National Health Care Institute

Final report
FAIR Data feasibility assessment

OKTOBER 2018

| Taking care of good health care |

Inhoud

Management summary

This is the final report of the FAIR Data part of the FAIR Data and Personal Health Train feasibility assessment. In this assessment, *Zorginstituut Nederland* applied the FAIR Data principles to its own data, and the Personal Health Train was subsequently implemented in practice. These activities were carried out according to agreements made between the Ministry of Health Welfare and Sport (VWS) and the *Zorginstituut*.

This final report describes what the *Zorginstituut* achieved by applying the FAIR Data principles – 'becoming FAIR'. The final report provides a summary of the results and findings of the first phase: FAIR Data. A separate final report was drawn up regarding the second phase, the Personal Health Train (PHT).

FAIR
Findable
Accessible
Interoperable
Reusable

There is a broad basis of support for FAIR throughout the world and also in the Netherlands. The advantages of being findable, accessible, interoperable and reusable are widely recognised. FAIR Data – and the underlying *linked data* technology – supplement the harmonisation and uniformity of data. It makes information comprehensible for systems without everyone having to speak the same language. It simplifies the linking of information both for research and for operational situations. FAIR Data can further boost data-interoperability and *big data* solutions.

'Becoming FAIR' is currently still quite a technical operation, with a steep learning curve and so far without user-friendly *tooling*. Expectations are that such tools will be made available. For instance, Amazon will be offering *linked data* services in the near future. Switching to FAIR is easy for a person who has experienced the learning curve and has access to data that is already available in a relatively structured form.

In just a few weeks' time, we made 50% of the *Zorginstituut*'s data elements on long-term care and support FAIR. Experience teaches that, as soon as you switch to FAIR, you can quickly start taking advantage of the main benefits of FAIR: linking and interpreting information from more than a single source. The advantages multiply as more data become available as FAIR Data.
After initially immersing ourselves in FAIR Data, we described the data elements from the iStandards – iWlz, iWmo and iJw – in terms of FAIR. During this phase, using FAIR, we were easily able to link data on long-term care with a completely different dataset, in this case the Basic Addresses and Buildings (BAG) register. With FAIR Data, interoperability can be organised at system level, without affecting the sovereignty of the parties or putting a lot of time and energy into realising uniformity of existing standards.

The conclusion of the feasibility assessment is that FAIR Data supplies a valuable contribution to making information and data accessible and legible. Concretely, the *Zorginstituut* is examining whether FAIR Data can be used as point of departure for on-going programmes, such as the iWlz Action Programme and the Nursing Home Care Quality Data Programme.

Introduction

One of the *Zorginstituut's* tasks is to interpret and research new developments in information management, the objective being to take the most promising applications a step further and ultimately improve health care.

The quantity of health care data is growing rapidly. Not only data of care providers and researchers, but also increasingly data of people who consume care. These are important data that can help to improve health care.

But how can they be accessed? And who can put them to good use? The *Zorginstituut* examined two closely related concepts that could help: FAIR Data and the Personal Health Train (PHT).
Between October 2017 and May 2018, the *Zorginstituut* worked on the FAIR Data & Personal Health Train feasibility assessment. These two concepts originated in the academic world and have received a lot of attention in recent months They are mainly studied and used in relation to research and development. By means of this feasibility assessment, *Zorginstituut Nederland* studied FAIR Data and PHT in more depth and used it in an operational health care environment. The assessment was carried out in two phases.

The focus of the first phase was on using the principles of FAIR Data and exploring its implementation in health care. The second phase realised a simulation of the Personal Health Train. This final report provides a summary of the results and findings of the first phase: FAIR Data.
A separate final report was drawn up regarding the second phase, PHT.

Contact

Questions about the FAIR Data & Personal Health Train can be sent to:

Zorginstituut Nederland
Wouter Franke, project manager, wfranke@zinl.nl
Information Standards Team (information management department)

(020) 797 89 48
info@istandaarden.nl
www.istandaarden.nl

Office address
Willem Dudokhof 1
1112 ZA Diemen

Postal address
Postbus 320
1110 AH Diemen

1 About FAIR Data

This section describes the background to FAIR Data. What are its underlying principles and what does the term 'FAIR' mean?

The principles of FAIR Data describe a meticulous and measurable set of qualities on which every good data publication should be based[1]. These principles and their elaboration help to provide solutions to the increasing challenges resulting from data, such as their abundance, the diversity of standards, interoperability and fragmentation.

1.1 FAIR Data principles

FAIR Data principles are general principles that apply to good data management. They have been in use for some time and are broadly acknowledged, recently also by the G20[2], G7[3], the American National Institutes of Health[4] and the European Union[5]. In order to study the implementation of FAIR Data, the European Commission set up an expert group[6], in addition to which FAIR is also used as guiding principle for the European Open Science Cloud'[7]. Lastly, over the course of time, the FAIR principles have been included in the feasibility assessment in the consultative version of the Principles on the Information system for health care[8].

1.2 The name FAIR

FAIR stands for:
Findable: Data are described and indexed properly and they permit a meta-data search.
Accessible: there is a clear description of how to access the data and they can be retrieved – including meta-data – via standard protocols.
Interoperable: Data and meta-data are described in such a way that machines can interpret them and clarity exists about their relationship with other (meta-) data.
Reusable: how data can be used and re-used is clear and the data have an abundance of characteristics.

1.3 FAIR Data principles

These principles are not new. For instance, FAIR Data principles were actually used in the Dutch Government Reference Architecture, though they were not explicitly referred to as 'FAIR'. An example is the Data on the web[9] page in the NORA-wiki.

The FAIR Data principles have been described to a high level of abstraction, and therefore they give few instructions on how to implement them. Explaining how someone can access data can simply mean giving them a telephone number to call, but it can also mean offering a fully automated process of authorisation and authentication.

The principles are particularly important for awareness of good data management – i.e. *data stewardship* – and providing a framework for further implementations. For instance, awareness of good data management could be evident from an obligation – in the event of research applications – to provide results according to FAIR Data principles, or to take the principles into account when creating information standards.

For the record, it is good to realise that FAIR Data should not be equated with *open data*, which are accessible to everyone. Data can be FAIR, without being 'open'.

---

1 https://www.nature.com/articles/sdata201618
2 http://europa.eu/rapid/press-release_STATEMENT-16-2967_en.htm
3 http://www.g7italy.it/sites/default/files/documents/G7%20Science%20Communiqu%C3%A9.pdf
4 https://commonfund.nih.gov/bd2k
5 http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf
6 http://ec.europa.eu/transparency/regexpert/index.cfm?do=groupDetail.groupDetail&groupID=3464
7 https://ec.europa.eu/research/openscience/index.cfm
8 https://www.informatieberaadzorg.nl/publicaties/publicaties/2018/3/26/principles-informatiestelsel-voor-de-zorg
9 https://www.noraonline.nl/wiki/Data_op_het_web

2 Implementing FAIR Data

In this phase of the feasibility assessment, the focus was on learning to use the principles of FAIR Dataand exploring their implementation in health care. We relied heavily on the use of existing technology, e.g. linked data.

The semantic web[10] was used, the best known aspect of which is *linked data*, in order to implement the FAIR Data principles, The semantic web is a set of technological standards[11] that can be used to create the Web of Data. The objective of the Web of Data is to allow computer systems to carry out significant and reliable interactions in a network.

2.1 Linked data

Because *linked data* focuses on gathering data from systems, the implementation of *linked data* is rather technical. This is evident later in the report, in the results of converting a dataset within the *Zorginstituut* to make it FAIR.

An important element when using *linked data* is that you are helping to support a certain degree of diversity. When organisations use different standards – e.g. vocabularies, terminology, classification systems, information standards – but want to use one another's data, the initial idea is often to realise a single standard.

Such a standard is regularly re-designed, because the common basis cannot be incorporated into existing standards. This often results in creating a new standard, while the older standards are not actually replaced.

2.2 Interoperability

Linked data offers a different approach. Existing information is related to one another, i.e. linked, in such a way that people understand one another without having to change their existing systems or standards. For instance, look at the situations in illustrations 1 and 2.

ILLUSTRATION 1

This shows two people whose meaning is identical, though each makes their own semantic choice, by their choice of words. *Linked data* solves this communication problem by registering, under definitions, that 'loo' and 'restroom' are the same concepts. This registration takes place so formally, that systems can automatically make this interpretation.

10 *For more information, see* https://www.w3.org/standards/semanticweb/
11 *These include methods and technologies such as OWL, SKOS, RDF, SPARQL, RIF and GRDDL.*

ILLUSTRATION 2

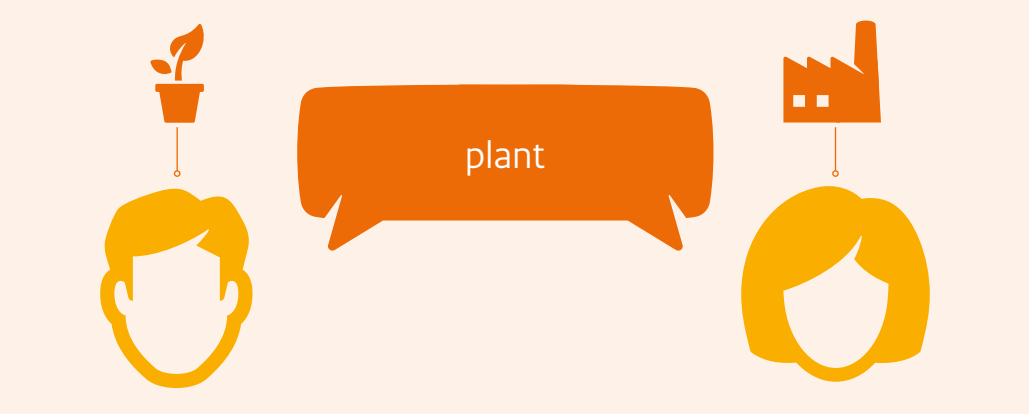


Illustration 2 shows two people who are using the same word, but with completely different pictures to represent their meaning. *Linked data* solves this communication problem by always including the reference to the definition of a data element in a dataset. This takes place with the aid of URIs: *universal resource identifiers*.

With *linked data*, interoperability can be organised at system level, without affecting the sovereignty of the parties or putting a lot of time and energy into realising uniformity of existing standards.

2.3 What stage has linked data currently reached?

The term *linked data* originated around 2006 when the semantic web was taking shape. Over the past ten years, it was subjected to a lot of research and is being used on a large scale. Major technology organisations, e.g. Google and Facebook, use it to connect and display information. Typical semantic web applications are the company information displayed by Google when you carry out a search or the Facebook-timeline.

Linked data and general Graph-technology are frequently used to record knowledge models. Such models are used to apply artificial intelligence, e.g., to improve Google-search results. Due to the rather technical nature of *linked data,* and because companies have often developed their own tools, *linked data* is not yet generally known. As long as it works well, no-one sees the technology – except for its developers.
This touches on another important development point for *linked data*. The tools that are publicly available to make an ontology[12] or to convert a dataset into a *linked data*set, has a steep learning curve and has not yet reached maturity.

Though it is true that large organisations often have their own tools, these are exclusively for their own use. *Tooling* that is available for a broader public is frequently developed as a hobby or due to scientific interest, and as a result is less user-friendly and lacks support. The learning curve is doable for people with a background in information architecture and an affinity for technology.

Making *linked data* and the corresponding technology more generally accessible requires that more user-friendly tools come onto the market with better support. *Zorginstituut Nederland* will be contributing to the spread of such *tooling* by entering into dialogue with parties about the demand.

12 *The semantic web defines an ontology as a description of reality that can be interpreted by computers (a representation of knowledge).*

3 The three steps of the feasibility assessment

In the feasibility assessment a dataset of Zorginstituut Nederland, i.e. the 'Wlz implementation information', was made FAIR in three steps. These data form the basis for information on the long-term care waiting lists that Zorginstituut Nederland publishes each month.

The dataset is described in the iStandards Information model[13]. As this dataset comprises information compiled from the Wlz-chain, the set covers about half of all the data elements in the iStandards: iWlz, iWmo and iJw. At a later stage, all the remaining data elements used within these iStandards were described according to 'FAIR'.

3.1 Step 1: make an ontology

First, we made an ontology of the information recorded in the file. In the ontology we describe the data elements, the relationships between the data elements and the definitions of the elements, and where necessary also in relation to external definitions. The ontology is important for making all meta-data of the dataset which are currently described in the iStandards Information model intelligible for people and for computer systems.

Based on the FAIR ideology, this step is important for making data findable, interoperable and re-usable. As the *Zorginstituut*'s dataset has already been given a structured description, this step is relatively simple. In fact, the model uses UML, textual descriptions and XML scheme definitions. The resulting ontology, which is recorded, technically, in an OWL-file, is published on the FAIR Data Point of the *Zorginstituut*, that was specifically designed for the feasibility assessment.

3.2 Step 2: conversion to linked data

In this step, we converted the dataset to *linked data*. We took the dataset in its current XML-format and converted it to RDF-format. RDF (Resource Description Framework) is the standard format for *linked data*. It describes the data in so-called *RDF triples*. Each triple is comprised of a subject, a predicate and an object. For example: the predicate for a client born in 1922 is 'has the date of birth', while the subject is the client and the object is 1922.

The ontology records 1) which relationships, i.e., which *triples* can exist and 2) the meaning of all concepts. Furthermore, to make the data 'linkable', for every subject the ontology records a URI, i.e., a *universal resource identifier*. Next, based on the ontology and an example of an XML-file, a Java-code is written that converts the data from XML to RDF.

There are various possibilities for converting data to RDF, e.g. using RDF Mapping Language (RML). However, the tools available do not yet support RML sufficiently.

3.3 Step 3: FAIR Data Point

The next step in making a dataset FAIR is to make a FAIR Data Point. To do this, the GoFAIR implementation team created simple *tooling*. During the feasibility assessment, a webserver was designed offering a Sparql-endpoint. Meta-information was published there, e.g. who owns the data and the conditions for using the data. After publication on the FAIR Data Point, the information can be found using the FAIR search engine. The goal is that eventually information will also be indexed and displayed by other search engines, e.g. Google.Illustration 3 gives an impression of the FAIR Data Point.

13 https://informatiemodel.istandaarden.nl/2018/

ILLUSTRATION 3

4 Results and challenges

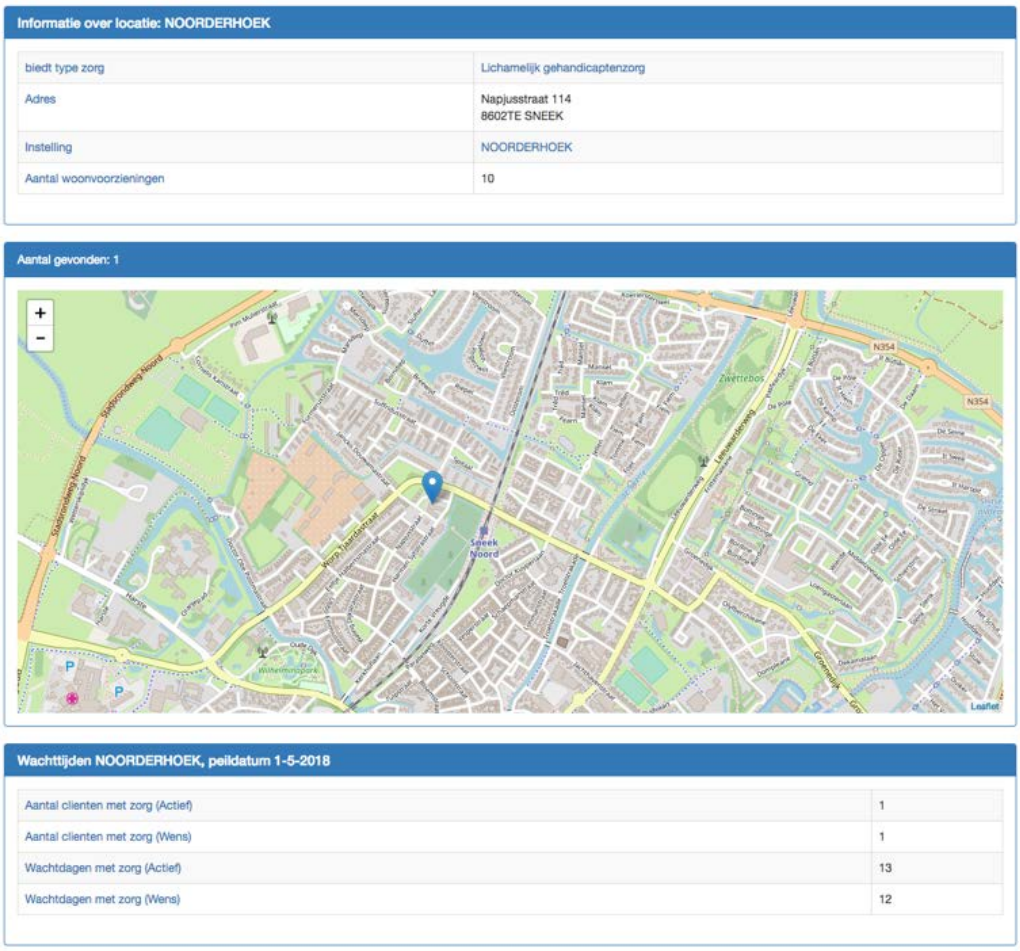
The dataset that was converted in three steps to *linked data*, can easily be linked to other datasets. This can result in all sorts of practical, useful applications that can be presented clearly.

To date, the process of making a dataset FAIR has not resulted in visually attractive results. The fact is that using *linked data* involves rather technical steps. Only then is it possible to get a clear picture of the potential.

4.1 Practical example

A dataset that has been made available as *linked data* can easily be linked to other datasets. As an example, in a short space of time we linked the Wlz implementation information with geographical and demographical information that is available as *linked data* (Illustration 4).

ILLUSTRATION 4



In this example, in which Wlz-data have been linked to the Basic Addresses and Buildings (BAG) register and geographical data, it was not necessary to download other data or first convert them to the same format as the *Zorginstituut*'s dataset.

4.2 Applications in health care

The more 'FAIR' data there are available, the greater the possibilities will be for linking data. FAIR and *linked data* offer considerable possibilities for big data applications within health care. Furthermore, one's own internal data can be enriched with external data without having to duplicate the external data. The same applies to data from various internal sources. This will prevent the same data being stored twice in different places.

4.3 Challenges

Technically, such data linking is done using Sparql queries. During the feasibility assessment, we discovered that the scalability of Sparql needs to be improved. This became apparent from the tool's performance and ease of use. Amazon recently stated that it would be offering environments for Sparql endpoints. Expectations are that this will result in large steps being taken in the field of scalability.

Another challenge is presented by the licences under which data are offered. Whether other parties can use the data, and how, and what control is retained by the owner of the data? Similar questions, which are also being asked regarding *open data*, are important to be clear about the re-use of data. Licences are essential to obtain a clear picture of the possibilities and limitations.

5 Conclusions

The FAIR Data & Personal Health Train feasibility assessment originated within the workspace of research and development. Zorginstituut Nederland has examined its applicability in a broad sense, in more operational environments – such as process and quality information. This section is where we draw conclusions based on the assessment.

5.1 General aspects

The principle of FAIR has been widely accepted. Various organisations, including the G7, G20, EU and hospitals, use FAIR Data because of the promise of data that is findable, accessible, interoperable, and re-usable and which systems are able to use directly. FAIR Data can give a significant impetus to data applications that involve the collation of data from all sorts of locations and of various origins. Based on the results of the feasibility assessment, we endorse such promises.

5.2 Availability of data

- FAIR Data will help us to make data available more effectively and faster, specifically due to its four principles which raise awareness of good data management.
- Making data 'Findable' and 'Accessible' means that information besides being findable and accessible, must also be consistent – they may not 'suddenly disappear' – and that people must put a lot of thought into licences under which data can be shared.
- 'Interoperable' and 'Reusable' are the principles behind working towards understanding one another's data. This demands data with rich descriptions, so that systems are better able to understand one another.

The aim of all these principles is to increase the availability and utility of data. While carrying out the feasibility assessment, it became clear that FAIR data can easily be used forlinking information from different domains, e.g. linking Wlz-data with information from the BAG-register.

FAIR Data contributes to the availability of data based on 1) the human element, by creating awareness of good data management and 2) from a technical perspective, by making data interoperable and findable for systems.

5.3 Transparency and convenience

Due to its transparency, FAIR Data contributes to the principles of re-usable health data[14] and one of the four outcome-objectives of the Information debate: 'once-only recording, multiple uses"15.

The FAIR Data principles are largely about:
- stating what data you have;
- providing rich definitions of the data;
- being clear about how a person can access the data.

Furthermore, FAIR information is also legible for machines. If offers parties who use FAIR data the possibility of computerised interpretation. This offers the extra perspective of using information from multiple sources as well as unifying and synchronising data. The user doesn't have to speak the same language to be able to interpret data electronically. Due to the ontology offered by the provider of information, including cohesion between the definitions used and other frequently used definitions, data elements do not have to be made uniform in advance. The user can interpret the data him/herself at the moment when they are used. This permits interoperability without affecting the sovereignty of the parties or having to put too much time and energy into bringing uniformity into existing standards that have already been implemented.

14 https://www.informatieberaadzorg.nl/publicaties/publicaties/2018/3/26/principes-informatiestelsel-voor-de-zorg
15 https://www.informatieberaadzorg.nl/outcomedoelen/vastleggen

5.4 Technical background

Linked data were used to implement FAIR data. The main added value of *linked data* is that it is independent of other *linked data* sources, and it has the possibility of linking definitions, etc., in other ontologies.

Providing an ontology for long-term care can help in creating ontologies in adjacent domains, such as hospital care or the municipal domain. Vice versa, the presence of ontologies in other domains helps to increase the added value of *linked data*.

5.5 Knowledge and expertise

Implementing FAIR Data and the underlying *linked data* is a fairly technical process. This is due to the focus on connecting data from systems. At the moment limited *tooling* is all that is available to conceal the technology from users. This results in quite a steep learning curve for people who first start to use it. It is doable for people with a background in information and who have an affinity for technology. The GoFair Office is taking various steps to realise training courses[16] and to encourage suppliers of *tooling*.

5.6 Implementation

In the first phase of the feasibility assessment, in a relatively short period of time – four weeks –, we succeeded in describing a large quantity of data elements (i.e. those of the iWlz, iWmo and iJw iStandards) based on 'FAIR'. In addition, we converted a dataset with Wlz implementation information to make them 'FAIR'.

What helped enormously is that the iStandards data already possess a great deal of structure. This means that the time and energy put into standardising and structuring information in the past give you a head start when you want to convert to FAIR.

With proper preparation, the technical process of becoming FAIR can be done quite quickly. Organisation-linked questions, such as under which licence data should be categorised, will take more time. It is important to remember that FAIR Data *is not* the same as *open data*.

16 *See also* https://www.go-fair.org/training/

6 Further steps

The feasibility assessment has shed light on the workings and implementation of FAIR Data. As stated earlier, the *Zorginstituut* is very positive towards these principles and their implementation in *linked data*.

We are currently examining whether the FAIR Data principles can be used as point of departure in various projects in which the *Zorginstituut* is intensely involved.

Examples of such trajectories are the iWlz Action Programme (modernising the provision of Wlz information) and the Nursing Home Care Quality Data programme.