


## Training exercise - Evaluating a dataset using FAIR-Aware (10.5281/zenodo.8089501)

This exercise was created by Maaïke Verburg  (DANS) - November 2021, and is available under a [CC by 4.0 licence](#)

FAIR-Aware was created by DANS in the FAIRsFAIR project<sup>1</sup> and is currently further developed in the FAIR-IMPACT project<sup>2</sup>.

### Trainer instructions

---

#### **This training exercise has three purposes:**

- To let participants learn where to look to extract information on the FAIRness of a dataset from its landing page and the information a repository publicly showcases;
- To reflect on what the implementation of FAIR looks like and how you can find evidence of FAIRness in datasets;
- To translate such information to realise what specific skills and actions they need to undertake to make their own data more FAIR.

#### **Set-up or programme of the training**

- Start with a general introduction to FAIR-Aware, its basic purpose, and where to find the additional guidance information in the tool
- Explain the exercise and its purposes
- Let the participants do the exercise (either in groups or individually)
- Report back to the full group in plenary to discuss findings, takeaways, and reflections (in case of individual exercise, it is recommended to first discuss findings in smaller groups before going plenary)

Suggested duration of the full exercise as described here is 60 minutes, but the programme is flexible for shorter or longer durations.

---

<sup>1</sup> FAIRsFAIR “Fostering FAIR Data Practices In Europe” has received funding from the European Union’s Horizon 2020 project call H2020-INFRAEOSC-2018-2020 Grant agreement 831558. The content of this document does not represent the opinion of the European Union, and the European Union is not responsible for any use that might be made of such content.

<sup>2</sup> FAIR-IMPACT “Expanding FAIR solutions across EOSC” is funded by the European Union.

## Materials needed

- Online access to: <https://fairaware.dans.knaw.nl/>
- The **worksheet for participants** to fill in (see below)
- **Datasets** to work on (see below, or select your own or let participants bring in their own dataset to assess)
- Preferably, a **worked example** of the exercise for each dataset you use (to make sure you have a balanced set of datasets, and to make sure you have the knowledge on the datasets to help participants out during the exercise or when reflecting at the end). The datasets listed have been worked out in a separate section below.

## Dataset options

This list shows a few options for datasets to use in this exercise, with some explanations behind them on why specifically they could be interesting to assess. Of course, this exercise can work with any dataset. Also linked are the relevant policies from the repositories the datasets are deposited in, which can give additional information on the potential FAIRness of the data.

When preparing this exercise, copy the link to the dataset and link to the policy document to the worksheet for the participants. You can let multiple people/groups assess the same dataset to see if they come to similar conclusions or not.

- Let participants use their own dataset *(most effective when it is deposited and published in a repository)*
- Verhoef, J., Rijksuniversiteit Leiden \* Leiden, Fac. sociale wetenschappen, vakgroep politieke wetenschappen (primary investigator); (1917): Dutch election data, 1888-1917. DANS. <https://doi.org/10.17026/dans-zz4-phyy> *(This is quite an old dataset, which makes it interesting to see if the relatively new FAIR principles can be applied here)*
  - DANS Preservation Policy: <https://dans.knaw.nl/en/preservationplan/>
  - DANS Data Station Policy: <https://dans.knaw.nl/wp-content/uploads/2023/04/DANS-Data-Stations-Policy-20230524.pdf>
- Rammstedt, Beatrice, Martin, Silke, Zabal, Anouk, Konradt, Ingo, Maehler, Débora, Perry, Anja, Massing, Natascha, . . . Helmschrott, Susanne (2016). Programme for the International Assessment of Adult Competencies (PIAAC), Germany - Reduced Version. GESIS Data Archive, Cologne. ZA5845 Data file Version 2.2.0, <https://doi.org/10.4232/1.12660> *(This dataset has had multiple version uploads, which makes it interesting to see if the FAIRness of the dataset was improved through new versions)*
  - Gesis information on data sharing: <https://www.gesis.org/en/datenservices/daten-teilen/how-to-guide-daten-teilen>
- Klages, Johann Philipp and colleagues; Expedition PS104 Scientists (2019): Apatite radiogenic isotopes and ages of sediment core 2R at site PS104\_20-2. PANGAEA, <https://doi.org/10.1594/PANGAEA.906168> *(This dataset gives a good example of the importance of provenance and vocabulary use, which is interesting as they can be difficult concepts to grasp)*
  - Pangaea data policy: [https://wiki.pangaea.de/wiki/Main\\_Page](https://wiki.pangaea.de/wiki/Main_Page) / [https://wiki.pangaea.de/wiki/Data\\_policy](https://wiki.pangaea.de/wiki/Data_policy)

## Participant Instructions and materials

*Copy this worksheet for each participant or group undertaking the exercise. Instructions or tips for trainers are included in the text below in italics and lighter grey font. Make sure to delete those additional instructions before using the materials for your participants.*

---

### Evaluating a dataset using FAIR-Aware

#### Goal of the exercise

What does it mean to put FAIR into practice? How can FAIR knowledge be translated into FAIR skills and how can this be applied to a dataset? This is what the FAIR-Aware tool aims to help you with.

The goal of this exercise is to reflect on:

- 1) What does the implementation of FAIR practices look like in a deposited dataset?
- 2) How does the implementation of FAIR practices aid reuse of the data?
- 3) How easy or difficult is it for humans to evaluate the FAIRness of a dataset?

With your group, evaluate the assigned dataset based on the **10 FAIR practices** detailed in the FAIR-Aware tool.

Use the **guidance texts** in the FAIR-Aware tool to discover how FAIR practices can be implemented (see the “How to do this?” sections). Use the dataset and the repository information to determine whether these practices are satisfied.

Gather **evidence**: How can you tell that this FAIR practice is or isn't satisfied?

Discuss **shortcomings**: How could the dataset be improved on this FAIR practice?

Discuss **importance**: How does this FAIR practice facilitate reuse or other important aspects of FAIR? Taking the perspective of a potential reuser of the data, how does the implementation of the FAIR practice help you in making your choice?

**FAIR-Aware:** <https://fairaware.dans.knaw.nl/>

**Group #:** (Assign the group numbers, delete in case exercise is done individually)

**Dataset:** (Put link to the assigned dataset here, or let participant fill in their own dataset)

**Repository policy:** (In case the repository where the dataset is deposited has a public link to their relevant policies, link them here to let participants use this source of information to better answer some of the questions)

FAIR Practice	Applied to data?	Evidence:	Room for improvement:	Comments:
1. Persistent identifier	Yes / No			
2. Discovery metadata	Yes / No			
3. Metadata for humans and machines	Yes / No			
4. Access control metadata	Yes / No			
5. Persistence of metadata	Yes / No			
6. Controlled vocabularies	Yes / No			

<b>7. Provenance information</b>	Yes / No			
<b>8. Community-endorsed metadata standards</b>	Yes / No			
<b>9. Preferred file formats</b>	Yes / No			
<b>10. Digital curation and preservation</b>	Yes / No			

## Reflection / highlights

After completing the exercise, please share your reflections and other takeaways from this exercise. *(You can also let the participants submit this feedback to a general place of collection (e.g., survey, polling, mentimeter))*

1. What does the implementation of FAIR practices look like in a deposited dataset?
2. How does the implementation of FAIR practices aid reuse of the data?
3. How easy or difficult is it for humans to evaluate the FAIRness of a dataset?

## Worked examples of the datasets

*Below, you can find the training exercise filled in for each of the example datasets listed above. These contain some of the possible answers participants can give, which can be used to help participants along in discussion and bring up interesting questions for the plenary discussion with the group. Please note that these suggestions may be outdated or incorrect if changes have been applied to the dataset since (date of check 27 June 2023). It is recommended that the trainer goes through these examples and potentially adds their own views and comments to be familiar with what your participants are working on. If you add other datasets, it is recommended to do the same for those.*

**Dataset:** Verhoef, J., Rijksuniversiteit Leiden \* Leiden, Fac. sociale wetenschappen, vakgroep politieke wetenschappen (primary investigator); (1917): Dutch election data, 1888-1917. DANS. <https://doi.org/10.17026/dans-zz4-phyu>

FAIR Practice	Applied to data?	Evidence:	Room for improvement:	Comments:
<b>1. Persistent identifier</b>	Yes / No	DOI is linked  'Other ID' and 'Data Vault Metadata' fields indicate other identifiers for the dataset	PID could also have been used for the creator (ORCID) and data collector (ROR)	
<b>2. Discovery metadata</b>	Yes / No	Creator, title, date submitted, description, subject keywords PID  Data content is described  Access rights are included	Terms for Access for Restricted Files aren't defined  If any, relevant relations could have been expressed more explicitly	Sharing options for social media to increase online presence

			More keywords e.g., about content would be good	
<b>3. Metadata for humans and machines</b>	Yes / No	DANS has OAI-PMH protocol in place		This information needs to be found on the repository level, not dataset level
<b>4. Access control metadata</b>	Yes / No	6/8 files are restricted access, this is clearly described in 'terms' tab  You can contact the data owner for access	Specifying terms of access (repository does have this field, but it wasn't used for this dataset)	
<b>5. Persistence of metadata</b>	Yes / No	Not applicable since data is still available.  In Data Station policy it is stated that 'DANS will always provide open access to all published metadata'	Explicitly state in preservation plan what the standard data retention period is for the archive	Since this is an old dataset, you can gather that generally, data remains available at DANS for a long time.
<b>6. Controlled vocabularies</b>	Yes / No	Since the Data Station SSH is new at this time, there is no overview of what vocabularies are supported and used yet	Explicit statement about which vocabulary was used for this dataset would help potential re-users	Hard to tell to what extent the (meta)data adheres to the standard



<p><b>7. Provenance information</b></p>	<p>Yes / No</p>	<p>Temporal &amp; Spatial coverage and submission dates</p> <p>Includes contributors</p> <p>Data generation is described in the dataset description</p> <p>Version log with information of changes</p>	<p>More information on methods, instruments, protocols would enrich the provenance information</p>	<p>‘Documentatie en codeboek’ is available as a file (restricted). Potentially, information from there could have been supplied in the metadata (which would be accessible)</p>
<p><b>8. Community-endorsed metadata standards</b></p>	<p>Yes / No</p>	<p>In DANS preservation policy it is detailed that the Dublin Core standard is used, as well as that this information is also mapped to other standards</p>		<p>Documentation on Data Station SSH not yet available, so this is based on what it was like in the previous system: EASY</p>
<p><b>9. Preferred data format</b></p>	<p>Yes / No</p>	<p>DANS has an overview of preferred data formats. It seems that the same data is uploaded in multiple formats</p>	<p>For a codebook, pdf doesn’t seem the most reusable format</p>	<p>It’s hard to tell whether the pdf files are pdf/A</p>
<p><b>10. Digital curation and preservation</b></p>	<p>Yes / No</p>	<p>DANS is a CTS certified repository</p>	<p>Since the data station is quite new, not all information and documentation is clearly findable yet at this time</p>	

		The data station is discipline-specific, tailored to the communities specific needs		
--	--	-------------------------------------------------------------------------------------	--	--

**Dataset:** Rammstedt, Beatrice, Martin, Silke, Zabal, Anouk, Konradt, Ingo, Maehler, Débora, Perry, Anja, Massing, Natascha, . . . Helmschrott, Susanne (2016). Programme for the International Assessment of Adult Competencies (PIAAC), Germany - Reduced Version. GESIS Data Archive, Cologne. ZA5845 Data file Version 2.2.0, <https://doi.org/10.4232/1.1266>

FAIR Practice	Applied to data?	Evidence:	Room for improvement:	Comments:
<b>1. Persistent identifier</b>	Yes / No	DOI is linked		
<b>2. Discovery metadata</b>	Yes / No	Title, abstract, PI, contributor, publisher, study number, PID, data content information, relations,	Data Access could me more explicit Relations could be more explicit (how do they relate?)	
<b>3. Metadata for humans and machines</b>	Yes / No	Unclear whether they use protocol in links		In terms of human readability, the design/lay-out of the page is not very human-friendly
<b>4. Access control metadata</b>	Yes / No	Information in “request data access”  Information on how to access the data includes a data usage contract, link to pricing and terms of use. The criteria for access are also detailed		Note that for FAIR, it doesn’t matter that this data is not free or openly accessible. It only matters that it is clearly and transparently described what the situation is

<b>5. Persistence of metadata</b>	Yes / No	Not applicable to dataset that is still available.  Unclear if this topic is covered in policy explicitly		
<b>6. Controlled vocabularies</b>	Yes / No	Unclear		
<b>7. Provenance information</b>	Yes / No	Very extensive, many metadata fields for provenance information.  Version history with change log and DOIs of previous versions		Questionnaire, codebook, and 'other documents' are available for download (free, no access request needed, in pdf format)
<b>8. Community-endorsed metadata standards</b>	Yes / No	Gesis uses DDI		
<b>9. Preferred data format</b>	Yes / No	Gesis provides information on preferred file formats on their website, based on different data types	The file format of a document only becomes clear after downloading it, not in advance	
<b>10. Digital curation and preservation</b>	Yes / No	GESIS certification has expired, so no easy direct indication	A clear webpage/document on what the repository does to be trustworthy could be a good indication even when no	Not being formally certified does not mean that the repository is automatically not trustworthy. It could also be that they're in the

			formal certification is obtained	process of re-certifying, which can take some time
--	--	--	----------------------------------	----------------------------------------------------

**Dataset:** Klages, Johann Philipp and colleagues; Expedition PS104 Scientists (2019): Apatite radiogenic isotopes and ages of sediment core 2R at site PS104\_20-2. PANGAEA, <https://doi.org/10.1594/PANGAEA.906168>

FAIR Practice	Applied to data?	Evidence:	Room for improvement:	Comments:
<b>1. Persistent identifier</b>	Yes / No	DOI is linked	ORCID's could be a good addition (Pangaea does support this PID)	
<b>2. Discovery metadata</b>	Yes / No	Title, creator, etc only in citation, not separately as metadata  Relations expressed to publications and projects  Some data content information, such as the amount of data points and the spatial covering (map view)	More explicit fields per element	There is information, but there could be more. This is a tricky part of assessing and scoring on the FAIR principles; when is metadata 'rich' enough?
<b>3. Metadata for humans and machines</b>	Yes / No	OAI-PMH protocol + Schema.org		
<b>4. Access control metadata</b>	Yes / No	Not explicit what access condition there is, only a direct download link	State rights holder and contact, state exact access conditions	

		Licence clearly added		
<b>5. Persistence of metadata</b>	Yes / No	Not applicable to dataset that is still available.  Unclear if this topic is covered in policy explicitly		
<b>6. Controlled vocabularies</b>	Yes / No	Long list of parameters, some with comments on source vocabulary / device	Use all terms from one vocabulary as much as possible (suggestion for depositor)	This is where you see that a repository can support good practices, but it's up to the researcher to implement it. Truly the collaboration between the two will facilitate FAIR data!
<b>7. Provenance information</b>	Yes / No	Extensive, coverage, events, map view of spatial coverage	Version history can be more clearly documented, e.g., in a change log	
<b>8. Community-endorsed metadata standards</b>	Yes / No	Unclear, list of metadata fields is provided, but no explicit mention of a standard is made	Mention explicit standard to adhere to	Don't reinvent the wheel, base yourself on standards so other people know what to expect
<b>9. Preferred data format</b>	Yes / No	Guidance page in Wiki provided. A bit limited in terms of data types included, but could be because the	A table to link data type to preferred formats gives the best overview	Efficient work, especially in domain-specific repositories, also means to not cover things outside of your domain or what you will encounter.

		data they receive doesn't come in many different types to begin with		However, what do you then when you do encounter a new situation? Some plans for that can be helpful.
<b>10. Digital curation and preservation</b>	Yes / No	Clear list of certifications in the website footer	It is always a nice addition (though not necessary) to include a more explicit dedication to being trustworthy and how this is approached	Also pilot-repository of F-UJI, so extra focus on FAIR data. Though this is not promoted much on the website.

**NB: People assessing their own data will have variable results depending on their own choices and practices.**