SYNOP Data Evaluation Using FAIR Maturity Model

Amina Annane, Mouna Kamel, Cassia Trojahn, Nathalie Aussenac-Gilles, Catherine Comparot, Christophe Baehr

SYNOP Data Evaluation Using FAIR Maturity Model

Amina Annane, Mouna Kamel, Cassia Trojahn,
Nathalie Aussenac-Gilles, Catherine Comparot and Christophe Baehr

March 2021

1 Introduction

SYNOP dataset is one of the open meteorological datasets privided by Météo-France on its data portal. The dataset includes observation data from international surface observation messages circulating on the Global Telecommunication System (GTS) of the World Meteorological Organization (WMO) . The choice of this dataset is motivated by the fact that these data are open and free, they concern several atmospheric parameters measured (temperature, humidity, wind direction and force, atmospheric pressure, precipitation height, etc.). These parameters are important for many scientific studies, particularly because SYNOP dataset are published by all WMO member states. SYNOP data is published as open data. On its web page[1], it is described by seven items: (i) *description* : natural language summary that describes the content of the dataset, (ii) *conditions of access* : Etalab license[2] for the data, (iii) *means of access* : specifies that the data can be accessed via direct download, (iv) *download* : possibility offered to the user to download the data in csv format for a given date, (v) *download archived data* : similar to the previous item, but for a given month, (vi) *station information* : list of stations (station id, name) accompanied by a map displaying the location of these stations, and (vii) *documentation*: three links that respectively reference (a) a pdf file that explains the acronyms (label, type, unit of measurement) present in the header of the SYNOP data files, (b) a csv file that describes the different meteorological stations of Météo-France (id_station, name, latitude, longitude, altitude), and (c) a JSON file containing the same information as the previous csv file. Table 1 shows an extract of the downloadable SYNOP data. The file contains 59 columns, the first two correspond to the station number and the date of the measurements made, the other 57 columns to the meteorological measurements.

[1] `https://donneespubliques.meteofrance.fr/?fond=produit&id_produit=90&id_rubrique=32`

[2] `https://www.etalab.gouv.fr/wp-content/uploads/2014/05/Licence_Ouverte.pdf`

Table 1: Extract from SYNOP data.

numer_sta	date	pmer	ff	t	...
7005	20200201000000	100710	3.200000	285.450000	...
7015	20200201000000	100710	7.700000	284.950000	...
7020	20200201000000	100630	8.400000	284.150000	...
7027	20200201000000	100770	5.500000	285.650000	...
...	...	...	...	...	...

2 FAIR Data Maturity Model

It has been developed by the RDA[3] FAIR Data Maturity Model working group [RDA, 2020]. The model is based on three main components: (i) indicators, (ii) priorities, and (iii) evaluation methods.

Indicators: the individual aspects of FAIRness that are evaluated. An indicator aims to measure the state or level of a digital resource with regard to a specific FAIR principle (e.g., F1, A2). Indicators are derived from the FAIR principles [Wilkinson et al., 2016]. The approach for developing the indicators was to create an indicator for each aspect that could be distinguished in the description of each principle. For instance from the principle *"F1. (Meta)data are assigned a globally unique and persistent identifier"*, four indicators are extracted:

- RDA-F1-01M Metadata is identified by a persistent identifier
- RDA-F1-01D Data is identified by a persistent identifier
- RDA-F1-02M Metadata is identified by a globally unique identifier
- RDA-F1-02D Data is identified by a globally unique identifier

Following this approach, 41 indicators have been extracted (see Tables 2 and 3).

Priorities: It consists in assigning the degree of importance to each indicator. Three levels are defined so that an indicator may be:

- Essential: FAIRness would be practically impossible to achieve if the indicator were not satisfied. e.g., *"Metadata is identified by a persistent identifier"*
- Important: its satisfaction, if at all possible, would substantially increase FAIRness. e.g. *"Metadata uses machine-understandable knowledge representation"*

[3] The Research Data Alliance (RDA) is a research community organization started in 2013 by the European Commission, the American National Science Foundation and National Institute of Standards and Technology, and the Australian Department of Innovation. Its mission is to build the social and technical bridges to enable open sharing of data. The RDA vision is researchers and innovators openly sharing data across technologies, disciplines, and countries to address the grand challenges of society.

Table 2: FAIR maturity model indicators (1)

N°		ID	INDICATORS	PRIORITY
1	F	RDA-F1-01M	Metadata is identified by a persistent identifier	Essential
2		RDA-F1-01D	Data is identified by a persistent identifier	Essential
3		RDA-F1-02M	Metadata is identified by a globally unique identifier	Essential
4		RDA-F1-02D	Data is identified by a globally unique identifier	Essential
5		RDA-F2-01M	Rich metadata is provided to allow discovery	Essential
6		RDA-F3-01M	Metadata includes the identifier for the data	Essential
7		RDA-F4-01M	Metadata is offered in such a way that it can be harvested and indexed	Essential
8	A	RDA-A1-01M	Metadata contains information to enable the user to get access to the data	Important
9		RDA-A1-02M	Metadata can be accessed manually (i.e. with human intervention)	Essential
10		RDA-A1-02D	Data can be accessed manually (i.e. with human intervention)	Essential
11		RDA-A1-03M	Metadata identifier resolves to a metadata record	Essential
12		RDA-A1-03D	Data identifier resolves to a digital object	Essential
13		RDA-A1-04M	Metadata is accessed through standardised protocol	Essential
14		RDA-A1-04D	Data is accessible through standardised protocol	Essential
15		RDA-A1-05D	Data can be accessed automatically (i.e. by a computer program)	Important
16		RDA-A1.1-01M	Metadata is accessible through a free access protocol	Essential
17		RDA-A1.1-01D	Data is accessible through a free access protocol	Important
18		RDA-A1.2-01D	Data is accessible through an access protocol that supports authentication and authorisation	Useful
19		RDA-A2-01M	Metadata is guaranteed to remain available after data is no longer available	Essential

Table 3: FAIR maturity model indicators (2)

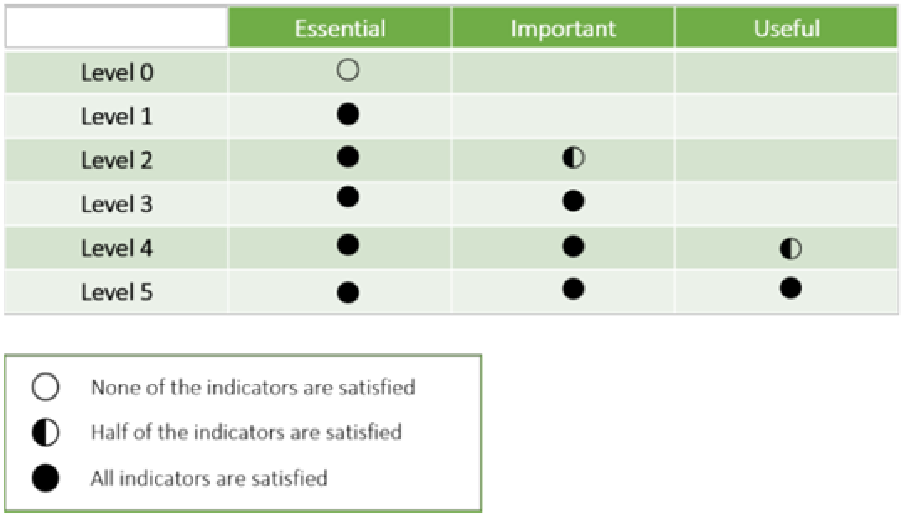
20	I	RDA-I1-01M	Metadata uses knowledge representation expressed in standardised format	Important
21		RDA-I1-01D	Data uses knowledge representation expressed in standardised format	Important
22		RDA-I1-02M	Metadata uses machine-understandable knowledge representation	Important
23		RDA-I1-02D	Data uses machine-understandable knowledge representation	Important
24		RDA-I2-01M	Metadata uses FAIR-compliant vocabularies	Important
25		RDA-I2-01D	Data uses FAIR-compliant vocabularies	Useful
26		RDA-I3-01M	Metadata includes references to other metadata	Important
27		RDA-I3-01D	Data includes references to other data	Useful
28		RDA-I3-02M	Metadata includes references to other data	Useful
29		RDA-I3-02D	Data includes qualified references to other data	Useful
30		RDA-I3-03M	Metadata includes qualified references to other metadata	Important
31		RDA-I3-04M	Metadata include qualified references to other data	Useful
32	R	RDA-R1-01M	Plurality of accurate and relevant attributes are provided to allow reuse	Essential
33		RDA-R1.1-01M	Metadata includes information about the licence under which the data can be reused	Essential
34		RDA-R1.1-02M	Metadata refers to a standard reuse licence	Important
35		RDA-R1.1-03M	Metadata refers to a machine-understandable reuse licence	Important
36		RDA-R1.2-01M	Metadata includes provenance information according to community-specific standards	Important
37		RDA-R1.2-02M	Metadata includes provenance information according to a cross-community language	Useful
38		RDA-R1.3-01M	Metadata complies with a community standard	Essential
39		RDA-R1.3-01D	Data complies with a community standard	Essential
40		RDA-R1.3-02M	Metadata is expressed in compliance with a machine-understandable community standard	Essential
41		RDA-R1.3-02D	Data is expressed in compliance with a machine-understandable community standard	Important

- Useful: nice to have but is not necessarily indispensable. e.g., *"Data uses FAIR-compliant vocabularies"*

From the 41 indicators, 20 (48.8%) are essential, 14 (34.1%) are important and 7 (17.1%) are useful. Classifying indicators into essential, important and useful may be changed from one domain to another as highlighted by the authors.

Evaluation methods the way that the results of the evaluation of the indicators can be interpreted/given a value. The model defines two evaluation methods:

1. Measuring progress: this method attempts to answer the question "*How the FAIRness of this data can be improved?*", hence it is suitable for data providers that want to improve the FAIRness of their data. It consists in assigning a maturity level (value from 0 to 4) for each indicator such that each value has the following meaning:
   - 0: not applicable
   - 1: not being considered this yet
   - 2: under consideration or in planning phase
   - 3: in implementation phase
   - 4: fully implemented
2. Measuring pass or fail: this method evaluates how a resource performs on meeting the indicators across the FAIR area. A binary answer (0 or 1) is assigned on each of the indicators. Finally, the method affects to each principle a FAIRness level taking into account the priorities as shown in Figure 1. This method is particularly suitable for external evaluators.



	Essential	Important	Useful
Level 0	○		
Level 1	●		
Level 2	●	◐	
Level 3	●	●	
Level 4	●	●	◐
Level 5	●	●	●

○	None of the indicators are satisfied
◐	Half of the indicators are satisfied
●	All indicators are satisfied

Figure 1: Fail or pass method: FAIRness levels

3 SYNOP data evaluation

We present the evaluation of SYNOP dataset against the indicators listed in Table 2 according to the RDA maturity model described previously. We adopt the pass/fail method since we are an external evaluator. More particularly, we present the value of each indicator (i.e., 0 or 1) before the generation of semantic metadata, and after their generation with a comment to explain the attributed value each time.

3.1 Findability

3.1.1 RDA-F1-01M. Metadata is identified by a persistent identifier

- (Before: 0) The URL of the home page may be considered as an identifier, but URLs are not persistent.
- (After: 0) The generation of persistent identifiers for Synop metadata should be managed by Météo-France using a third-party.

3.1.2 RDA-F1-01D. Data is identified by a persistent identifier

- (Before: 0) No identifier is assigned to the dataset, unless we consider the home page URL which is not persistent.
- (After: 0) The generation of persistent identifiers for Synop data should be performed by Météo-France

3.1.3 RDA-F1-02M. Metadata is identified by a globally unique identifier

- (Before: 0) No metadata records.
- (After: 0) The generation of unique identifiers for metadata records should be managed by Météo-France.

3.1.4 RDA-F1-02D. Data is identified by a globally unique identifier

- (Before: 0) The URL of the Synop data, which can be considered as an identifier, offers a form to select one dataset from all the Synop datasets (one dataset per month), hence it is not unique.
- (After: 0) The generation of unique identifiers for metadata records should be managed by Météo-France.

3.1.5 RDA-F2-01M. Rich metadata is provided to allow discovery

- (Before: 0) To assess this indicator, we have to define what rich metadata is. Note that we are evaluating the findability, hence we are interested in metadata that supports the findability such as keywords, theme, spatial coverage, etc. Most of these metadata are not available, hence we gave the value of zero. In addition, metadata are not machine actionable and therefore they are not indexed by search engines like google dataset search, which compromises the findability of the dataset.
- (After: 1) We use GeoDCAT-AP vocabulary to describe SYNOP datasets. This vocabulary offers a plurality of metadata that promotes their findability.

3.1.6 RDA-F3-01M. Metadata includes the identifier for the data

- (Before: 0) Synop datasets have no identifier.
- (After: 0) GeoDCAT-AP has a metadata element that points the data identifier, but Synop datasets have no identifier.

3.1.7 RDA-F4-01M. Metadata is offered in such a way that it can be harvested and indexed

- (Before: 0) No machine-actionable metadata are available.
- (After: 0*) We have generated machine-actionable metadata, but we have not yet published them on the WEB.

3.2 Accessibility

3.2.1 RDA-A1-01M. Metadata contains information to enable the user to get access to the data

- (Before: 1*) Download form is available on the landing page, but no machine actionable metadata with the download link.
- (After: 1) A dcat:Distribution instance to describe each Synop monthly file, with the download link.

3.2.2 RDA-A1-02M. Metadata can be accessed manually (i.e. with human intervention)

- (Before: 1) Few metadata are available in natural language and can be accessed manually on the landing page
- (After: 1) We have generated richer semantic metadata, but we have not published them on the web

3.2.3 RDA-A1-02D. Data can be accessed manually (i.e. with human intervention)

- (Before: 1) Synop data can be download manually using the form on the landing page.
- (After: 1)

3.2.4 RDA-A1-03M. Metadata identifier resolves to a metadata record

- (Before: 0) No metadata record.
- (After: 0) Generation of metadata identifiers should managed by Météo-France.

3.2.5 RDA-A1-03D. Data identifier resolves to a digital object

- (Before: 0) Synop datasets do not have unique identifiers, they are accessible using a form on the landing page.
- (After: 0) Data identifiers should be generated for Synop datasets, eventually using a third-party to ensure that identifiers are persistent and globally unique.

3.2.6 RDA-A1-04M. Metadata is accessed through standardised protocol

- (Before: 0) No metadata record for each Synop dataset
- (After: 0) we have not published the generated semantic metadata yet

3.2.7 RDA-A1-04D. Data is accessible through standardised protocol

- (Before: 1) Data is accessible via the form on the landing page using http protocol

- (After: 1) We have represented the download link of each synop dataset, which allows humans and machine to access data.

3.2.8 RDA-A1-05D. Data can be accessed automatically (i.e. by a computer program)

- (Before: 0) the user has to manually select the day or the month for which he wants SYNOP data.

- (After: 1) semantic metadata that we have generated allow to access data automatically.

3.2.9 RDA-A1.1-01M. Metadata is accessible through a free access protocol

- (Before: 0) No metadata records

- (After: 0) We have not published the semantic-metadata yet.

3.2.10 RDA-A1.1-01D. Data is accessible through a free access protocol

- (Before: 1) Data is accessible to download with the form on the landing page using the free access protocol http.

- (After: 1)

3.2.11 RDA-A1.2-01D. Data is accessible through an access protocol that supports authentication and authorisation

- (Before: 1) as explained previously data is accessible using http that supports authentication and authorisation

- (After: 1) No change

3.2.12 RDA-A2-01M . Metadata is guaranteed to remain available after data is no longer available

- (Before: 0) machine-actionable metadata record does not exist

- (After: 0) semantic metadata are generated, but we have not deposited it in a warehouse that guarantees its persistence yet.

3.3 Interoperability

3.3.1 RDA-I1-01M . Metadata uses knowledge representation expressed in standardised format

- (Before: 0) machine-actionable metadata record does not exist
- (After: 1) we have generated machine-actionable metadata record with FAIR vocabularies, mainly GeoDCAT-AP, using RDF format.

3.3.2 RDA-I1-01D . Data uses knowledge representation expressed in standardised format

- (Before: 0) Data is in CSV format without any knowledge representation
- (After: 1) Data schema is represented using FAIR ontologies, mainly RDF data cube, and many other domain ontologies.

3.3.3 RDA-I1-02M . Metadata uses machine-understandable knowledge representation

- (Before:0) No semantic metadata
- (After: 1) We have generated semantic metadata using FAIR ontologies in RDF format

3.3.4 RDA-I1-02D . Data uses machine-understandable knowledge representation

- (Before: 0) CSV files with no machine-understandable knowledge representation of SYNOP datasets
- (After: 1) Data schema is represented using ontologies, hence the meaning of data is represented in a machine-understandable format.

3.3.5 RDA-I2-01M . Metadata uses FAIR-compliant vocabularies

- (Before: 0) semantic metadata.
- (After: 1) Semantic metadata using FAIR vocabularies, namely DCAT.

3.3.6 RDA-I2-01D . Data uses FAIR-compliant vocabularies

- (Before: 0) No knowledge representation of data
- (After: 1) Data schema is represented using FAIR ontologies, mainly RDF data cube.

3.3.7 RDA-I3-01M. Metadata includes references to other metadata

- (Before: 0) No references to other metadata.
- (After: 1) Usage of multiple vocabularies for metadata properties.

3.3.8 RDA-I3-01D . Data includes references to other data

- (Before: 1*) "numer_sta" column references the station dataset, but no semantic referencing
- (After: 1+) the "numer_sta" has been represented in a semantic manner as a foreign key.

3.3.9 RDA-I3-02M . Metadata includes references to other data

- (Before: 0) No references to other data
- (After: 1) references the station dataset with the new property ":require"

3.3.10 RDA-I3-02D . Data includes qualified references to other data

- (Before: 0) No qualified references, the single reference is "numer_sta" but not explicit relation as a foreign key.
- (After: 1) explicit qualified reference with csvw vocabulary property "csvw:foreignkey"

3.3.11 RDA-I3-03M . Metadata includes qualified references to other metadata

- (Before: 0) No references.
- (After: 1) multiple qualified references to other vocabularies specified by GeoDCAT-AP. In addition, "qb:concept" references to domain ontologies.

3.3.12 RDA-I3-04M. Metadata include qualified references to other data

- (Before: 0) no references to other data.
- (After: 1) qualified reference ":requires" to the station dataset.

3.4 Reusablility

3.4.1 RDA-R1-01M . Plurality of accurate and relevant attributes are provided to allow reuse

- (Before: 0) The description of the data columns, given as a PDF document, is not clear, nor precise. For instance, "t" acronym is explained to be "temperature" but which kind of temperature : air temperature?, soil temperature?

- (After: 1) a semantic documentation of the different measures included in the dataset using RDF data cube and domain ontologies. Plus, a semantic representation of Distribution structure using csvw vocabulary.

3.4.2 RDA-R1.1-01M . Metadata includes information about the licence under which the data can be reused

- (Before: 1) A link to the license under which the data may be used is available on the landing page.
- (After: 1+) a semantic qualified reference to the used license

3.4.3 RDA-R1.1-02M . Metadata refers to a standard reuse licence

- (Before: 1) the data refers to Etalab license which is a standard license
- (After: 1) No change.

3.4.4 RDA-R1.1-03M . Metadata refers to a machine-understandable reuse licence

- (Before: 0) data refers to a licence in a PDF file which is not a machine-understandable format[4].
- (After: 0)

3.4.5 RDA-R1.2-01M. Metadata includes provenance information according to community-specific standards

- (Before: 0) No explicit provenance metadata.
- (After: 1) semantic metadata using PROV-O properties as a part of GeoDCAT-AP specification

3.4.6 RDA-R1.2-02M . Metadata includes provenance information according to a cross-community language

- (Before: 0) No provenance metadata
- (After: 1) Provenance metadata using PROV-O properties. PROV-O ontology is a W3C recommendation since April 2013.

[4] `https://www.etalab.gouv.fr/wp-content/uploads/2014/05/Licence_Ouverte.pdf`

3.4.7 RDA-R1.3-01M . Metadata complies with a community standard

- (Before: 0) No structured or semantic metadata following a standard schema, unless few natural language metadata.
- (After: 1) Usage of FAIR metadata vocabulary DCAT which is a community standard.

3.4.8 RDA-R1.3-01D Data complies with a community standard

- (Before: 1) SYNOP data are standard data for meteorologists at the international level
- (After: 1) No change.

3.4.9 RDA-R1.3-02M. Metadata is expressed in compliance with a machine-understandable community standard

- (Before: 0) No machine understandable metadata.
- (After: 1) Yes, metadata is expressed with GeoDCAT-AP.

3.4.10 RDA-R1.3-02D. Data is expressed in compliance with a machine-understandable community standard

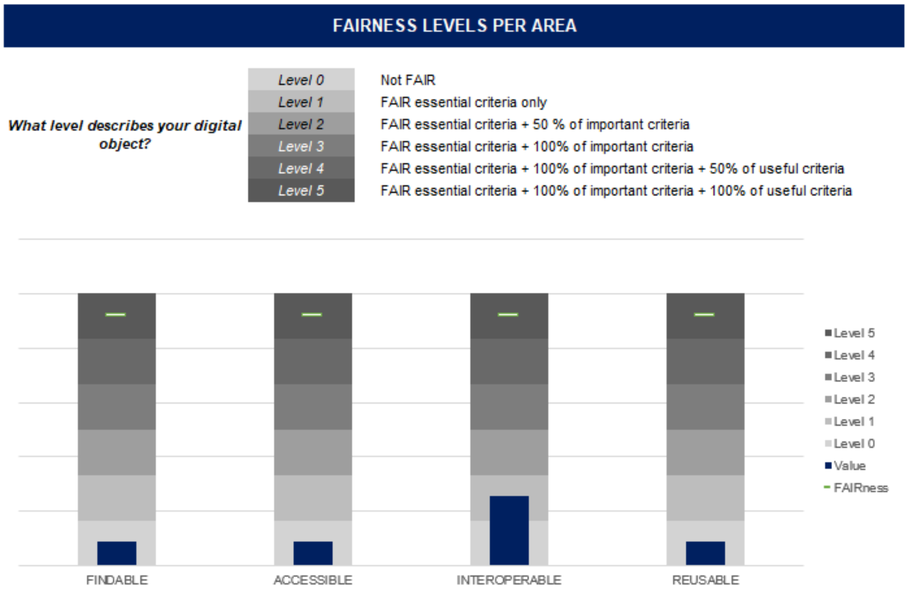
- (Before: 0) No machine understandable format.
- (After: 1) Data is not transformed into RDF, but the semantic model of data is represented using RDF Data Cube and domain ontologies.

4 Evaluation Results

Figure 2[5] shows the summary of the pre-evaluation as suggested by the RDA WG. As we can see, the evaluation gives the level 0 for the three principles 'F', 'A' and 'R' since one or more essential indicators are not satisfied for each of them. 'I' is the single principle that has the level 1 while it doesn't have any satisfied indicator. This is because there is no essential indicator for the I principle. Finally, and based on our evaluation using the FAIR maturity model detailed hereafter, we may conclude that the SYNOP dataset has a very low FAIRness degree.

Figure 3, suggested as a visualisation of the measuring progress method, is more informative. It shows the different levels of maturity per principle. Here, we have only used the two values: *not being considered this yet* (1) and *fully implemented* (4). As can be seen, most accessibility indicators are met, but no interoperability indicator is met.

[5] Figure 2, and Figure 3 are generated with the RDA maturity level tool.



FAIRNESS LEVELS PER AREA

What level describes your digital object?

Level 0	Not FAIR
Level 1	FAIR essential criteria only
Level 2	FAIR essential criteria + 50 % of important criteria
Level 3	FAIR essential criteria + 100% of important criteria
Level 4	FAIR essential criteria + 100% of important criteria + 50% of useful criteria
Level 5	FAIR essential criteria + 100% of important criteria + 100% of useful criteria

Figure 2: SYNOP dataset FAIRness levels per area (pre-evaluation i.e., with Before values)

The SYNOP dataset has been re-evaluated after generating the semantic metadata that describe it . This metadata significantly improves the FAIRness level, especially for the "I" and "R" principles (see Fig. 4 and 5). However, the "F" and "A" principles require the generation of persistent and globally unique identifiers, and the publication of the generated metadata on the web.

The goal of our work is to describe existing datasets with semantic metadata, in order to make them FAIR without making changes to the datasets. However, modifications are necessary to allow the reuse of published data. For example, in the station dataset provided in the SYNOP data documentation, only one location (latitude, longitude, and altitude) is provided for each station. However, in the PDF sheets describing the stations (one sheet per station)[6], we find the history of the stations' locations. These sheets are in PDF format and contain a lot of information, which makes it difficult to use them or to combine them with other data. We therefore recommend that Météo-France improves the station dataset – provided in the documentation – by enriching it with the history of locations so that users can have the precise location of each measurement over time. The history of the instruments used during the measurements is also included in the data sheets, which represents relevant metadata. Another recommendation would be to provide an instrument dataset per station and date interval.

[6] `https://donneespubliques.meteofrance.fr/?fond=contenu&id_contenu=37`

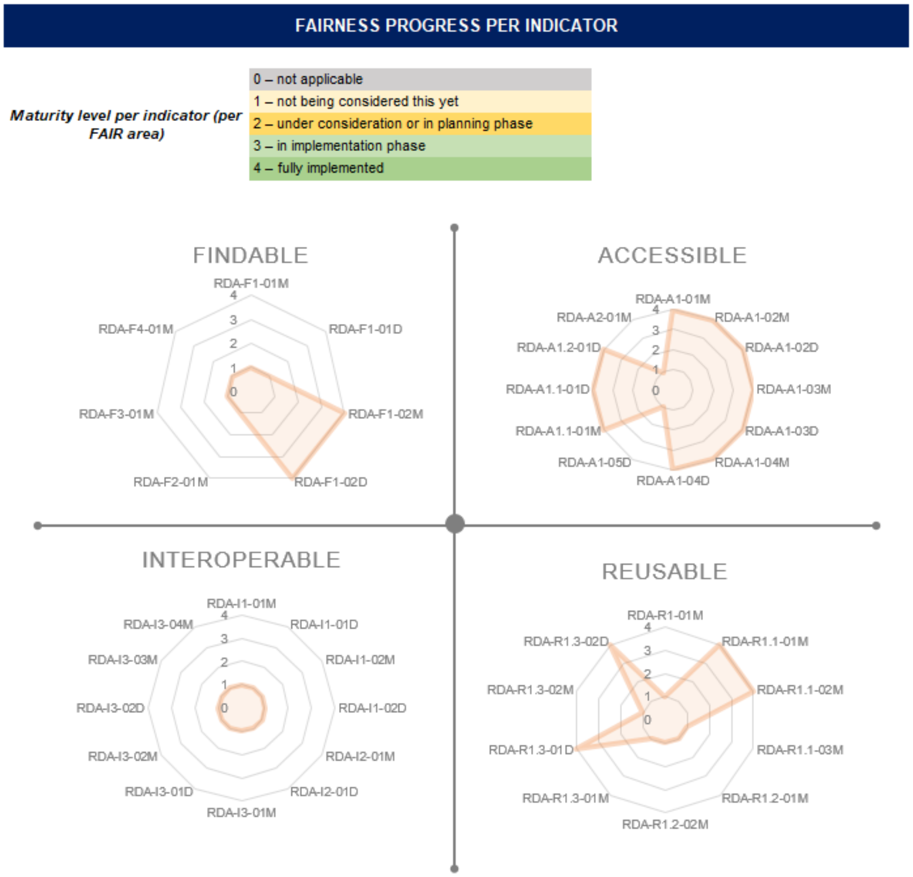
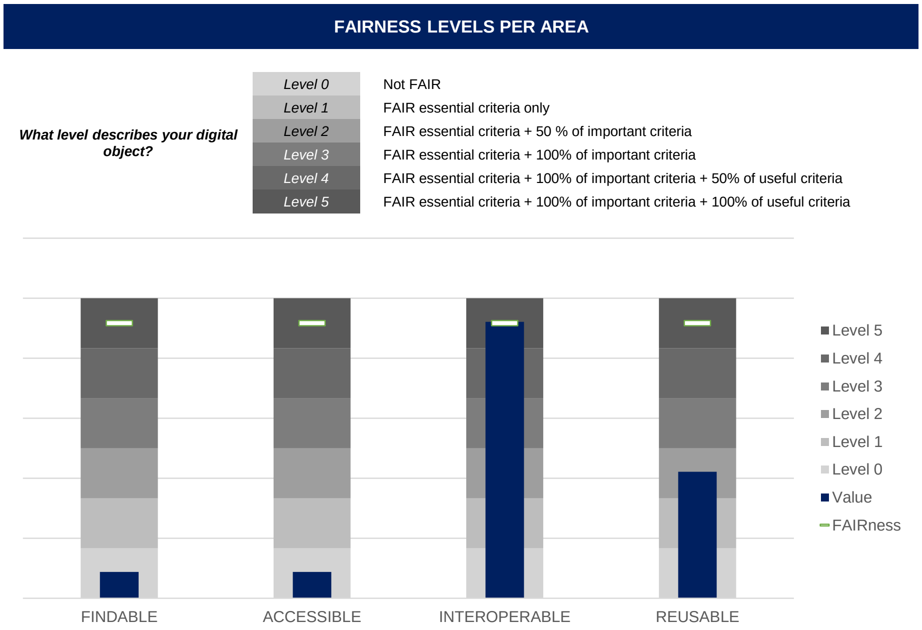


Figure 3: SYNOP dataset FAIRness progress per indicator (pre-evaluation)

References

[RDA, 2020] RDA, F. D. M. M. W. G. (2020). FAIR Data Maturity Model. Specification and Guidelines.

[Wilkinson et al., 2016] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., et al. (2016). The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9.



FAIRNESS LEVELS PER AREA

What level describes your digital object?

Level 0	Not FAIR
Level 1	FAIR essential criteria only
Level 2	FAIR essential criteria + 50 % of important criteria
Level 3	FAIR essential criteria + 100% of important criteria
Level 4	FAIR essential criteria + 100% of important criteria + 50% of useful criteria
Level 5	FAIR essential criteria + 100% of important criteria + 100% of useful criteria

Figure 4: SYNOP dataset FAIRness levels per area (post-evaluation i.e., with After values)

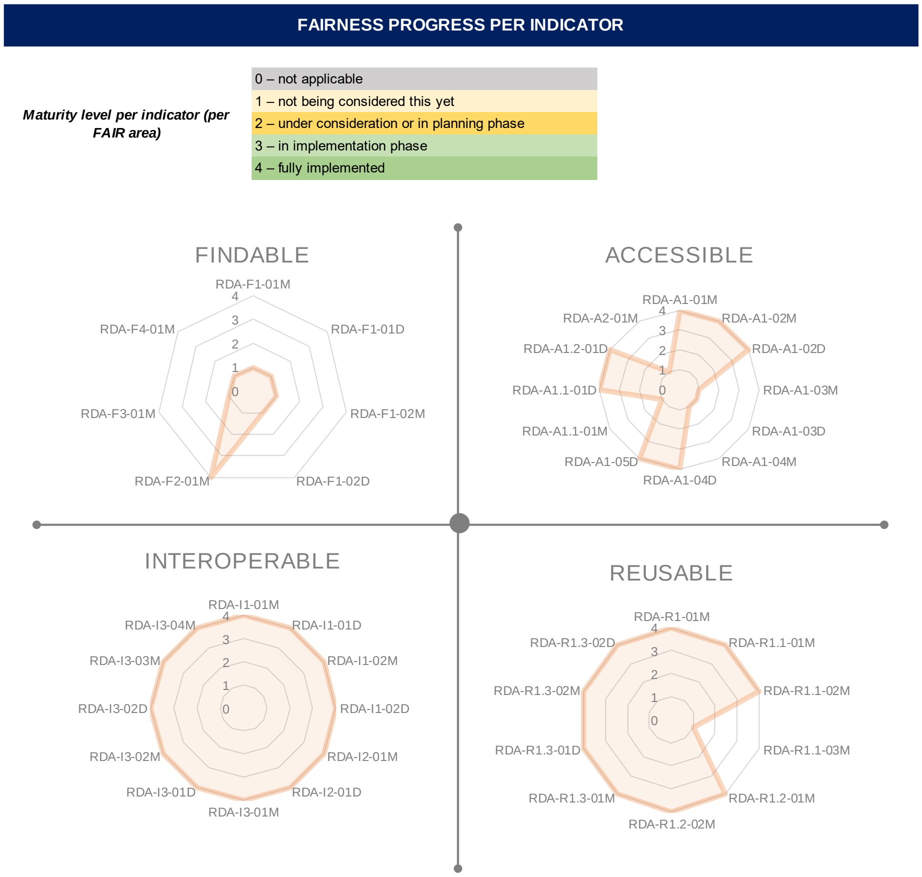


Figure 5: SYNOP dataset FAIRness progress per indicator (post-evaluation)