FAIR data assessment tool specification

What is the purpose of this document?
This document has the purpose of accompanying the FAIR data assessment tool (survey) designed for re-users of datasets to review data based on our operationalised FAIR principles. Following dataset reviews, each dataset which has been reused will have a published review like you would see on a website for product reviews i.e. amazon. In such product reviews, it is normally the case that star ratings are given by users of the product to rate their experience with, and perceived quality of the item. We want to implement this process in the same way for reused datasets.

The specific aim of this document is to provide explanation and examples for reviewers using the assessment tool on the five criteria for each of the three principles: Findable, Accessible and Interoperable. Once a reviewer has rated the dataset to give it a score on each of these principles, an overall or 'Resultant' star score can be calculated to reflect the general 'FAIRness' of the dataset.

Who should use this document/who are we intending to address?
The FAIR data assessment tool addresses people who have recently re-used a dataset from a trusted repository and who now want to provide a review on their experience with the data. Additionally, archive/repository staff who wish to score a dataset are also able to use this document to establish the 'FAIRness' rating of a dataset.

How to read and navigate through this document
The general structure of this document follows the structure of the new operationalisation of the FAIR principles. There are now five criteria under each of the three main principles (FAI) which represent the 5 star rating scales within each principle.

For example, each of the 5 criteria outline the necessary conditions for a dataset to score that number of stars, so if a dataset only meets the conditions under the first criterion for the first principle (Findable) then the dataset will score 1 star for that principle. If a dataset meets the conditions under the first, second and third criteria of the first principle for example, then it will score 3 stars for that principle. The criteria for each principle tend to be cumulative so that a score of 5 stars reflects a dataset meeting all of the conditions in criteria 1-5.

This document simply provides definitions and examples at every level of each criterion in order to avoid confusion or ambiguity, and aims to provide a very clear outline of the necessary conditions which need to be present in order for a dataset to score at the different star levels. Once the reviewer has read through all of these criteria in detail, he/she should have an idea about how the dataset can be scored on the five star scales, which can be confirmed once he/she has completed the FAIR Data Assessment Tool (SurveyMonkey).

The 5 criteria outlined for each principle are displayed in the table below:

Findable (defined by metadata (PID included) and documentation)
1. No PID nor metadata/documentation 2. PID without or with insufficient metadata 3. Sufficient/limited metadata without PID 4. PID with sufficient metadata 5. Extensive metadata and rich additional documentation available
Accessible (defined by presence of user license)
1. Metadata nor data are accessible 2. Metadata are accessible but data is not accessible (no clear terms of reuse in license) 3. User restrictions apply (i.e. privacy, commercial interests, embargo period) 4. Public access (after registration) 5. Open access unrestricted
Interoperable (defined by data format)
1. Proprietary (privately owned), non-open format data 2. Half of the data files are in a proprietary format 3. Non-proprietary, open format = 'preferred format' 4. As well as in the preferred format, data is standardised using a standard vocabulary format (for the research field to which the data pertain) 5. Data additionally linked to other data to provide context

NB: Scale examples for the three principles, where a score of 5 is the best possible score and a score of 1 is the lowest a dataset can score.

Table of Contents

Glossary

Acceptable formats - file formats that are widely used, and can be moderately to reasonably usable, accessible and robust in the long term.

Copyright - this places a set of restrictions and limitations on content, and so the licenses are a means to open up these restrictions and allow licensees the opportunity to use the content in ways defined by the license type or Terms of Use.

Documentation - a description of the dataset which is sufficient enough to know and understand what it is about and what has been done.

Interoperability - describes the extent to which systems and devices can exchange data, and interpret that shared data, without any restriction on access and implementation of the data.

User **license** - a mechanism that defines the permission to reuse data and other material which is protected by copyright or database right.

Linked Data - it should be possible to link one dataset to another, even if they are stored in different places and in different formats, in order to browse semantically similar datasets.

Metadata - can be described as data about the data, and specifically, the formal description of the resource, i.e. title, date, author, description, location.

Ontology - the formal name or definition of the entities that exist for a particular domain of discourse.

PID ('persistent identifier') - provides a permanent citable reference to a certain dataset, just like an ISBN number does for a certain book. A PID always has a fixed reference to the dataset, no matter where it is located online.

Preferred formats - file formats of which DANS is confident will offer the best long-term guarantees in terms of usability, accessibility and sustainability.

Proprietary format - a file format belonging to a company, organisation or individual which is stored according to a particular encoding-scheme, designed by the company or organisation.

Resource Description Framework (RDF) - a standard model for data interchange on the Web. It is based upon the idea of making statements about resources in the form of subject–predicate–object expressions: The subject denotes the resource (i.e. represented by URI), and the predicate denotes traits or aspects of the resource and expresses a relationship between the subject and the object.

URI ('uniform resource identifier') - a sequence of characters which identifies an abstract or physical resource online. You may also recognise the term 'URL' which refers to a subset of URIs which provide the resource's network location.

Vocabulary standardisation - the use of the same set of codes and vocabulary to encode and describe data throughout a system.

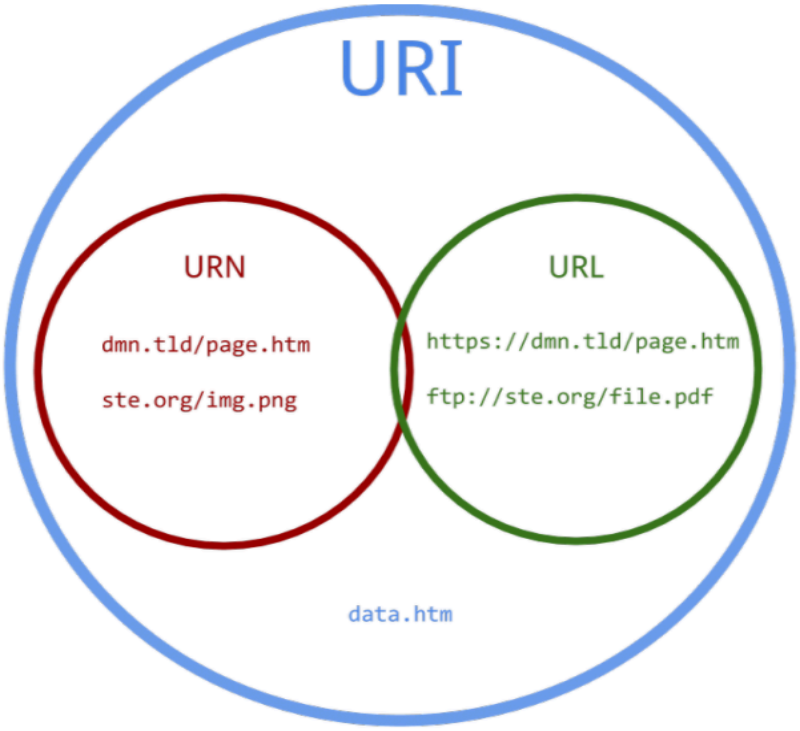
Findable

F1. No PID **AND** no metadata/documentation

1a. Does the dataset have a PID (persistent identifier)?

Explanation:

- A PID or 'persistent identifier' provides a permanent citable reference to a certain dataset, just like an ISBN number does for a certain book. A PID always has a fixed reference to the dataset, no matter where it is located online.[1] [2]
    - For more information and examples on this, you can refer to the links in footnotes 1 and 2.
- Additionally, a URI or 'uniform resource identifier', is a sequence of characters which identifies an abstract or physical resource online. You may also recognise the term 'URL' which refers to a subset of URIs which provide the resource's network location.[3]
    - The image below is taken from footnote 3 to demonstrate this.



    - You can also refer to footnote 3 for more information on this.
- Note! For the criterion of F1, it is specified that there is NO URI or PID.

1b. Is there any metadata or documentation available?

[1] See: https://www.surf.nl/en/services-and-products/data-persistent-identifier/index.html and http://www.rsp.ac.uk/documents/briefing-papers/technical-persistentids.pdf

[2] http://www.persistent-identifier.de/english/204-examples.php

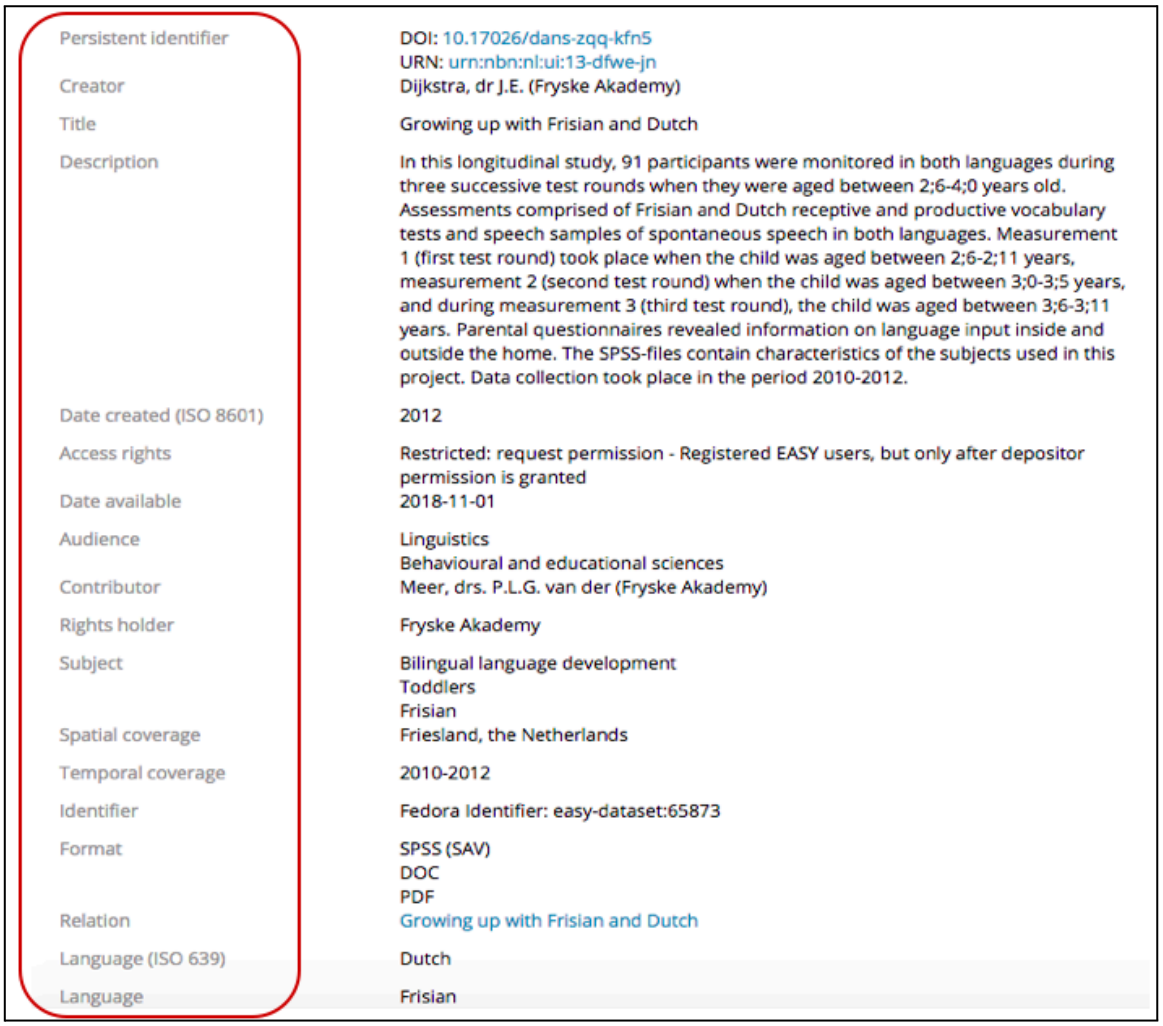
[3] https://danielmiessler.com/study/url-uri/#gs.=4FFgLw

Explanation of metadata:

- Metadata can be described as 'data about the data', and specifically, we refer to this term as meaning the formal structured description of the resource.[4]
- An example of some simple book metadata can be seen in the table below[5]:

Document type	Book
Last name of author	Bloomfield
First name of auther	Leonard
Year of publication	1933
Title	Language
City	New York
Publisher	Holt, Rinehart & Winston

- The following example is of the metadata fields collected for datasets deposited in the EASY repository at DANS[6]:



Persistent identifier	DOI: 10.17026/dans-zqq-kfn5 URN: urn:nbn:nl:ui:13-dfwe-jn
Creator	Dijkstra, dr J.E. (Fryske Akademy)
Title	Growing up with Frisian and Dutch
Description	In this longitudinal study, 91 participants were monitored in both languages during three successive test rounds when they were aged between 2;6-4;0 years old. Assessments comprised of Frisian and Dutch receptive and productive vocabulary tests and speech samples of spontaneous speech in both languages. Measurement 1 (first test round) took place when the child was aged between 2;6-2;11 years, measurement 2 (second test round) when the child was aged between 3;0-3;5 years, and during measurement 3 (third test round), the child was aged between 3;6-3;11 years. Parental questionnaires revealed information on language input inside and outside the home. The SPSS-files contain characteristics of the subjects used in this project. Data collection took place in the period 2010-2012.
Date created (ISO 8601)	2012
Access rights	Restricted: request permission - Registered EASY users, but only after depositor permission is granted
Date available	2018-11-01
Audience	Linguistics Behavioural and educational sciences
Contributor	Meer, drs. P.L.G. van der (Fryske Akademy)
Rights holder	Fryske Akademy
Subject	Bilingual language development Toddlers Frisian
Spatial coverage	Friesland, the Netherlands
Temporal coverage	2010-2012
Identifier	Fedora Identifier: easy-dataset:65873
Format	SPSS (SAV) DOC PDF
Relation	Growing up with Frisian and Dutch
Language (ISO 639)	Dutch
Language	Frisian

[4] http://www.niso.org/publications/press/UnderstandingMetadata.pdf
[5] http://www.language-archives.org/documents/gentle-intro.html
[6] https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:65873/tab/1; DOI: 10.17026/dans-zqq-kfn5 URN: urn:nbn:nl:ui:13-dfwe-jn

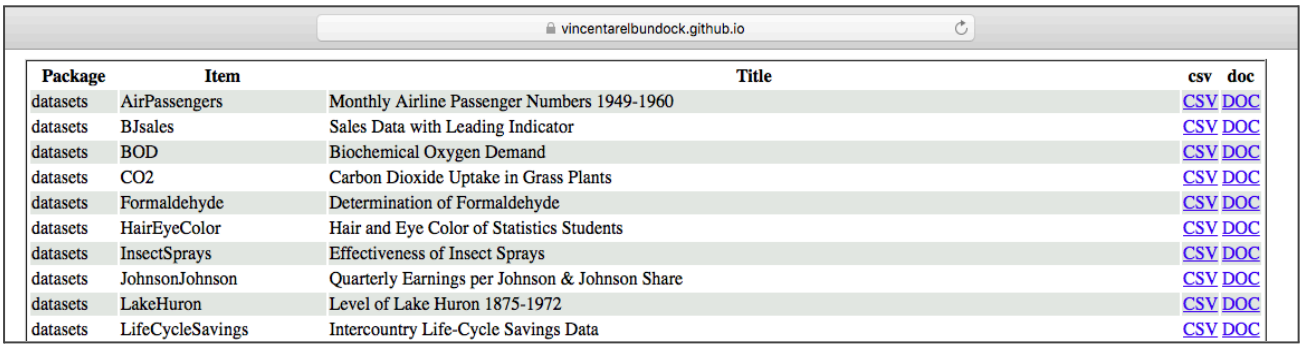
- Although the metadata fields required can vary between different repository standards, there is generally a similar structure to all metadata required i.e. formal and structured fields to describe different aspects of the data, as seen in the examples above.

Explanation of documentation:

- We use the term data documentation to mean a description of the dataset which is sufficient enough to know and understand what it is about, and which does not hold a formal metadata structure.
- This could come in the form of additional code books or articles with extra information about the data which might describe details which aren't obvious at first glance i.e. context of methods of collection, descriptions of variables, algorithms used to transform the data etc.

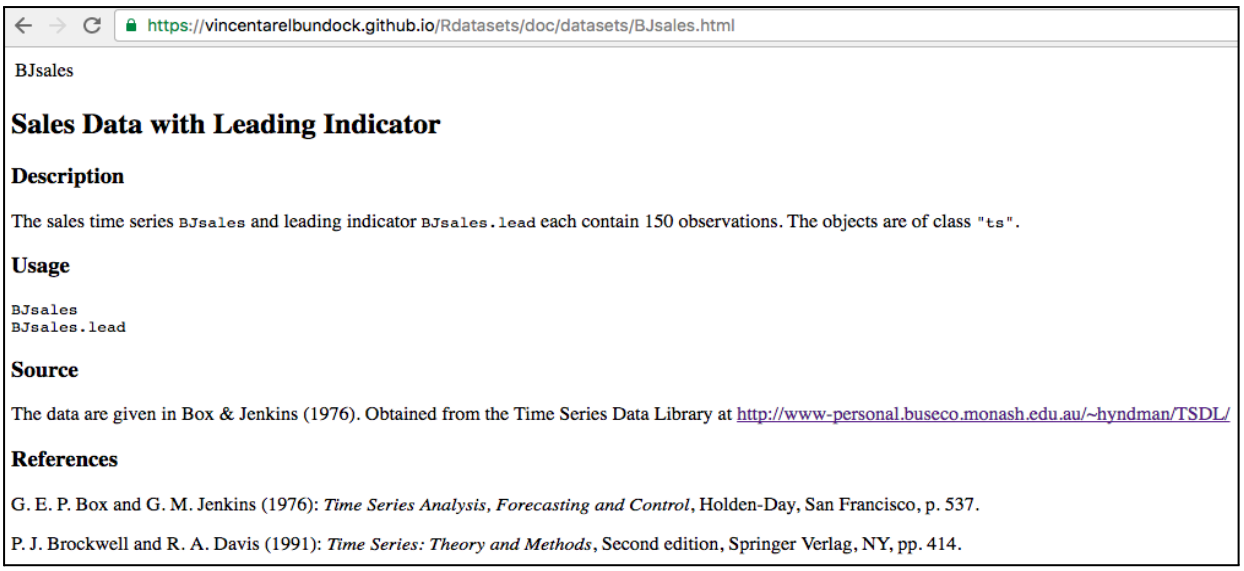
Examples:

- A dataset published outside of a repository may be posted for example on a webpage, blog or online forum without any further information about what the dataset represents.
- For example, you can click this link which takes you to a blog page containing some datasets for a programming tutorial: https://vincentarelbundock.github.io/Rdatasets/datasets.html (preview below)



Package	Item	Title	csv	doc
datasets	AirPassengers	Monthly Airline Passenger Numbers 1949-1960	CSV	DOC
datasets	BJsales	Sales Data with Leading Indicator	CSV	DOC
datasets	BOD	Biochemical Oxygen Demand	CSV	DOC
datasets	CO2	Carbon Dioxide Uptake in Grass Plants	CSV	DOC
datasets	Formaldehyde	Determination of Formaldehyde	CSV	DOC
datasets	HairEyeColor	Hair and Eye Color of Statistics Students	CSV	DOC
datasets	InsectSprays	Effectiveness of Insect Sprays	CSV	DOC
datasets	JohnsonJohnson	Quarterly Earnings per Johnson & Johnson Share	CSV	DOC
datasets	LakeHuron	Level of Lake Huron 1875-1972	CSV	DOC
datasets	LifeCycleSavings	Intercountry Life-Cycle Savings Data	CSV	DOC

- On the right hand side there is access to 'doc' which contains the documentation for the dataset.
- When opening the dataset 'Sales Data with Leading Indicator' (second row), the following 'documentation' is displayed:



https://vincentarelbundock.github.io/Rdatasets/doc/datasets/BJsales.html

BJsales

Sales Data with Leading Indicator

Description

The sales time series `BJsales` and leading indicator `BJsales.lead` each contain 150 observations. The objects are of class `"ts"`.

Usage

```
BJsales
BJsales.lead
```

Source

The data are given in Box & Jenkins (1976). Obtained from the Time Series Data Library at http://www-personal.buseco.monash.edu.au/~hyndman/TSDL/

References

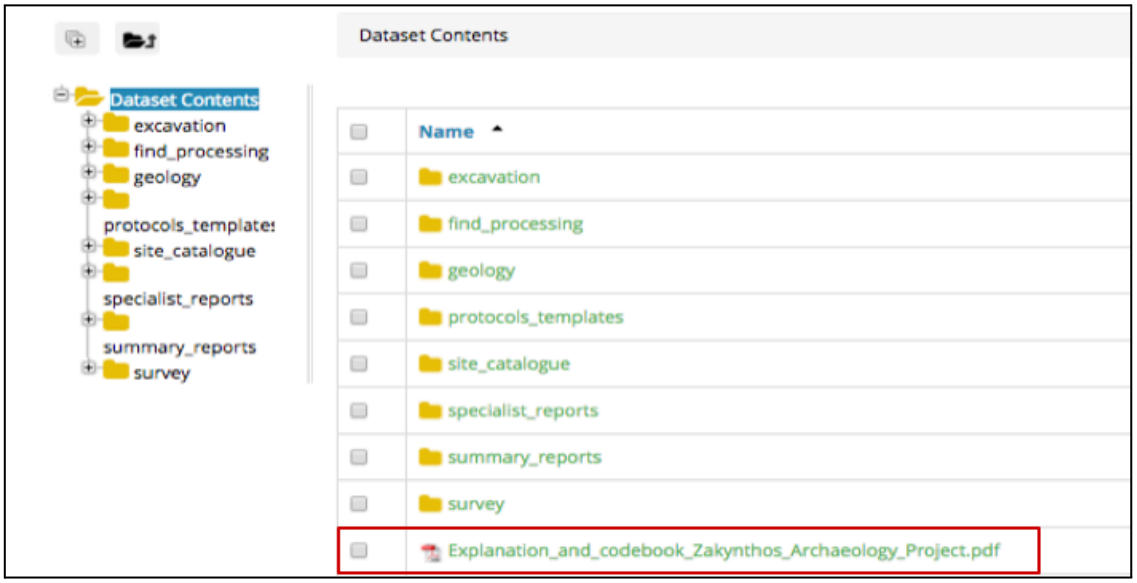
G. E. P. Box and G. M. Jenkins (1976): *Time Series Analysis, Forecasting and Control*, Holden-Day, San Francisco, p. 537.

P. J. Brockwell and R. A. Davis (1991): *Time Series: Theory and Methods*, Second edition, Springer Verlag, NY, pp. 414.

- This documentation is very lacking in detail and description. There is almost no description that can be understood, or which explains what the dataset is about. Also, the additional link provided refers you to a page which is not currently active.
- Thus, it would be considered that the dataset via this website has very limited documentation which does not tell us anything about the dataset represents.

Example 2:

- An alternative example taken from the EASY archive at DANS shows a dataset with an explanation and code book document[7]:



[7] https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:64592/tab/2; DOI: 10.17026/dans-27f-zd2u URN: urn:nbn:nl:ui:13-g6wg-il

Explanation and Codebook
Zakynthos Archaeology Project

Preparation of dataset in 2016 by I. (Ivan) Kisjes MA (University of Amsterdam) and I.F. (Ilona) von Stein MA (University of Amsterdam and DANS)

Content Explanation and Codebook:
- Explanation dataset (e.g. participants, research methods, methodology)
- Folder structure
- Photos / photolists
- Abbreviations
- Databases

- In this example the accompanying documentation to the dataset is very descriptive and gives explanation for additional aspects not covered by the metadata (as can be seen in the content described above)
- This is an example of documentation which is sufficient for someone not familiar with the data to understand what it is about.
- Note! For the criterion of F = 1 it is specified that there is NO documentation or metadata, which means that there is no descriptive information about the dataset, or is insufficient in enabling understanding of what the data represents.

Scoring the dataset:

1. Now you can think about how you would rate the dataset based on the criteria: 'URI/PID' and 'Data Documentation'
2. Make sure you have read the criteria for F1 carefully
3. Decide whether the dataset qualifies for a score of 1 for the principle of 'Findability'
4. If the dataset doesn't meet the criteria of a score of 1 (i.e. does have a PID or metadata/documentation), then you can proceed onto the next set of criteria (refer to the routing options below for specific instructions)

Remember: the AND is important!

Routing
IF:
1a = no AND 1b = no ⇒ F = 1
1a = no AND 1b = yes ⇒ F = 3 (go to F3 to check)
1a = yes AND 1b = no ⇒ go to F2
1a = yes AND 1b = yes ⇒ go to F2

F2. PID present but without metadata/insufficient metadata

2a. Does the dataset have a URI or PID (persistent identifier)?

Explanation: See description in 1a.

- Note! For the criterion of F2 it is specified that there IS a URI or PID available for the dataset.

2b. Does the dataset have insufficient or no metadata?

Explanation of insufficient metadata criteria:

- See explanation of metadata in 1b.
- Insufficient metadata is defined in the online archive of DANS (EASY)[8], as not having all of the 4 minimum details required:
    - Creator name
    - Title
    - Description (1 to 2 paragraphs describing the data to someone who has never worked with this data before)
    - Intended audience/discipline
- In application of other repositories outside of EASY, it is also possible to follow the rule that if at least 25% of the metadata fields of a repository are not filled in then the metadata is not sufficient
    - In EASY there are around 16 metadata fields (see diagram above) which can be filled in before depositing, of which 25% (4 criteria) are necessary for sufficient (but limited) metadata

Example of INSUFFICIENT metadata:

- Here you can see an example of metadata which would be insufficient[9]:

[8] https://easy.dans.knaw.nl/ui/home, see also https://blog.datacite.org/datacite-mandatory-properties/ for information on mandatory metadata fields.

[9] https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:23075/tab/1; DOI: 10.17026/dans-xry-9z5w, URN: urn:nbn:nl:ui:13-nvh-24h
Additional examples of human viewable metadata descriptions of a dataset can be found at: https://www.w3.org/TR/dwbp/dwbp-example.html

Persistent identifier	DOI: 10.17026/dans-xry-9z5w URN: urn:nbn:nl:ui:13-nvh-24h
Archis Onderzoeksmeldingsnr.	17424
Title	Archeologisch onderzoek Golfodrome te Amstelveen
Alternative title	Bureauonderzoek GAR 287
Creator	Soetens, L.; Osinga, M.
Date created (ISO 8601)	2007-12-18
Rights holder	Grontmij
Publisher	Grontmij
Description	onderzoeksrapport
Spatial point	RD (in m.) X: 120225 Y: 478375
Spatial coverage	25D/25G Golfodrome Amstelveen
Identifier	208862 (Grontmij) Archis onderzoek: 17424 ISSN: 1573-5710 eDNA-project: a03186 AipId: twips.dans.knaw.nl-2347791999001429903-1264014492984 Archis onderzoeksmeldingsnr. 17424 Fedora Identifier: easy-dataset:23075
Relation	is part of • *Grontmij Archeologische Rapporten*
Type (DCMI resource type)	Text
Language (ISO 639)	Dutch
Date submitted	2010-01-20
Access rights	Restricted: archaeology group - Access restricted to registered 'archaeology' group members
Date available	2010-01-19

- As you can see, the title, PIDs and keywords allow for some identification of the type of study and discipline, however, the description is uninformative and does not explain anything which can be used to understand what the dataset is about, hindering its findability.
    - Metadata fields have also been used in an unstandardised manner, where semicolons have been used to separate entities instead of using the repeatable fields option for multiple rows.
    - A province has not been provided in the Spatial details which makes the dataset also harder to find.
    - There is no audience/discipline

Note! For the criterion of F = 2 it is specified that there is NO or INSUFFICIENT metadata accompanying the dataset.

- Reminder: if the metadata does not include a minimum of the 4 descriptions above, or does not have at least 25% of the metadata fields filled in properly ⇒ it is NOT sufficient
- Now you can use the routing options below to decide whether your dataset qualifies for a score of 2 for 'Findability'

Routing
IF:
2a = yes AND 2b = yes ⇒ F = 2 2a = no AND 2b = no ⇒ F = 3
2a = yes AND 2b = no ⇒ go to F4

NB: An answer of 'yes' to 2b would mean that there is insufficient or no metadata. When there is insufficient metadata, then the data can not score higher than F2.

F3. Sufficient metadata without PID

3a. Does the dataset have a URI or PID (persistent identifier)?

Explanation: See description in 1a.

- Note! For the criterion of F3 it is specified that there is NO URI or PID.

3b. Is there just sufficient metadata available?

Explanation of sufficient metadata available:

- Following from the guidelines for insufficient metadata in section F2, metadata should be sufficient when it includes at least the 4 main details:
  - Creator name, title, description (1 to 2 paragraphs describing the data to a non-specialist) and intended audience/discipline
- Additionally, sufficient metadata can also include the following 'recommended elements' as outlined in EASY's metadata criteria (as can also be seen in the sample metadata from EASY in section F2):
  - Contributor/publisher
  - Rights holder
  - Subject
  - Spatial coverage
  - Temporal coverage
  - Type of data
  - Clear version numbers
  - Provenance (how it was collected)
  - Keywords provided

- In application of other repositories, it is also possible to follow the rule that if 25% - 75% of the metadata fields of a repository are filled in then the metadata is sufficient
- It is important to understand that ALL sufficient metadata include enough information for someone who has not had experience with the dataset before to understand what the dataset is about.
  - Discipline-dependence: It may be that the description of the data is understandable for specialists only, but sufficient metadata always needs to be understandable for non-specialists as well.

Examples of SUFFICIENT metadata:

- This example displays the metadata which is sufficient to describe what the dataset is about, as well as providing a number of additional details to aid its findability[10]:

Creator	Meroño Peñuela, Dr A. (Vrije Universiteit Amsterdam) Ashkpour, A. (Internationaal Instituut voor Sociale Geschiedenis) Guéret, Dr C. (BBC)
Title	CEDAR RDF database, the Dutch historical censuses (1795-1971) as Linked Open Data
Description	This are the results of the CEDAR project. The deposit contains, among other relevant information, the final database of the harmonized Dutch historical censuses, encoded using the Resource Description Framework (RDF), a suitable format for linking data on the Web. The deposit contains two different versions: a complete RDF conversion only partially harmonized; and partial RDF conversion, completely harmonized.
Date created (ISO 8601)	2016-04-30
Access rights	Open access - Unrestricted access (CC0 Waiver No Rights Reserved - https://creativecommons.org/about/cc0)
Date available	2016-06-23
Audience	Humanities History Social sciences Computer science
Contributor	Valk, Dr R. de (DANS) Scharnhorst, Dr A. (DANS) Hoekstra, Dr R. (Vrije Universiteit Amsterdam) Schlobach, Dr S. (Vrije Universiteit Amsterdam) Mandemakers, Prof.Dr. K. (Internationaal Instituut voor Sociale Geschiedenis) Harmelen, Prof.Dr. F. van (Vrije Universiteit Amsterdam)
Subject	RDF, Linked Data, semantic web Historische Volkstellingen, Census Data historical censuses, dutch history
Spatial coverage	Nederland (alle gemeentes) The Netherlands (all municipalities)
Temporal coverage	1795-1971
Source	The Dutch historical censuses, 1795-1971 http://volkstellingen.nl/
Identifier	Fedora Identifier: easy-dataset:65225

- This example is sufficient because it includes detail of all of the essential details as well as some recommended elements (as mentioned above), entered in the standardised way which makes the dataset highly findable.

Note! For the criterion of F = 3 it is specified that there is SUFFICIENT metadata accompanying the dataset.

- Reminder: if the metadata does include a minimum of the 4 descriptions outlined, or has at least 25% of the metadata fields filled in ⇒ it is sufficient

[10] https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:65225/tab/1; DOI: 10.17026/dans-xpk-wj5w
URN: urn:nbn:nl:ui:13-c5ac-jm
Specific metadata standards can be found here:
http://rd-alliance.github.io/metadata-directory/standards/

- NB: If you think the metadata exceeds the standard of sufficient metadata described here, please go on to read the criteria for a score of F5 which describes a higher level of rich and extensive metadata.
- Now you can use the routing options below to decide whether your dataset qualifies for a score of 3 for 'Findability'.

Routing IF: 3a = no AND 3b = yes ⇒ F = 3 3a = yes AND 3b = yes ⇒ Go to F4

NB: Any dataset without a PID cannot score higher than F = 3.

F4. PID present with sufficient metadata

4a. Does the dataset have a URI or PID (persistent identifier)?

Explanation: See description in 1a.

- Note! For the criterion of F4 it is specified that there IS a URI or PID present.

4b. Is there just sufficient metadata available?

Explanation: See description for sufficient metadata in 3b.
Note! For the criterion of F = 4 it is specified that there is SUFFICIENT metadata accompanying the dataset.

- Reminder: if the metadata does include a minimum of the 4 descriptions outlined as well as some of the recommended elements, or has at least 25% of the metadata fields filled in ⇒ it is sufficient
- Now you can use the routing options below to decide whether your dataset qualifies for a score of 4 for 'Findability'.

Routing IF: 4a = yes AND 4b = yes ⇒ F = 4 4a = yes AND 4b = metadata is more rich than only just sufficient ⇒ go to F5

F5. Extensive metadata and rich additional documentation available

5a. Is there extensive metadata and rich additional documentation for the dataset available?

Explanation for extensive metadata:

- Extensive metadata includes all the criteria for 'sufficient' metadata (including all of the recommended elements) but has additional information which goes beyond sufficient and really gives a detailed description of the dataset

- Metadata must be machine and human understandable for autonomicity
- Some of the additional elements that should be included for a score of F5 are:
    - Language of dataset
    - File format
    - Relation or linked to other datasets based on
    - Pakbon (applicable to archeology datasets)

Explanation of rich additional documentation:

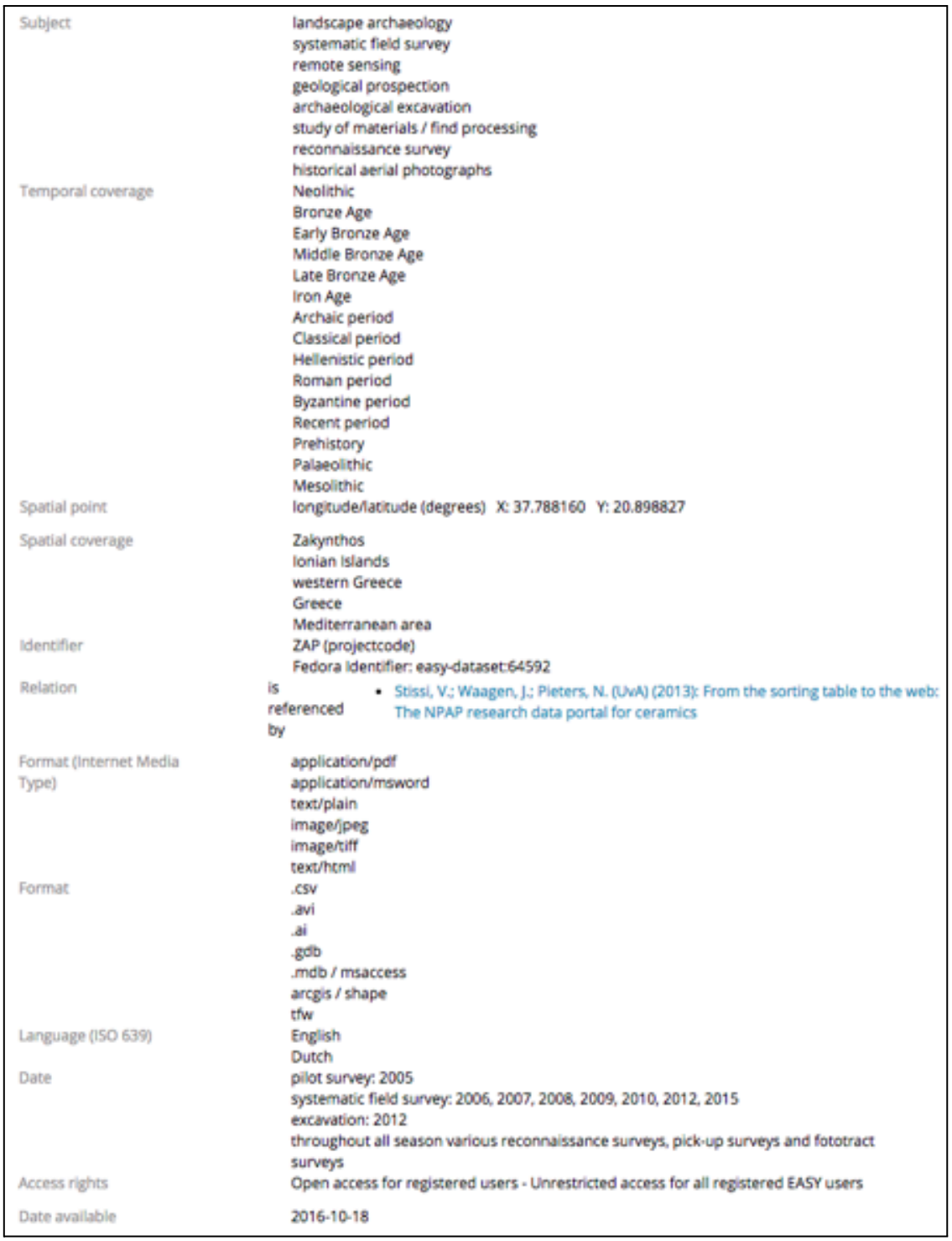
- Information about how the data was used and explanation about its reuse possibilities for future use
- Information about precision of measurement

Examples of extensive and rich metadata/documentation:

- This example demonstrates very high quality metadata descriptions and provides a lot of additional and very detailed descriptions under almost all of the available fields[11]:

[11] https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:64592/tab/1; DOI: 10.17026/dans-27f-zd2u URN: urn:nbn:nl:ui:13-g6wg-il

Persistent identifier	DOI: 10.17026/dans-27f-zd2u URN: urn:nbn:nl:ui:13-g6wg-il
Title	The Zakynthos Archaeology Project
Creator	Wijngaarden, dr. G.J.M. van (University of Amsterdam, The Netherlands) DAI: info:eu-repo/dai/nl/110236629
Contributor	Merkouri, dr. C. (Ephorate of Antiquities Zakynthos, Greece) Stissi, prof. dr. V.V. (University of Amsterdam, The Netherlands) DAI: info:eu-repo/dai/nl/158748204 Kourtesi-Philippakis, prof. dr. G. (National and Kapodistrian University of Athens, The Netherlands) Kars, prof. dr. H. (Vrije Universiteit Amsterdam, The Netherlands) DAI: info:eu-repo/dai/nl/070504601 Zananiri, dr. I. (Institute of Geology and Mineral Exploration, Greece) Avramidis, P. (Technological Educational Institute of Mesolonghi (TEI), Greece) Pieters, N. (University of Amsterdam, The Netherlands)
Date created (ISO 8601)	2015-06-30
Rights holder	University of Amsterdam Ephorate of Antiquities at Zakynthos
Description	The island of Zakynthos (Greece) has a rich and varied history, but is notoriously lacking in archaeological remains. This "scarcity of archaeology" is primarily due to the extensive destruction of the ancient landscapes by earthquakes and intensive agriculture and building. Starting point for this project is our conviction that we need to be better informed about archaeologically marginal areas such as Zakynthos in order to understand the ancient world. The Zakynthos Archaeology Project (2005-2015) is an interdisciplinary research project with the aim to relate the distribution of archaeological finds to the dynamic landscape of the island. To achieve this, scholars and students from various Dutch and Greek universities carry out a range of researches: intensive archaeological survey, remote sensing, geological prospection, limited archaeological excavations (test trenches), study of materials and reconnaissance surveys. The project is a cooperation between the Netherlands Institute in Athens and the Ephorate for Antiquities at Zakynthos. It is directed by Dr Gert Jan van Wijngaarden (University of Amsterdam) and by Christina Merkouri (Ephorate for Antiquities at Zakynthos). This dataset for EASY of DANS consists of all primary data of the archaeological survey, excavation and find processing collected between 2005 and 2015. A wide variety of research data types are included, such as field photos and drawings, find photos and drawings, GIS-files, databases, reports, etc. The dataset is prepared for archiving by I. (Ivan) Kisjes MA and I.F. (Ilona) von Stein MA.



Subject	landscape archaeology systematic field survey remote sensing geological prospection archaeological excavation study of materials / find processing reconnaissance survey historical aerial photographs
Temporal coverage	Neolithic Bronze Age Early Bronze Age Middle Bronze Age Late Bronze Age Iron Age Archaic period Classical period Hellenistic period Roman period Byzantine period Recent period Prehistory Palaeolithic Mesolithic
Spatial point	longitude/latitude (degrees) X: 37.788160 Y: 20.898827
Spatial coverage	Zakynthos Ionian Islands western Greece Greece Mediterranean area
Identifier	ZAP (projectcode) Fedora identifier: easy-dataset:64592
Relation is referenced by	• Stissi, V.; Waagen, J.; Pieters, N. (UvA) (2013): From the sorting table to the web: The NPAP research data portal for ceramics
Format (Internet Media Type)	application/pdf application/msword text/plain image/jpeg image/tiff text/html
Format	.csv .avi .ai .gdb .mdb / msaccess arcgis / shape tfw
Language (ISO 639)	English Dutch
Date	pilot survey: 2005 systematic field survey: 2006, 2007, 2008, 2009, 2010, 2012, 2015 excavation: 2012 throughout all season various reconnaissance surveys, pick-up surveys and fototract surveys
Access rights	Open access for registered users - Unrestricted access for all registered EASY users
Date available	2016-10-18

- Additionally, this dataset also has available an ‘Explanation and codebook’ document for the project under the folder ‘Data Files’ in EASY. This provides detailed information about how the data is coded and further explanation about the methodology, origin and structure of the dataset.
- Each individual data file also has additional ‘details’ which can be accessed via the ‘view details’ button in the top-left area when viewing the files on EASY. This provides even more detail on an individual file level.

Note! For the criterion of F = 5 it is specified that there is rich documentation AND extensive metadata which goes well beyond the descriptions of 'sufficient' metadata accompanying the dataset.

- It is possible to apply the rule to this criterion, in which metadata including above 75% of the repository fields and are all filled in accurately, as well as fulfilling the additional/extensive criteria mentioned above, will gain a findability score of 5 here.
- Note: If there is additional documentation i.e. codebook, but the metadata is not extensive then the dataset will score F4, because both documentation and metadata should be rich for a score of F5.
- Now you can use the routing options below to decide whether your dataset qualifies for a score of 5 for the principle 'Findable'.

Routing IF: 5a = yes ⇒ F = 5 For answers 'no' to either of these criteria, go back to the previous steps.

Now you should have a score for the dataset which reflects its quality based on the principle 'Findable'. Continue onto the next page to rate the data based on how 'Accessible' it is.

NB: keep your score in mind to compare to the score which you are given on the FAIR data assessment tool. If these two don't match then something has gone wrong, perhaps try to understand why there could be a mismatch and complete the tool again.

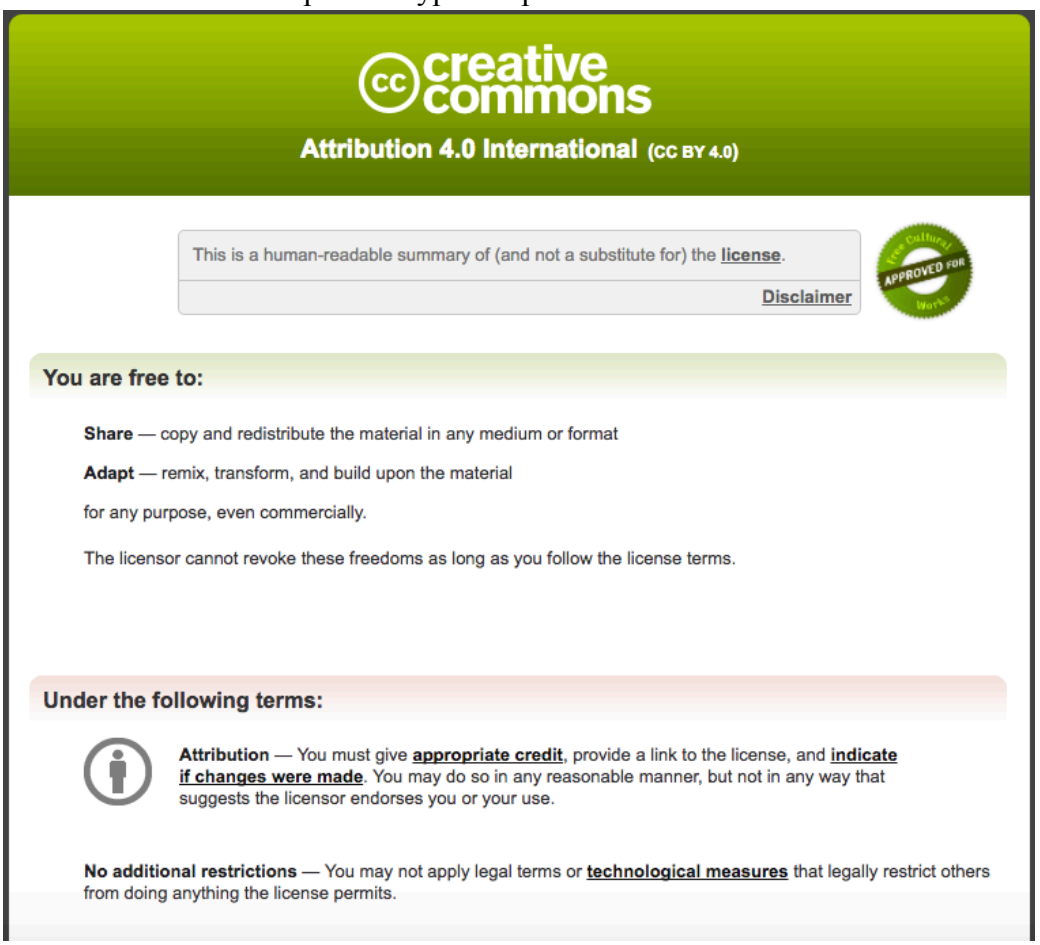
Accessible

A1. Metadata nor data is accessible

_____1a. Does the dataset have a user license?

Explanation of user licenses:

- A user licence is a mechanism that explains the extent to which people and organisations have permission to reuse the dataset and other material which is protected by copyright or database right.[12]
    - Copyright places a set of restrictions and limitations on content, and so the licenses are a means to open up these restrictions and allow licensees the opportunity to use the content in ways defined by the license type or Terms of Use.[13]
- Therefore, it is appropriate that a licence should provide clarity on what users and re-users are permitted to do with the dataset, and outline any restrictions on the extent of that permission.
- Below is a sample of a type of open user license from 'Creative Commons':[14]



12 http://www.nationalarchives.gov.uk/documents/information-management/uk-government-licensing-framework.pdf

13 http://vertnet.org/resources/datalicensingguide.html

14 https://creativecommons.org/licenses/by/4.0/

CC-BY can be considered the most liberal Creative Commons license since it allows for unrestricted reuse of the creative work, subject only to the requirement that the work is appropriately attributed. The full version of this license can be viewed in the footnote below, including the full terms of use[15].

NB: If a dataset has no user license then this makes the data automatically inaccessible. Furthermore, if there is unclarity about the terms in which the data can be used then it is also deemed to be inaccessible.

Note! For the criterion of A1 it is specified that there is NO user license available for the dataset which means that the data is not available for re-use.

1b. Are the metadata accessible?

Explanation:

- See (F)1b for explanation of what metadata is.
- Metadata is accessible if it is available on the web, even if the accompanying dataset is no longer accessible

Note! For the criterion of A=1 it is specified that metadata is NOT available

- NB: Any dataset without accessible metadata cannot score higher than a score of A1 for this principle 'Accessible'.

Routing IF: 1a = no AND 1b = no ⇒ A = 1 1a = no AND 1b = yes ⇒ A = 2 1a = yes AND 1b = no ⇒ go to A3 1a = yes AND 1b = yes ⇒ go to A3

A2. Metadata are accessible but data is not accessible (anymore)

2a. Are the metadata accessible?

Explanation

- See description in (A)1b
- NB: If the data is no longer available e.g. it has been erased for an important reason (i.e. medical), the metadata (also referred to as tombstone in this case) should still be available to describe the original dataset in the standardised way also explaining why the data was erased.

Note! For the criterion of A2 it is specified that there IS metadata available which accompanies a recorded dataset, even if this actual data is no longer available.

2b. Does the dataset have a user license? Are there clear conditions of reuse?

Explanation

[15] https://creativecommons.org/licenses/by/4.0/legalcode

- See (A)1a for explanation and an example of user licenses for accessibility and reuse of datasets.
- If it is unclear to the re-user what is permitted to be done with the dataset and/or what the restrictions are stated under a license, then the dataset can be seen as inaccessible for reuse legally.

Note! For the criterion of A = 2 it is specified that there is either UNCLEAR TERMS (insufficient to understand reuse conditions) or NO user license available for the dataset which means the data is not available for re-use.

Routing IF: 2a = yes AND 2b = no ⇒ A = 2 2a = yes AND 2b = yes ⇒ go to A3

A3. User restrictions apply

3a. Does the license have user restrictions? (no public access)

Explanation of license restrictions

- See (A)1a for explanation of user licenses and an example.
- If the user license has some restrictions on its access then it can only gain a maximum score of 3 for 'Accessible'.

Examples of restrictions:

- Privacy settings - ethical issues restrict access
  - NB: when there are restrictions of access to data because of privacy issues, metadata should be explicit and detailed in describing these cases including why these restrictions occur; see section (A)2a.
- Software tools needed to access data
- Commercial interests (paying a journal for access)
- Embargo periods (access for only 24 hours etc.)
- Specific group access only login (only people belonging to a certain group can log in, i.e. academics, companies, data organisations)

Example of Creative Commons license[16] which grants public access but under some user restrictions for reuse and redistribution:
- No derivatives (nd)
  - Others can copy, distribute, display and perform only original copies of the work. If re-users want to modify the work then they must get the owner's permission first.

Note! It is important to remember that restrictions to public access of any kind qualify for this category of scoring, even if they are ones not mentioned in the examples section. For the criterion of A3 it is specified that there IS a user license but which describes some restriction to the access of the data.

[16] https://creativecommons.org/share-your-work/licensing-types-examples/

Routing IF: 3a = yes ⇒ A = 3 3a = no ⇒ go to A4

A4. Public access (after registration)

4a. Is the dataset available for public access but holds only slight restrictions for access?

Explanation of criteria:

- A public access license allows anyone in the general public to access the dataset for free, but this usually requires the person to register on the website or login with their registered account (so that the person is known to whoever is providing the data) in order to access the data.
- An example of this is in the online EASY archive at DANS in which access to datasets often require the person to register in order to access and download the dataset.

Examples of Creative Commons[17] licenses which grant public access but hold very slight restrictions on the terms of reuse and redistribution:

- Attribution (by)
  - It is required by CC licenses that people who use another person's work in any way must give them credit the way they request. If re-users want to use this work without giving the original owner credit or for endorsement purposes, they must get the owner's permission first.
- Share alike (sa)
  - Others can copy, distribute, display, perform, and modify work, as long as they distribute any modified work on the same terms. If they want to distribute modified works under other terms, they must get the owner's permission first.
- Non-commercial (nc)
  - Others can copy, distribute, display, perform, modify and use work for any purpose other than commercially unless they get the owner's permission first.

Note! For the criterion of A = 4 it is specified that there is a license for public access of the dataset but this may contain slight restrictions for reuse, such as that registration is needed in order to access the data.

Routing

[17] https://creativecommons.org/share-your-work/licensing-types-examples/

IF:
4a = yes ⇒ A = 4
4a = no (in that it is public but has no restrictions) ⇒ go to A5

A5. Open Access (unrestricted)

_____5a. Is the dataset completely open access with no restrictions?

Explanation of the criteria:

- The license for the dataset must be completely open access to anyone in the public and with no restrictions of reuse or redistribution
- The license should be internationally recognised and machine readable
- Any documentation about the software needed to access the data should be included and described in detail

An example of a Creative Commons license which is completely open access with no restrictions:
Zero (CC0)[18]

- CC0 enables scientists, educators, artists and other creators and owners of copyright- or database-protected content to waive those interests in their works and thereby place them as completely as possible in the public domain, so that others may freely build upon, enhance and reuse the works for any purposes without restriction under copyright or database law.
- Empowers the 'no rights reserved' alternative license

Note! For the criterion of A = 5 it is specified that there is a license for open public access of the dataset containing no restrictions of reuse or modification. Registration before accessing the data should not be necessary under this criteria.

Routing
IF:
5a = yes ⇒ A = 5
For answers of 'no' to this criteria, refer back to previous steps.

Now you should have a score for the dataset which reflects its quality based on the principle 'Accessible'. Continue onto the next page to rate the data based on how 'Interoperable' it is.

Interoperable

[18] https://creativecommons.org/share-your-work/public-domain/cc0/

Some definitions of interoperable:

- Interoperability describes the extent to which systems and devices can exchange data, and interpret that shared data[19], without any restriction on access and implementation of the data
- Data interoperability reflects our ability to let computers find, access and utilise data from physically separate and heterogeneous data repositories.[20]
    - Must be machine readable = the ability for a computer to extract a description of the terms and conditions from a licence document[21] in order to compare and combine to similar data sets
- Interoperability requires a level of standardisation

I1. Proprietary, non-open format data

1a. Is the data file/are all of the data files in a proprietary format?

Explanation of proprietary format:

- A data file in a proprietary format is a file format belonging to a company, organisation or individual which is stored according to a particular encoding-scheme, designed by that company or organisation, i.e. Microsoft.
- If the software firm owning the right to that format stops making the software to read the data, then it is possible that all information is lost in those files[22]

Current examples of proprietary formats:

- Photoshop files (.psd)
- Microsoft Word (.doc) – versions prior to 1997
- Microsoft Excel (.xls) – versions prior to 1997
- Microsoft PowerPoint (.ppt) – versions prior to 1997
- WMF/EMF (Windows Metafile/Enhanced Metafile)

Note! For the criterion of I = 1 it is specified that the data format IS proprietary which means the dataset is in a closed format, in the sense that specific software is needed to read the data.

Routing IF: 1a = yes ⇒ I = 1 1a = no ⇒ go to I2 and I3

I2. Half of the data files are in a proprietary format

[19] Ole-Anders Turi, SAAMI Council

[20] Øystein Godøy, Norwegian Met Office

[21] http://www.nationalarchives.gov.uk/documents/information-management/uk-government-licensing-framework.pdf

[22] https://en.wikipedia.org/wiki/Proprietary_format

2a. Please indicate how many of the files are proprietary.

- Most of the files are proprietary
- Half of the files are proprietary
- Few of the files are proprietary
- None of the files are proprietary, all are preferred

Explanation:

- See I1 for explanation of proprietary formats.
- It is assumed that if a data file is non-proprietary that it is consistent with a 'preferred' archival format, see I3.
- Note: if there are two versions of the same data file in different formats (e.g. original proprietary version and an updated preferred version) then only score the data file which is in the preferred format (i.e. mark that file as having only one preferred file format in this case).

Note! For the criterion of I = 2 it is specified that the data format is only partially proprietary when there is a multi-file dataset.

- NB: make sure to check out the preferred formats document if your dataset file format is not listed in the examples above.

Routing IF: 2a = options 1-3 ⇒ I = 2 If 2a = option 3 ⇒ go to I3

I3. Non-proprietary, preferred open format (= "*preferred format*")

3a. Is the files/are all of the data files in a 'preferred' open format?

Explanation of preferred formats:

- Preferred formats[23] are file formats of which DANS is confident will offer the best long-term guarantees in terms of usability, accessibility and sustainability.
- These are preservation-friendly formats[24] which requires the format to follow some criteria for example:
  - The format has a large community of users
  - Is open so that the source code is available to developers
  - Format's structure and content should be human and machine readable
  - Should not depend on specific hardware or software in order to open the data
  - The format shouldn't be expected to become obsolete or need transcoding in the near future

[23] DANS File formats https://dans.knaw.nl/en/deposit/information-about-depositing-data/file-formats/file-formats

[24] http://www.digitalpreservation.gov/formats/sustain/sustain.shtml

Current examples of preferred formats:

- Unicode (.txt)
- Web pages (.html, .xhtml)
- Comma-separated values file (.csv)
- SQL (.sql)
- SPSS (.sav)
- JPEG (.jpeg)
- Portable Network Graphics file (.png)

Note! For the criterion of I = 3 it is specified that the data format is NOT proprietary but instead the data is in the preferred openly readable format which means there is no need for specific software to read the data.

- For example, unicode text files (.txt) hold only basic characters of text which means they can be read by almost any program on any platform in the computer industry, demonstrating a truly open and non-proprietary format
- Again more examples of preferred formats can be found at the direct document provided by DANS (see previous footnote).
- NB: make sure to check out the preferred formats document if your dataset file format is not listed in the examples above.

Routing IF: 3a = yes ⇒ I = 3 NB: I3 is the minimum criteria required for gaining a higher 'I' score of 4 or 5 (the criteria are cumulative), so if your data qualify for I = 3 then also see I4 and I5 incase it qualifies for a higher score.

I4. As well as in the preferred format, data is additionally standardised using a standard vocabulary (for the research field to which the data pertain)

4a. Is the dataset in the 'preferred' open format?

Explanation: See description in (I)3a for preferred data formats.
Note! For the criterion of I4a it is specified that the data IS in the preferred format.

4b. Does the dataset use a standardised format, coding scheme or markup/tagging schema?

Explanation:

- Organisations face the challenge of exchanging, comparing or integrating data among its multiple systems, and with external organisations. Unfortunately, the data collected is often in a non-standard or even non-coded (text) form, resulting in a lack of interoperability.[25]

[25] http://bok.ahima.org/doc?oid=59324#.WC2piqIrLBI

    - Data standardisation refers to the use of the same set of codes to encode data throughout a system.
- In order to connect datasets from different projects meaningfully, it is important that they use common ways of coding the data values. Many research communities have defined standard vocabularies, ontologies, coding/tagging/markup schemes, thesauri, classification systems and other semantic knowledge graphs to achieve this.

Examples of possible unstandardised vocabulary:

- Gender variables coded in non-standardised or varying manners, i.e. men = 0 and women = 1 or only using M or F to describe this.
    - If all data about sex is coded consistently according to specific vocabulary, the data should always be understandable and usable for analysis in a longitudinal fashion and across populations.[27]
- Dates consistent across countries and formats (i.e. dd/mm/yyyy or mm/dd/yyyy; dd-mm-yyyy, dd.mm.yyyy or dd/mm/yyyy)
- Country codes (i.e. always NL for Netherlands but uncertain UK/GB/EN?)
- Place names (varying based on main language data)
- English/American spellings
- Spatial co-ordinates order

Examples of standardised vocabularies which can be used:

- Locale-neutral formatting - using locale-neutral representations allows the data values to be processed accurately without the need to establish specific interchange rules that vary according to the language or location of the user.
    - For example, rather than storing "€2000,00" as a string, it's strongly preferred to exchange a data structure such as:

    - ```
      "price" {
        "value": 2000.00,
        "currency": "EUR"
      }
      ```
    - For each data field that can contain locale-affected or natural-language text, there should be an associated language tag used to indicate the language and locale of the data.[26]

Examples of vocabulary standards in various different fields:

- History of art:
    - Iconclass, see: http://www.iconclass.nl
    - Art & Architecture Thesaurus, see: http://www.getty.edu/research/tools/vocabularies/aat/
- Social, behavioural, economic and health sciences:
    - DDI (data documentation initiative), see: http://www.ddialliance.org/
- Textual studies:
    - TEI (Text Encoding Initiative), see: http://www.tei-c.org/

[26] https://www.w3.org/TR/dwbp/#LocaleParametersMetadata
      ```

- Life sciences:
    - TMO: Translational Medicine Ontology, see: https://bioportal.bioontology.org/ontologies/TMO
- Social Science History:
    - HISCO (Historical International Classification of Occupations), see: https://collab.iisg.nl/web/hisco

Resources for finding standards and ontologies:

- https://biosharing.org/ = searchable database of metadata standards, markup languages, taxonomies, and other resources for biological and life sciences.
- http://bioportal.bioontology.org/ = extensive repository of biomedical ontologies, including a recommender tool to help choose the best ontology for your research.
- http://rd-alliance.github.io/metadata-directory/standards/ = collaborative, open directory of metadata standards applicable to scientific data. Subject areas include arts and humanities, engineering, life sciences, physical sciences & mathematics, social & behavioral sciences, and general research data (multidisciplinary).
- http://lov.okfn.org/dataset/lov = provides a searchable repository of vocabularies and ontologies used to describe many different disciplines and domains.

When data vocabulary is standardised and is also explained in a codebook then it also qualifies for a Interoperable score of 4.

Note! For the criterion of I = 4 it is specified that the data DOES use a standardised format relevant to the field of research, so that integration or comparison to similar datasets is possible on human and computer levels.

Routing IF: 4a = yes AND 4b = yes ⇒ I = 4 NB: I4 is the minimum criteria for gaining a higher 'I' score of 5 (the criteria are cumulative), so if your data qualify for I = 4 then also see I5 incase it qualifies for a higher score.

NB: In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, you should provide links to more commonly used ontologies (for more information refer also to I5 - Linked Data)

- Note: if the dataset provides appropriate links to more commonly used vocabularies in this case, then it can qualify for a score of I = 4

I5. Dataset is additionally linked to other data to provide context

5a. Is the data linked to any other data which is semantically similar?

Explanation:

- Tim Berners-Lee (inventor of the World Wide Web) first sparked the idea of 'linked data' so that when you have some data on the web, you can find other related data.[27]
    - The concept of linked data is mainly needed for analytical purposes in that it provides a context for the area of research involved with the dataset; so that different representations of the same content can be found, and links made between related resources[28]
- With Linked Data, it should be possible to browse datasets, and link from one dataset to another, even if they are stored in different places and in different formats[29]
- The overall objective of 'Linked Data' is to create a global information space where the data is described following a common model (i.e. RDF - see description below), and linked to other relevant data by active links which can be used by machines as well as humans.[31]
    - Resource Description Framework (RDF) is a standard model for data interchange on the Web. It is based upon the idea of making statements about resources in the form of subject–predicate–object expressions: The subject denotes the resource (i.e. represented by URI), and the predicate denotes traits or aspects of the resource and expresses a relationship between the subject and the object.[30]
    - Using this approach, it is possible to create links between resources that are described following different models, as long as the common grammar is RDF.[31]
        - Vocabularies play an important role in the Linked Data, especially when it comes to linking data from several different domains. (this is why the criteria for 'Interoperable' are cumulative, so standardised vocabularies are required for this level as well as the linked data here)
- Please refer to the links in the footnote for more information on Linked Data and common models such as RDFs.
- Example of RDF links?

Note! For the criterion of I = 5 it is specified that the data IS linked to other semantically similar data sets to provide a context of which can be used and accessed by machines as well as humans.
NB: For a score of I5, the dataset must hold all of the qualities stated in the preceding criteria (I1-4) as well as the criteria under I5a.

---

[27] https://www.w3.org/DesignIssues/LinkedData.html and http://5stardata.info/en/
[28] http://labs.europeana.eu/api/linked-open-data-introduction
[29] http://www.ifla.org/past-wlic/2011/149-bermes-en.pdf
Convergence and Interoperability: a Linked Data perspective - Emmanuelle Bermes (2011)
[30] https://en.wikipedia.org/wiki/Resource_Description_Framework

Routing IF: 5a = yes ⇒ I = 5 For answers of ‘no’ to this criteria, refer back to previous steps.

Now you should have a score for the dataset which reflects its quality based on the principle ‘Interoperable’. Continue onto the next page for additional information about rating your dataset and what to do next.

Final rating for your dataset

- Now you should have a rating for each of the principles:
    - Findable
    - Accessible
    - Interoperable
- Keep these scores in mind to compare to the score which you are given on the FAIR data assessment tool.
- The original ‘R’ principle in FAIR as ‘Reusable’ we will now propose as a ‘Resultant’ factor of the three previously rated principles (an average taken of the three preceding scores)
    - This is because we believe that having a high score for the first three principles ‘Findable’ ‘Accessible’ and ‘Interoperable’ actually leads to the dataset being more reusable.
    - Therefore, we propose that the score of the ‘Resultant’ principle can be calculated by taking an average of the scores of the first three principles, and also serves as a kind of ‘Overall FAIRness’ rating at the same time.
- Once this has been calculated, your stars scores of the dataset may look something like the image below:

---------- End of Document ----------