DATA FAIR ASSESSMENT FRAMEWORK FOR DATA MANAGEMENT GOOD PRACTICES

IRTA

Generalitat de Catalunya
Government of Catalonia

CERCA
Research Centres
of Catalonia

Autor: Carmen Reverté Reverté. Coordinadora de gestió de dades de recerca
Col·laborador: Miguel Ángel López González. Tècnic de gestió de dades de recerca

IRTA

Index

1. Introduction

This framework is an internal document created by the research data management support service in IRTA. This framework is created with the aim of assisting the assessment of FAIR data and data management good practices in different environments (data portals, data repositories or research projects), where is needed to share, manage and disseminate.

This framework includes two relevant documents, a template to be able to evaluate datasets with a short guide to know what is being evaluated, and a second document where we find the complete guide for the evaluation of the datasets with a detailed explanation of each indicator that is taken into account, following the guidelines of the FAIR principles and European protocols to make a good data management plan.

2. Template to framework for FAIR data assessment and data management good practices

INSTITUTION	
EVALUATOR	
DATASET SAMPLE TITLE (significant name)	
DATE	
VERSION Nº	

FAIR	Indicators	Assessment			Comments
FAIR Code	Description	Level 1 (3)	Level 2 (2)	Level 3 (1)	Notes and examples
Administrative information (Findability)					
F1	ID (datasets have a unique identification number)				
F2	Funder (if applicable) + project name + project nº described (data precedence)				
F2, F5, R3	Context: dataset is described in its context? (Briefly summarize the type of study/studies to help others to understand the data)				
F2	Authors/Researchers identified + ORCID				
F2	Datasets Version (first, last)				
F1, F5	Data collection Policies (existing procedures, guidelines, etc.)				
Data Collection: which data is colleted and how?					
F1, F2, R3	Data Type (experimental, observational, simulation, derived/compiled data)				
F3, I1	Standard formats use (.xml, .xls, .sql,)				
F4	Data volume (size)				
F1, F2, F5	Software (if applicable)				

F3, I1	Open Software				
F5, R3	Software documentation				
F1, F2, R1, R3	Data description (including any existing data or third parties)				
F1, F2	Standards or Methodologies for data collection described? Or other quality procedures?				
F1, F2, A4	Data location described (structure, naming conventions, folders, servers, repositories)				
Data Access					
F1, A1, A2, A7, R4	Access Conditions Specified?				
A2, A3, A7, R4, R5	Open Data Access?				
A2, A3, A7, A8, R4, R5	Data restrictions Access defined?				
A4	Are mentioned software tools needed to data access?				
F2, A4	Is proprietary software described?				
F4, A5, A6	Storage System defined?				
F1, A4	Does it specify **where** data and associated metadata and documentation or code are deposited?				
Data Interoperability					
F3, I1	Use of Open Formats				
F2, I2, I3,	Use of standard metadata, vocabularies, or				

R1, R2	protocols for data description				
F2, I4, R3	In case of lack of standards, information about metadata or data description is known				
Data Reusability					
F2, A2, A3, A8, R1, R4, R5	Reusability defined?				
R2	Open License?				
F5, I4, R3	Documentation available (readme, data dictionaries)				
A4, A5, A6, R6	Data Preservation protocols defined (time, place, and responsibility)				
Results					

*28 parameters = 84 points (maximum level of Fairness), 56 (medium level), 28 (minimum = not optimal)

2.1 Basic instructions for parameters scoring

Different levels of completeness are defined to evaluate the FAIR Data status of different data collection. These parameters also have correspondence to the data management plan EC guidelines.

Level 1: complete = 3 points (the parameter is described, or the answer is yes)

Level 2: medium = 2 points (parameter is not complete enough – we don't have the complete information)

Level 3: not complete / not exist = 1 point (the parameter it is not defined, or answer is not complete)

In case the description **is not applicable**, please code as: **0**

2.2 Brief guide to description of the framework's FAIR parameters

Brief guide to the evaluation and description of the parameters that appear in the template to evaluate FAIR data and how these should be managed in a system or data repository.

Administrative data (Findability):

    F1. General information about what and how datasets are collected or generated and its identification & organization for its findability.

Data collection (Findability)

    F2. Have a good data description (e.g. metadata DC: title, creator, subject, description, publisher, contributor, date, type, format, identifier (PIDs), source, language, relation, coverage and rights)

    F3. Use of standard & open formats (e.g. .CSV instead of .xls)

    F4. Data volume (size determine the storage system)

    F5. Documentation: is data well documented for its reproducibility and reuse? What documentation and metadata will accompany the data?

Data Access (Accessibility):

    A1. How will you manage access and security?

    A2. How will you manage ethical issues?

    A3. How will you manage copyright and Intellectual Property Rights (IPR) issues?

    A4. How will the data be stored and backed up during the research?

    A5. Which data should be retained, shared, and/or preserved?

    A6. What is the long-term preservation plan for the dataset?

    A7. How will you share the data?

    A8. Are any restrictions on data sharing required?

Data Interoperability:

    I1. Are the data produced in the project interoperable, that is allowing data exchange and re-use between researchers, institutions, organizations, countries, etc. (i.e. adhering to standards for formats) facilitating re-combinations with different datasets from different origins)?

    I2. What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?

    I3. Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability?

    I4. In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies?

Data Reusability:

    R1. Are (Meta) data (richly described? (use of community standards)?

    R2. Are (Meta)data released with a clear and accessible data usage license?

    R3. Are the data well documented (readme files, data dictionaries…)?

    R4. When will the data be made available for re-use? If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

R5. Are the data produced and/or used in the project useable by third parties, after the end of the project? If the re-use of some data is restricted, explain why.

R6. Do you have a data preservation policy? How will the data be stored after the end of the project (data sustainability)? How long can the data be (re)used?

3. Complete guideline to support FAIR data assessment for data management good practices´

This guide is created with the aim of assisting the assessment of FAIR data and data management good practices, and to understand the indicators scope of the "Data FAIR Assessment Framework for Data Management Goof Practices Template". This Framework is a manual tool that helps you to assess the FAIRness of research datasets and how they should be managed in a data system or data repository, regardless of the open access conditions of the data.

The usefulness of this tool is that integrates both the FAIR principles, and the data management practices recommended by the European Commission (EC, 2022).

Administrative information (Findability)

ID (datasets have a unique identification number)
F1. General information about what and how datasets are collected or generated and its identification & organization for its findability.
Funder (if applicable) + project name + project nº described (data precedence)
F2. Have a good data description (e.g. metadata DC: title, creator, subject, description, publisher, contributor, date, type, format, identifier (PIDs), source, language, relation, coverage and rights)
Context: dataset is described in its context? (Briefly summarize the type of study/studies to help others to understand the data)
F2. Have a good data description (e.g. metadata DC: title, creator, subject, description, publisher, contributor, date, type, format, identifier (PIDs), source, language, relation, coverage and rights)
F5. Documentation: is data well documented for its reproducibility and reuse? What documentation and metadata will accompany the data?
R3. Are the data well documented (readme files, data dictionaries…)?
Authors/Researchers identified + ORCID
F2. Have a good data description (e.g. metadata DC: title, creator, subject, description, publisher, contributor, date, type, format, identifier (PIDs), source, language, relation, coverage and rights)
Datasets Version (first, last)
F2. Have a good data description (e.g. metadata DC: title, creator, subject, description, publisher, contributor, date, type, format, identifier (PIDs), source, language, relation, coverage and rights)
Data collection Policies (existing procedures, guidelines, etc.)
F1. General information about what and how datasets are collected or generated and its identification & organization for its findability.
F5. Documentation: is data well documented for its reproducibility and reuse? What documentation and metadata will accompany the data?

Data Collection: which data is colleted and how?

Data Type (experimental, observational, simulation, derived/compiled data)
F1. General information about what and how datasets are collected or generated and its identification & organization for its findability.

F2. Have a good data description (e.g. metadata DC: title, creator, subject, description, publisher, contributor, date, type, format, identifier (PIDs), source, language, relation, coverage and rights)

R3. Are the data well documented (readme files, data dictionaries...)?

Standard formats use (.xml, .xls, .sql,)

F3. Use of standard & open formats (e.g. .CSV instead of .xls)

I1. Are the data produced in the project interoperable, that is allowing data exchange and re-use between researchers, institutions, organizations, countries, etc. (i.e. adhering to standards for formats) facilitating re-combinations with different datasets from different origins)?

Data volume (size)

F4. Data volume (size determine the storage system)

Software (if applicable)

F2. Have a good data description (e.g. metadata DC: title, creator, subject, description, publisher, contributor, date, type, format, identifier (PIDs), source, language, relation, coverage and rights)

F5. Documentation: is data well documented for its reproducibility and reuse? What documentation and metadata will accompany the data?

Open Software

F3. Use of standard & open formats (e.g. .CSV instead of .xls)

I1. Are the data produced in the project interoperable, that is allowing data exchange and re-use between researchers, institutions, organizations, countries, etc. (i.e. adhering to standards for formats) facilitating re-combinations with different datasets from different origins)?

Software documentation

F5. Documentation: is data well documented for its reproducibility and reuse? What documentation and metadata will accompany the data?

R3. Are the data well documented (readme files, data dictionaries...)?

Data description (including any existing data or third parties)

F2. Have a good data description (e.g. metadata DC: title, creator, subject, description, publisher, contributor, date, type, format, identifier (PIDs), source, language, relation, coverage and rights)

R1. Are (Meta) data (richly described? (use of community standards)?

R3. Are the data well documented (readme files, data dictionaries...)?

Standards or Methodologies for data collection described? Or other quality procedures?

F1. General information about what and how datasets are collected or generated and its identification & organization for its findability.

F2. Have a good data description (e.g. metadata DC: title, creator, subject, description, publisher, contributor, date, type, format, identifier (PIDs), source, language, relation, coverage and rights).

Data location described (structure, naming conventions, folders, servers, repositories)

F1. General information about what and how datasets are collected or generated and its identification & organization for its findability.

A4. How will the data be stored and backed up during the research?

Data Access

Access Conditions Specified?

A1. How will you manage access and security?

A7. How will you share the data?

R4. When will the data be made available for re-use? If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

Open Data Access?

A5. Which data should be retained, shared, and/or preserved?

A6. What is the long-term preservation plan for the dataset?

A7. How will you share the data?

R4. When will the data be made available for re-use? If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

Data restrictions Access defined?

A2. How will you manage ethical issues?

A3. How will you manage copyright and Intellectual Property Rights (IPR) issues?

A8. Are any restrictions on data sharing required?

Are mentioned software tools needed to data access?

A4. How will the data be stored and backed up during the research?

A7. How will you share the data?

Is proprietary software described?

F3. Use of standard & open formats (e.g. .CSV instead of .xls)

F2. Have a good data description (e.g. metadata DC: title, creator, subject, description, publisher, contributor, date, type, format, identifier (PIDs), source, language, relation, coverage and rights)

Storage System defined?

F4. Data volume (size determine the storage system)

A4. How will the data be stored and backed up during the research?

A5. Which data should be retained, shared, and/or preserved?

Does it specify where data and associated metadata and documentation or code are deposited?

F1. General information about what and how datasets are collected or generated and its identification & organization for its findability.

A7. How will you share the data?

A4. How will the data be stored and backed up during the research?

Data Interoperability

Use of Open Formats

F3. Use of standard & open formats (e.g. .CSV instead of .xls)

I1. Are the data produced in the project interoperable, that is allowing data exchange and re-use between researchers, institutions, organizations, countries, etc. (i.e. adhering to standards for formats) facilitating re-combinations with different datasets from different origins)?

Use of standard metadata, vocabularies, or protocols for data description

I2. What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?

I3. Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability?

R1. Are (Meta) data (richly described? (use of community standards)?

In case of lack of standards, information about metadata or data description is known

I4. In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies?

Data Reusability

Reusability defined?

A2. How will you manage ethical issues?

A3. How will you manage copyright and Intellectual Property Rights (IPR) issues?

A8. Are any restrictions on data sharing required?

R5. Are the data produced and/or used in the project useable by third parties, after the end of the project? If the re-use of some data is restricted, explain why.

R2. Are (Meta)data released with a clear and accessible data usage license?
Documentation available (readme, data dictionaries)
F5. Documentation: is data well documented for its reproducibility and reuse? What documentation and metadata will accompany the data?
R3. Are the data well documented (readme files, data dictionaries…)?
Data Preservation protocols defined (time, place, and responsibility)
A4. How will the data be stored and backed up during the research?
A5. Which data should be retained, shared, and/or preserved?
A6. What is the long-term preservation plan for the dataset?
R6. Do you have a data preservation policy? How will the data be stored after the end of the project (data sustainability)? How long can the data be (re)used?

4. Bibliographic references

Checklist for a Data Management Plan (2013, August 2024). Retrieved from https://www.dcc.ac.uk/guidance/how-guides/data-management-plans

Comission, E. (2022). *Template for a Data Management Plan from the EU Funding and Tenders* Retrieved from https://cohortcoordinationboard.eu/toolkit/horizon-europe-data-management-plan-template/

Rocca-Serra, P., Gu, W., Ioannidis, V., Abbassi-Daloii, T., Capella-Gutierrez, S., Chandramouliswaran, I., . . . the, F. C. C. (2023). The FAIR Cookbook - the essential resource for and by FAIR doers. *Scientific Data, 10*(1), 292. doi:10.1038/s41597-023-02166-3

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., . . . Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data, 3*(1), 160018. doi:10.1038/sdata.2016.18