing, or sometimes rather not?

'Investments in working according to the FAIR principles or extensive data management are always a balance. The added value is not always clearly visible. A lot is automated, but if you don't know whether a dataset is worth reusing beforehand, how much more effort do you want to put in? Graduate students are under pressure to finish their research as quickly as possible. You

can't put a data manager next to every researcher. This still needs to be considered very carefully, not only in our field. What do you do with large datasets when you can be almost certain they are not interesting enough ever to use again? Do you pay a lot for persistent storage in an international repository, or do you leave it in cheap local storage, despite this being less findable?

I don't know if this is already available, but for this category of data you actually want a place where you can save your data without doing all kinds of processes, but where it does get a persistent identifier, for example. You can then postpone the decision about whether the amount of effort needed to make everything nicely reusable weighs up against the chance that this set will be relevant. You just indicate: yes, we have this set. If there is a reason to look at this set again, then we can put in the effort to make it available and reusable. A sort of DOI minter.'

It works well to automate
'FAIR will remain relevant as a guideline for all domains in the coming years. Making sure that people can work according to the FAIR principles is expensive. It works well to automate: for example to develop tools that fit into a particular type of research and which automatically generate all kinds of metadata. Resources need to be available for this work and also for storing datasets in the long term. These kinds of questions still need to be considered carefully. If the attention that FAIR is now generating means that we can invest in easier access to data, better descriptions of the research process and training in making data reproducible for people who are able to analyse the data, that is a major step forward. But as we have said, we absolutely do not see the added value of assessing datasets for their interoperability 'using a formal language for knowledge representation' and attaching consequences to this.'

Links:

A gallery of interesting Jupyter Notebooks: https://github.com/jupyter/jupyter/wiki/A-gallery-of-interesting-Jupyter-Notebooks#reproducible-academic-publications
An emerging standard for making, storing and sharing the statistical models that lead from data to results that Nikhef is working on is ROOFIT:
https://www.nikhef.nl/~verkerke/talks/chep03/chep2003_v4.pdf and
https://cds.nyu.edu/projects/collaborative-statistical-modeling/
SWAN (Service for Web based ANalysis) https://swan.web.cern.ch/ is a platform to perform interactive data analysis in the cloud, which has features like:
- Analyse data without the need to install any software
- A Jupyter notebook interface as well as shell access from the browser
- Access experiments' and user data in the CERN cloud
- Document and preserve science **-** create catalogues of analyses: encourage reproducible studies and learning by example

The Leiden Centre for Digital Scholarship (CDS)
The CDS works with researchers, faculties, national and international colleagues and expertise centres to facilitate and support Digital Scholarship. The CDS organises meetings and workshops and is *the* place where researchers can go for information, questions, advice and training about data management, open access, publication advice, copyright, virtual research environments and advanced datamanagement (e.g. datamodelling, text and data mining, annotating and enrichment of digital objects).

https://www.library.universiteitleiden.nl/research-and-publishing/centre-for-digital-scholarship

CDS: 'An integrated approach with the right support is essential'

Laurents Sesink
Head of the Centre for Digital Scholarship (CDS)
Leiden University Libraries

'Leiden University adopted Data Management Regulations in 2016. They don't mention FAIR by name, but the regulations do specifically state that the data has to be findable, accessible, comprehensible, reusable and archived for the long term ... which matches the intention of FAIR. In addition, the new general data protection regulation (GDPR) has an influence on data management practices. We need to translate the regulations according to the practical situations of every discipline. A data management implementation programme has been set up to support the faculties with this. Our aim is to work toward practical protocols tailored for each discipline or research institute. These protocols do not have to be binding, but if you deviate from them, you do have to explain why.'

Support and awareness are essential
'If the FAIR principles are just imposed, and there is resistance to FAIR because it is only seen as extra work, there is the chance that researchers will be alienated. So it is very important for us to provide good support. Not every researcher works with huge data sets and at some faculties people are still completely unaware that FAIR also applies to their type of data, such as interview data. It is essential to raise awareness amongst researchers that making their research **F**indable and **A**ccessible offers them only advantages. If this is done well then it will be much easier to structure research data management properly in that domain.

Currently, advice tends to be focused on saving the data in a FAIR-compliant way after a data project has ended, so that it can be reused by others. It is better to shift the focus to working according to FAIR from the initial setup of a research project. This will make it possible to do better analyses. Working with digital data has its benefits; let us explain what you can get out of it during your research if you do this in a structured way.'

As simple as possible
'It is also important to make it as easy as possible for researchers to comply with any extra expectations or requirements. We envisage establishing protocols and methodology for each discipline on the basis of the Data Management Regulations. This will also need facilities, instructions and training to make adoption and use as easy as possible. Data stewards will play an important role in this, researchers don't have to be able to do everything themselves.

We want our researchers to have the best support possible so we very much welcome that FAIR is now considered an integral element of data management. This does require good coordination and harmonisation. But all parts of the research life cycle that involve data are so interwoven that there is no longer any other way.'

Ben Companjen
Digital Scholarship Librarian
Centre for Digital Scholarship (CDS)
Leiden University Libraries

'In addition to our other work in Digital Scholarship, such as support with data management and Open Access, the CDS helps individual researchers with advanced datamanagement, for example processing and modelling data and text and data mining technologies. In general, this involves researchers who want to know more about the potential of new digital methods and technologies. For advanced applications and research that involves working with large quantities of data, researchers can go to the Leiden Centre for Data Science; for information about the cricitical use of digital technology and computational approaches in disciplines of the humanities, researchers can get support from the Leiden University Centre for Digital Humanities (LUCDH). Sometimes we explain what data analysis tools can offer to researchers who have never worked with them. The questions we get are about every stage of the research: from planning to analysis to publication. They often involve small datasets. These may not require machines to generate and analyse the data. So the researcher may not perceive a very clear reason to apply the work method from step one in the research process. It is important that we explain and raise awareness of the advantages.'

Custom work
'We provide information about the use of digital data, answer questions and give advice. Sometimes we also provide hands-on support, which is often customised. We do not have the capacity to do this with a large number of researchers. For instance, we are working with a few researchers to transfer the data from their work environment (virtual research environment (VRE)) to our repository. In this way, the data also remains findable, interoperable and accessible after the end of the research. We have done this a number of times, so we recognise a certain workflow. But because these VREs are all different, we cannot set a general script for this. So providing this advice is time consuming.'

FAIR as part of good Research Data Management
'FAIR is an important starting point for us, as part of good Research Data Management. A number of people in the team have taken training in FAIR. We are keeping track of developments in the field of FAIR data. The principles of FAIR are being used in our Research Data Management training. Hopefully, with training and also with ever-better tools, it will become standard practice for our researchers to handle their data according to FAIR. To achieve this, it does need to become easier. If you do not have an IT background, it is difficult to think like a computer. When we explain what the letters FAIR stand for, it usually sounds very logical to people. But the underlying principles are often meaningless without being applied to their work.

FAIR is logical and is getting more important. But the question we're more likely to get here is: 'what should I do to describe everything properly in my data management plan?' The questions that need to be answered in a data management plan are often still very general in nature, and there is not yet enough guidance or examples showing how this should be done. Researchers

themselves should not need in depth knowledge of FAIR, but this knowledge underpins our support and advice which can sometimes just be help with how you fill in a spreadsheet.'

Early stage
'A researcher may come to us at an early stage of a project and we can immediately show the extra potential of creating and using a good spreadsheet, for example. Once people see everything that can be done with their data, they are quickly motivated to work just a little bit differently. I may demonstrate running a script on it and converting it to internationally used data formats such as RDF, JSON or HTML. Sometimes researchers are not aware that parts of the research can be easy to process with a computer. This can save a lot of time. If a group comes here at a later stage, for example with a question about filing a database in a repository, this time saving is no longer possible.

People still often fail to consider the importance of recording where certain information comes from. An example could be a dataset of plant names attached to old drawings, where the English names have been added from another dataset. That makes it important to record the set that data comes from, especially when you will be making this new dataset available. Hopefully, people will increasingly understand what a computer needs to be able to work with the data. Once that happens, new research questions can also emerge. We try to make this potential clear in our training sessions.'

Increasing interoperability
 'Different domains and institutions are currently working on finding ways to implement the principles in practice. The interoperability of datasets is increasingly taken into account. A standard such as IIIF for image material is being used by a growing number of libraries and cultural institutions. IIIF is also being expanded to other audio-visual and even 3D material. The fact that such a standard is being used on an increasingly large scale hugely strengthens interoperability and reuse. But, for example, the setup of a FAIR data point (a repository for FAIR data) for GO FAIR Netherlands is not compatible with the *W3C Recommendations* for the Linked Data Platform. So it would be good if more examples became available.'

Manuals per domain
'It is very helpful to hear about other people's practical experiences. A lot of this kind of information exists for biology and life sciences, for example on fairsharing.org. It would be very helpful to have manuals to support researchers in other fields. Examples are important: when you work with this kind of data and you use this data model with these tools, you can do those calculations, or visualise your data in a new way. Easy access to standard data models is also a great help. We also try to contribute to this with our training programs, for example by drawing up a practical description of how you can get from a spreadsheet to Linked Data. There is still plenty to do.'

////text box////

University-wide programme
Leiden University recently performed an inventory of all the research support offered. It was found that when you are drawing up a research data management plan, you may need advice from six different experts or support staff. Issues include: privacy legislation, ethical guidelines, funder requirements and also the FAIR principles. The support and expertise are distributed over different services. We now plan to set up Research Support so that there is one virtual place where researchers can go for information, questions, help and referrals. Leiden University also wants to look in more detail at how researchers can be supported better. This involves the

expansion of IT support for research, increased support for advanced data science, and more general support for open science, which also covers the "FAIRification" of data. To be able to offer optimal support to Leiden researchers, this support is being designed in more detail as part of a university-wide programme. This is based on satisfying far-reaching requirements and regulations and supporting advances in digitisation of all kinds of research.

///text box/////

Dr Krista Murchison
Lecturer in Medieval English Literature
University of Leiden

'I recently won a grant to catalogue and analyse a group of manuscripts from medieval England. These can still be read and those ideas from 500 to 1000 years ago can still be studied by us today. But in my digital humanities training, I learned that most digital formats are not stable at all. I approached CDS to ask if they had facilities to deposit my data in a sustainable way and if they could advise me on best practices.

In the initial stages, Ben advised me on how to develop a strategy for structuring the project data in a way that would follow the FAIR principles. He also gave me crucial information about the university's web hosting options and pointed me toward a resource about formatting my data that has proven very useful. Aside from this, he developed a script that will eventually help me analyse the data I've collected. It was helpful that the CDS set aside time to understand my project, its goals, and my own background with digital scholarship, and to develop a strategy that was tailored to the needs of my project. This is a sign that the university recognizes the importance of responsible data management strategies and is willing to support researchers in this.

The underlying idea behind the FAIR principles was really what led me to seek out the CDS in the first place. I have had some training in text encoding, so I know about the importance of making data interoperable and reusable. But I learned the FAIR acronym from the CDS, and by working with them I developed a stronger understanding of how the four principles fit together to support good data management practices. I am convinced that as well as helping current researchers, ensuring that data follows the FAIR principles will prove beneficial to researchers in the future.'

Links:
Leiden University https://www.universiteitleiden.nl/en
Leiden University Data Management Regulations
https://www.universiteitleiden.nl/binaries/content/assets/algemeen/onderzoek/research-data-management-regulations-leiden-university
The General Data Protection Regulation (GDPR) https://ec.europa.eu/info/law/law-topic/data-protection_en

This use case was delivered via the Dutch GO FAIR office as an example of their work in the biomedical and healthcare field. It is an abridged version of an interim report of a practical test conducted by the Dutch National Health Care Institute in cooperation with GO FAIR. This use case therefore has a somewhat different structure.

National Health Care Institute
The National Health Care Institute strives to ensure that every citizen in this country can be sure of good health care, no more and no less than necessary. Information and information management play an important role in the Health Care Institute's legal tasks and the sustainability of the healthcare system. Information and healthcare can no longer be seen as separate from each other. Getting good information is essential in the process of being, becoming and staying healthy. The Health Care Institute facilitates the use of information technology for an effective and efficient healthcare system. As part of this, the Health Care Institute explicitly looks for new concepts and technologies that can play a role in the information supply. The work method for this focuses on practical tests in which a new concept or new technology is implemented in practice. To do this, we work with other organisations to look at innovation from different perspectives (technical, legal, privacy, organisational). That enables the Health Care Institute to learn what innovation really means for healthcare and what you need to take into account for the implementation. For instance, the FAIR Data and Personal Health Train project was started in the past year.

https://english.zorginstituutnederland.nl/

A practical test for FAIR data

Wouter Franke MSc MCM
Information management advisor
National Health Care Institute

In October 2017, the Health Care Institute started a practical test of FAIR data & the GO FAIR Personal Health Train. In the first phase, the focus was on learning to apply the FAIR data principles and exploring the application and implementation of these principles in healthcare. In the second phase, we will test a scenario in which we use the Personal Health Train. The goal of the second phase is to explore what is needed to implement the Personal Health Train in healthcare.

Results of first phase of practical test
In the first phase, practical work was done to make a dataset from the Health Care Institute FAIR. The dataset we used for this is the Long Term Care Act (Wlz) process information. The Health Care Institute draws on this data to publish information on the waiting lists in long term care each month. The dataset is written in the iStandaarden information model.

The first thing we did was to make an ontology of the information that is present in the file. Because the dataset from the Health Care Institute has already been written in a structured way in the information model, this step was relatively simple. In the second step, we converted the actual dataset to Linked Data. As a basis, we took the dataset in the current form (XML) and converted it to RDF format. To make the data 'linkable', a URI (universal resource identifier) is also determined in the ontology for every subject. Then the ontology and a sample XML file were used as a basis to write a piece of code (in the programming language Java) to convert the data from XML to RDF.

The next step in making the dataset FAIR was setting up a FAIR Data Point. The tooling that was available from the GoFAIR implementation team made this easy. Technically, this means setting up a web server on which a Sparql endpoint is provided. After it is published, this information is also findable using the FAIR search engine. The goal of this is that over time, this information is also indexed and displayed, for example by Google or other search engines.

Possibilities and challenges
The more data there is available in a FAIR form, the greater the possibilities to link data. FAIR and Linked Data together offer vast possibilities for 'big data' applications in health care. Sparql queries allow the connections to be scaled better.

The implementation of the FAIR principles as described here relies heavily on the use of Linked Data. To make Linked Data (and the associated technologies) accessible to a wider audience, we need more user-friendly and better-supported tooling. We will contribute to this from the Health Care Institute by engaging with other parties to share our needs in this area.

A challenge is the licences under which data is offered. How can other parties use the data and what control do you have over the data as the owner? Licences are important to clarify the possibilities and limitations of reusing the data.

Conclusions
Following from the first phase of the practical test, we can draw the following conclusions.
- FAIR has been broadly adopted as a principle. Various bodies (e.g. G7, G20, EU, hospitals) are working on FAIR data based on the promise of data that is findable, accessible and reusable, that systems can use directly.
- FAIR data can give a huge boost to Big Data-type applications in which data is brought together from many different places and origins.
- FAIR data allows one to connect data sources that cannot otherwise talk to each other. In this way, you can achieve interoperability without the need to all speak exactly the same language.
- FAIR Data is fairly technical to implement. It focuses on connecting data from systems and currently has limited tooling to keep the technology concealed from users.
- To become FAIR, it is a great help to have data that is already structured. Time and energy that have been put into standardising and structuring information give you a head start when you want to become FAIR.
- The learning curve to implement FAIR is quite steep. For people with a background in information provision and an affinity for technology, it certainly can be done.
- The technical process to become FAIR can be done relatively quickly. The Health Care Institute made a dataset FAIR in four weeks, and this will keep getting faster with experience. Organisational questions, such as the licence under which data are shared, will take more time. It helps to remember that FAIR data is not the same as Open data.
- Standards in long-term healthcare can be made FAIR with a little effort. This can help the field itself to become FAIR. That is why in the last steps of this phase, the Health Care Institute is working on making FAIR all the iStandaarden standards that are used in healthcare & support. Parties in the field will then have the specification of the data which will allow them to make the switch to FAIR quite quickly.
- The added value of Linked Data primarily depends on the possible links to other Linked Data sources and the possibility of linking to definitions etc. in other ontologies. The availability of Linked Data and ontologies is still limited for the dataset used. Making an ontology available for long-term care can also help with drawing up ontologies in adjoining domains, such as hospital care or the municipal domain.

Continuation

To complete the first phase, the standards which the Health Care Institute manages are now being made FAIR. These descriptions will then be published on the Fair Data Point of the Health Care Institute. We will also look internally at whether more datasets present at the Health Care Institute can also be considered for this process. We also explicitly look at the privacy aspects (e.g. using Privacy Impact Assessments) and licencing possibilities within FAIR.

Outside of the Health Care Institute, the inventory will be broadened. The implications for other organisations and programmes will be examined. This will involve coordination with the programme Registratie aan de Bron [Registration at the Source], MedMij, the standardisation organisation HL7 and with DICA. Outside healthcare, we are looking for collaboration with organisations such as the LIACS for technology, the government prosecutor from a legal perspective and the Dutch Data Protection Authority from the privacy perspective. In this way, we also expect to be able to describe the implications of FAIR data and the Personal Health Train for healthcare more broadly. This will happen in the second phase of the practical test.

Preparations for second phase

In addition to completing the first phase, we are now preparing the second phase, in which we will test a practical scenario of the Personal Health Train. We are in discussion with other parties in the field to achieve this. With this, we are focusing on three different scenarios (use cases), including Intra-Arterial Thrombectomy (IAT). As with the Health Care Institute's other projects, there is collaboration with the government prosecutor for legal testing, the Dutch Data Protection Authority for privacy testing and a university for technical testing at an early stage. We are working toward a working environment for the Personal Health Train by May 2018.

/////textbox////////

GO FAIR

GO FAIR is developing a bottom-up open implementation strategy for the first phase of the European Open Science Cloud. GO FAIR's user-led strategy includes an early phase of 'federating the gems'. Crucial FAIR activities will begin without delay, working with motivated early movers who have already been identified and organised into "implementation networks". New implementation networks can be added at any time and the strategy can be adapted by the participating networks. The GO FAIR consortium is open, inclusive and stakeholder driven. Its mission is to contribute to and coordinate the coherent development of "the Internet of FAIR Data & Services" through community-led initiatives.

The GO FAIR strategy has three key interactive elements:

- creating the socio-cultural change required for Open Science to flourish (GO CHANGE)
- training the data stewards needed for data stewardship plans, including FAIR data and services (GO TRAIN)
- designing and building the technical standards, best practices and infrastructure components needed to create the network of FAIR data & services (GO BUILD).

The GO FAIR website explains each of the FAIR principles, gives examples and provides links to resources. The Netherlands is one of the co-founders of GO FAIR, and at the moment there are five Dutch implementation networks actively involved. The following use case is a practical exploration of the implementation of the FAIR principles by Zorginstituut Nederland, in collaboration with the Dutch National GO FAIR office.

https://www.go-fair.org/

////textbox////

**Intra-Arterial Thrombectomy scenario**

Intra-Arterial Thrombectomy (IAT) is a relatively new treatment for a cerebrovascular accident (CVA), or stroke. This is a highly complex treatment and twelve centres in the Netherlands are designated to deliver this care. The treatment must be done within a few hours after the stroke occurs, and the time between the various steps in the patient process is crucially important. The sample process we are using for this case is given schematically below.
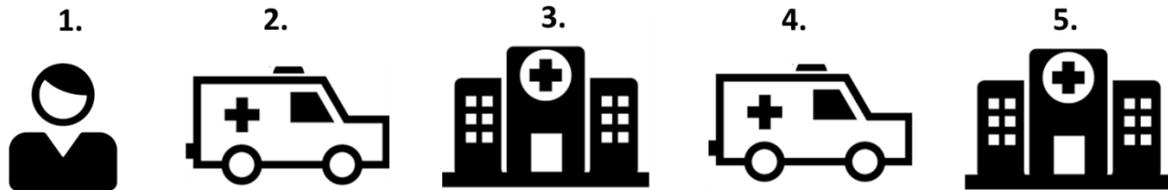
1. A patient has a stroke. He calls 112 to be taken to the hospital urgently.
2. The ambulance picks up the patient and takes him to the hospital.
3. In the hospital, the patient is diagnosed as having had a stroke, and he gets intravenous thrombolysis (IVT). The hospital then suspects that the patient is eligible for an IAT treatment and a second ambulance is arranged.
4. The second ambulance takes the patient from the hospital to a hospital that does IAT treatment.
5. The IAT treatment is done in this hospital. This hospital is currently responsible for entering all the turnaround times of the various organisations in this healthcare supply chain. This is done using the existing Mr Clean registration.

The Personal Health Train provides ways to collect this information afterward. A scientific association, a researcher or a quality organisation is interested in the turnaround times for IAT treatments. The steps the patient has gone through and how much time each step took. Where in the process can we make improvements? To be able to ask this question and answer it using a Personal Health Train, we propose the following example of a process with the elements of a Personal Health Train added.
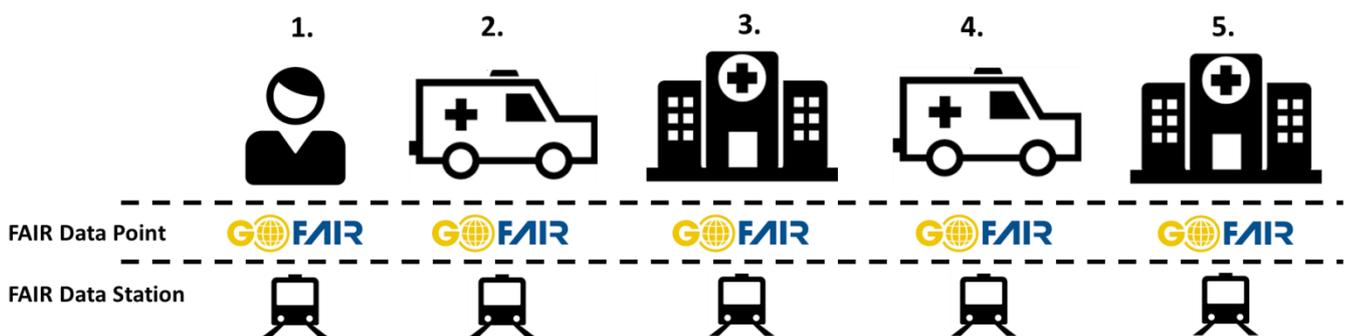
1. A patient has a stroke. He calls 112 to be taken to the hospital urgently.
2. The ambulance (Friesland service) picks up the patient and takes him to the hospital in Leeuwarden. Relevant information from the ambulance service's own registration system is published in a FAIR Data Point. NB: This data point can only be consulted by those who are permitted to access it.
3. In the Leeuwarden hospital, the patient is diagnosed as having had a stroke, and he gets intravenous thrombolysis (IVT). The hospital then suspects that the patient is eligible for

an IAT treatment and a second ambulance is arranged. Relevant information from its own registration system is again published in a FAIR Data Point.
4. The second ambulance (Groningen service) takes the patient from the hospital to the UMCG (an IAT treatment hospital). Relevant information from its own registration system is again published in a FAIR Data Point.
5. The UMCG performs the IAT treatment and only publishes its own relevant information in a FAIR Data Point.

Ultimately, the Personal Health Train can be used to obtain information from this supply chain about the turnaround times of the process. A 'train' runs from the UMCG or another organisation that wants to have this information. This train runs along the different stations and picks up the relevant information. During the time the train is underway, BSN data are taken along to be able to track a patient, but ultimately only the indicators related to turnaround time are made available.

Links:
https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud
https://informatiemodel.istandaarden.nl/2018/https://www.dtls.nl/fair-data/personal-health-train/

(Preliminary) Conclusions

The six use cases in this report describe how different actors from different domains are implementing FAIR in their policies and practice, or how they plan to do so in the future. FAIR is gaining momentum, both in the Netherlands and internationally. As mentioned in the introduction, we are aware that six use cases are not enough to do justice to all the FAIR developments in the Netherlands. However, there are interesting points and commonalities across the different use cases. We list some of them below.

In the *FAIR data Advanced Use Cases: from principles to practice in the Netherlands* workshop on the 22nd of May 2018 in Utrecht, these points will be discussed. The outcomes of this discussion will then be fed back into the final conclusions of the report.

1. FAIR is seen as part of a larger culture change

FAIR is seen as an element of good research data management or data stewardship. And even though FAIR data is not the same as open data, it is seen as part of a larger culture change towards more openness in research and interdisciplinary cooperation. It raises awareness, opens up new innovative ways of working and boosts transfer of knowledge between different domains.

'Going FAIR' is generally seen as a series of improvements. There are always steps ahead that can improve reuse even further. Machine readability is sometimes one of those next steps. The importance of machine readable data is acknowledged in all use cases, but for less data-driven communities there is a tendency to focus on human interoperability first.

FAIR is seen as a driver for innovation. Together with developments such as new national and international privacy regulations and policies and the fact that more and more research communities are using digital data, FAIR highlights the need to update policies and to invest in support and awareness activities, new infrastructures, software and tools.

2. There is a tension between domain specific needs and maximum interoperability

No matter the maturity of the community, there is a tension felt between trying to build on existing domain-specific principles and workflows on the one hand, while trying to get to a maximum level of interoperability with others on the other. The use studies suggest two strategies for alleviating this tension:
- For interoperability, a consensus on minimal cross domain standards is important. For instance, it is important that FAIR is compatible with other broadly shared standards like the W3C Recommendations for Linked Data. Domain specific solutions can build on that.
- It helps to share FAIR examples from different domains to build a better understanding of the potential to align standards and workflows and handling of sensitive data.

3. Policies can't be about FAIR compliance alone

The nature of the different FAIR principles varies. They are not drafted as rules. It is not easy to derive a set of metrics from the principles, especially in the non-life sciences domains. Therefore, to assess compliance with the 15 FAIR guiding principles is seen as difficult. From the examples in the different use cases it becomes apparent that any policy in which FAIR is mentioned should be open to discipline specific solutions which at least satisfy the overarching requirements of **F**indable, **A**ccessible, **I**nteroperable and **R**eusable.

The FAIR principles are often seen as helpful guidance but not a goal in itself. For instance, two of the organisations described in the use cases, Leiden University and Nikhef, have recently drafted new data policies along the lines that data should be Findable, Accessible, Interoperable and Reusable but FAIR is not specifically mentioned in these data policies.

4. **A way forward: integrated approaches with domain specific guidance**

To make it as easy as possible for researchers, it is said that the FAIR data principles need to be translated into more practical guidance. There is a tendency to take an integrated approach when doing so, in which domain-specific needs are leading. Several use cases mention the need to offer well-resourced and innovative support, including domain specific guidance and examples of good practice.

Approaches include: investing in non-mandatory domain specific protocols that have a broader scope than the FAIR principles alone; including FAIR in community focused support for research data management; and including automated tracking of changes, metadata and provenance in research support tools.

5. **FAIR takes effort, but it is worth it**

Another important message coming from different use cases is that it takes effort to get to a certain level of FAIRness. Whether you are developing internationally shared ontologies or software or you want to have trained data stewards in place, it takes resources.

Some communities have already achieved a lot. Their use cases show that once you have reached that level of FAIRness, a whole world of possibilities opens. For instance, the climate data MAGIC project or the international projects in the field of the high energy physics would not have been possible without large investments into **F**indability, **A**ccessibility and (cross-domain) **I**nteroperability over a period of years.

In other domains, allocating large amounts of resources to adjusting their way of working is more challenging. Still, it is generally seen as the only way forward, now that research is becoming more and more data driven. The trend is towards more interdisciplinary collaboration and openness, so that we can all benefit from sharing knowledge.

6. **The Future: recommendations for further exploration**

A few points were raised that are not easily answered. Many issues are linked to the question of how to enable maximum interdisciplinary interoperability. There is a need to align or integrate FAIR with the different regulations that concern data, both on the national and the international level. It will be increasingly important to look at efficient ways of making some data findable without depositing everything in data repositories and also choosing which data to discard.

Other open questions are linked to the long-term business case for the implementation of FAIR: how much effort are we willing to put into, for instance, machine-readability and do we really want to reuse everything? How much effort should go into preparing data for re-use and long-term preservation of datasets? The scope of the financial business case for implementing FAIR still seems unclear and these issues need further exploration**.**

But as one of the use cases points out: if the attention that FAIR is generating ensures that we can all invest in easier access to knowledge, improved reproducibility and more robust research data management, we will have made enormous progress already.