Module 1 : Intro to FAIR Principles- Metrics

Luigia Cristiano, HMC Hub Matter

Helmholtz Institute Berlin

Why this training

Familiarize end users with FAIR data curation

- How to extract, format and standardize data description
- How to stage the data in institutional repository
- Discuss licences and data copyright

- Overall we want to show the benefits of FAIR data sharing and foster their use in the daily research activity

We ease FAIR you use FAIR

Why getting you on board

Researchers difficulties in making data available in repositories

- Allocate time to describe the data
- Finding trusted repository
- Lack of specific tools and curation expertise
- Insecurity in terms of misuse, licencing and authorship
- Data publication not as first class product

HMC can support the curation of your research data

01 FAIR Principles and PaN data life cycle

Benefits aka

Why a good data management in research is important

- Encourage high quality data
- Foster/enable the data discovery and reuse and support efficient data usage with machine readability and data linking
- Enhance visibility of the research results and increase funding and cooperation opportunities
- Access to data and processing codes is mandatory for paper submission
- EU projects are incentivating FAIR by funding the efforts
- Promote citizen science
- Ensure trasparency and reduce costs

Data Statement	Funding Agency			
	H2020	ERC	DFG	BMBF
Open Access Policy	as open as possible, as closed as necessary	as open as possible, as closed as necessary	as open as possible, as closed as necessary	as open as possible, as closed as necessary
DMP requested	within first six months of project	as part of the proposal	as part of the proposal	as part of the proposal
Template available?	yes	yes	yes	yes
Which data should be made available?	all data and metadata	data collections and metadata	reusable raw and structured (meta)data	
When should data be made available?	as soon as possible, embargo possible	as soon as possible, embargo possible	as soon as possible	as soon as possible, at the latest 9 months after completion of project

https://dataservices.gfz-potsdam.de/portal/drr.html

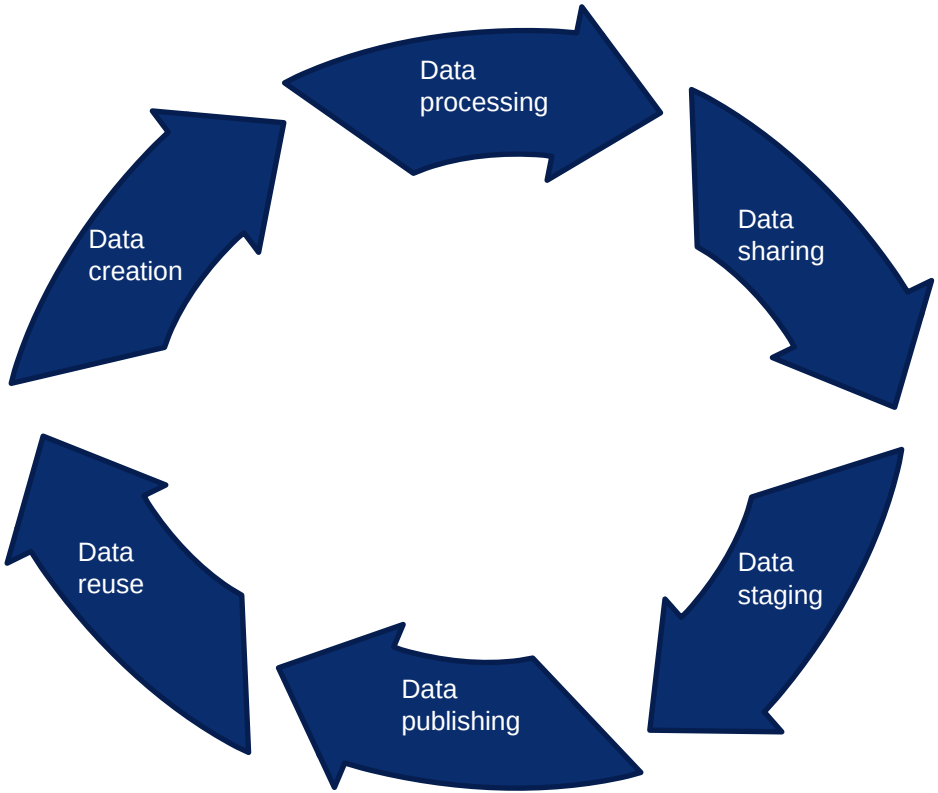
Research Data Management (RDM)

- Protocols and actions to optimize the research data usage
- The implementation of RDM:

  technical, organization and legal aspects must be accounted

- Consultancy is offered from data managers to researchers on:

  - Optimization of data staging in distributed locations
  - Development of tools and workflows for data curation
  - Guidineline to implement data management
  - Scale down the complexity



How to start

DO FAIR

Data management can boost the impact of your research data

Reduce the complexity by using machine readable metadata and **standard** formats,

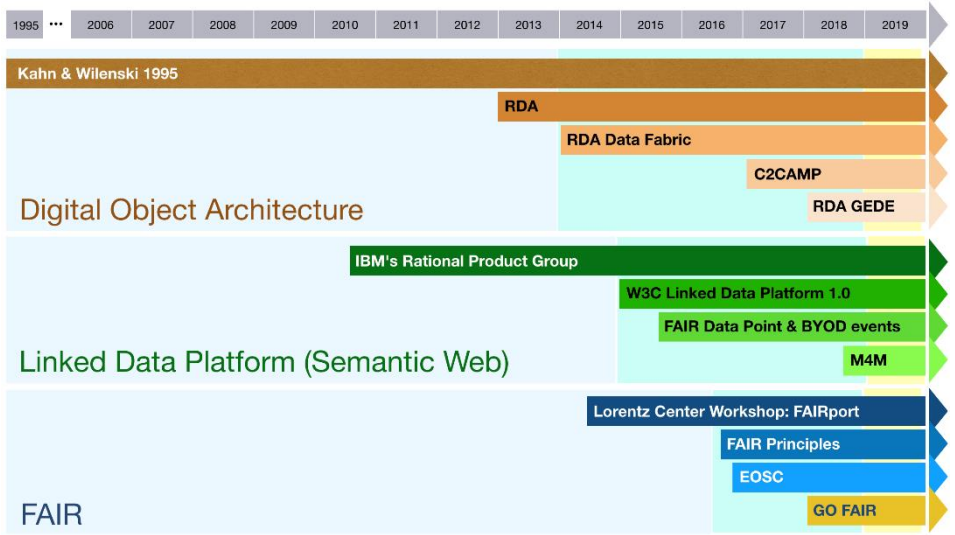
Ensure the accessibility to your data and implement practice to data reuse (**licences**)

Use a Trusted **data repository** and **author** your data assigning **PID**s

FAIR: Turning data to knowledge.

History of the FAIR initiative and ~~how realize that high quality, accessible data support innovation~~

- Lorentz workshop „jointly designing a Data FAIRport, 2014
- Mark D. Wilkinson et al., ‘The FAIR Guiding Principles for Scientific Data Management and Stewardship,’ Scientific Data 3 (March 15, 2016): 160018.)
- European task force /action plan document 2018
- Fair data maturity model, 2020



Timeline of developments in the convergence of various approaches to data infrastructures: conceptualization (blue background), design (green background), and implementation (yellow background)

Why FAIR

- FAIR is not Open Access
- Different solutions and different paces of implementation are suggested for the different communities
- **Findable**

  Persistent ID

  Metadata online

  Online Repository
- **Accessible**

  Community or generic repository
- **Interoperable**

  Community or generic standards

  Open file formats
- **Reusable**

  Rich, complete and clear documentation

  Licenced

Guidelines for promoting data visibility, reuse and ensure data quality

Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3:160018 doi: 10.1038/sdata.2016.18 (2016).

Box 2 | The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

How to FAIR: Implementation Challenges for Researchers

- How to plan
- How to create suitable structures to host data, ensure the access, the privacy and licensing of the data and the versioning
- How to allocate time and resources for it

FAIR Principles in detail

The principles refer to three types of entities: any digital object, metadata and infrastructure

Meta(data) should be easy to find and readable for humans and computers.

F1. (*Meta)data are assigned a globally unique and persistent identifier*.

F2. *Data are described with rich metadata*.

F3. *Metadata clearly and explicitly include the identifier of the data they describe.*

F4. *(Meta)data are registered or indexed in a searchable resource*.

On how meta(data) can be accessed, possibly including authentication and authorisation.

A1. *(Meta)data are retrievable by their identifier using a standardised communications protocol*.

A1.1 *The protocol is open, free, and universally implementable.*

A1.2 *The protocol allows for an authentication and authorisation procedure, where necessary*.

A2. *Metadata are accessible, even when the data are no longer available*.

I1. *(Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.*

I2. (*Meta)data use vocabularies that follow FAIR principles.*

I3. *(Meta)data include qualified references to other (meta)data.*

R1. *(Meta)data are richly described with a plurality of accurate and relevant*

attributes.

R1.1. *(Meta)data are released with a clear and accessible data usage license.*

R1.2. *(Meta)data are associated with detailed provenance*.

R1.3. *(Meta)data meet domain-relevant community standards*.

https://www.go-fair.org/wp-content/uploads/2022/01/FAIRPrinciples_overview.pdf

FAIR Metrics

- How to measure how FAIR are my data ?
- What has been done in this sense so far
- Let‘s have a look at it in details

https://fairsharing.org/search?q=&selected_facets=type_exact%3Ametric&fairsharingRegistry=Standard

https://fairdata.services:7171/FAIR_Evaluator/

https://metrics.api.fair-enough.semanticscience.org/docs#/R1/Conforms_to_community_standards_tests_r1_community_standards_get

FAIR Data Maturity Model: specification and guidelines

- How to measure how FAIR are my data ?
- What has been done in this sense so far
- Let‘s have a look at it in details
- Indicators and application examples

https://www.rd-alliance.org/groups/fair-data-maturity-model-wg

FAIR	ID	Indikator	Priorität	
F1	RDA-F1-01M	Metadaten werden durch einen persistenten Identifikator identifiziert	●●●	Wesentlich
F1	RDA-F1-01D	Daten werden durch einen persistenten Identifikator identifiziert	●●●	Wesentlich
F1	RDA-F1-02M	Metadaten werden durch einen global eindeutigen Identifier identifiziert	●●●	Wesentlich
F1	RDA-F1-02D	Daten werden durch einen global eindeutigen Identifier identifiziert	●●●	Wesentlich
F2	RDA-F2-01M	Umfangreiche Metadaten werden bereitgestellt, um ein Auffinden zu ermöglichen	●●●	Wesentlich
F3	RDA-F3-01M	Metadaten beinhalten den Identifikator der Daten	●●●	Wesentlich
F4	RDA-F4-01M	Metadaten werden so angeboten, dass sie abgefragt und indexiert werden können	●●●	Wesentlich
A1	RDA-A1-01M	Metadaten enthalten Informationen, die es dem Nutzer ermöglichen, auf die Daten zuzugreifen	●●	Wichtig
A1	RDA-A1-02M	Metadaten können manuell (d. h. mit menschlicher Beteiligung) abgerufen werden	●●●	Wesentlich
A1	RDA-A1-02D	Daten können manuell (d. h. mit menschlicher Beteiligung) abgerufen werden	●●●	Wesentlich
A1	RDA-A1-03M	Der Metadatenidentifikator führt zu einem Metadatensatz	●●●	Wesentlich
A1	RDA-A1-03D	Der Datenidentifikator führt zu einem digitalen Objekt	●●●	Wesentlich
A1	RDA-A1-04M	Der Zugriff auf Metadaten erfolgt über ein standardisiertes Protokoll	●●●	Wesentlich
A1	RDA-A1-04D	Der Zugriff auf Daten erfolgt über ein standardisiertes Protokoll	●●●	Wesentlich
A1	RDA-A1-05D	Der Zugriff auf Daten kann automatisch (d. h. durch ein Computerprogramm) erfolgen	●●	Wichtig
A1.1	RDA-A1.1-01M	Der Zugriff auf Metadaten kann über ein freies Zugriffsprotokoll erfolgen	●●●	Wesentlich
A1.1	RDA-A1.1-01D	Der Zugriff auf Daten kann über ein freies Zugriffsprotokoll erfolgen	●●	Wichtig
A1.2	RDA-A1.2-01D	Der Zugriff auf Daten kann über ein Zugriffsprotokoll erfolgen, das Authentifizierung und Autorisierung unterstützt	●	Nützlich
A2	RDA-A2-01M	Metadaten bleiben garantiert verfügbar, wenn die Daten nicht mehr verfügbar sind	●●●	Wesentlich
I1	RDA-I1-01M	Metadaten verwenden eine in einem standardisierten Format ausgedrückte Wissensdarstellung	●●	Wichtig
I1	RDA-I1-01D	Daten verwenden eine in einem standardisierten Format ausgedrückte Wissensdarstellung	●●	Wichtig
I1	RDA-I1-02M	Metadaten verwenden eine maschinenlesbare Wissensdarstellung	●●	Wichtig
I1	RDA-I1-02D	Daten verwenden eine maschinenlesbare Wissensdarstellung	●●	Wichtig
I2	RDA-I2-01M	Metadaten verwenden FAIR-konforme Vokabularien	●●	Wichtig
I2	RDA-I2-01D	Daten verwenden FAIR-konforme Vokabularien	●	Nützlich
I3	RDA-I3-01M	Metadaten enthalten Verweise auf andere Metadaten	●●	Wichtig

Indicators and prioritization, how are they applied ?

	Principle				
Priority	Findable	Accessible	Interoperable	Reusable	Grand Total
Essential	7	8	0	5	20
Important	0	3	7	4	14
Useful	0	1	5	1	7
Grand Total	7	**12**	**12**	**10**	**41**

Figure 3: Distribution of priorities per FAIR area as defined by the RDA Working Group

FAIR for Software

- How to measure how FAIR is software.

Interactivity section on state of the art

	FAIR for data	FAIR for software	Operation
F1	(Meta)data are assigned a globally unique and persistent identifier.	Software and its associated metadata have a global, unique and persistent identifier for each released version.	Rephrased
F2	Data are described with rich metadata.	Software is described with rich metadata.	Rephrased
F3	Metadata clearly and explicitly include the identifier of the data it describes.	Metadata clearly and explicitly include identifiers for all the versions of the software it describes.	Rephrased and extended
F4	(Meta)data are registered or indexed in a searchable resource.	Software and its associated metadata are included in a searchable software registry.	Rephrased
A1	(Meta)data are retrievable by their identifier using a standardized communications protocol.	Software and its associated metadata are accessible by their identifier using a standardized communications protocol.	Rephrased
A1.1	The protocol is open, free, and universally implementable.	The protocol is open, free, and universally implementable.	Remain the same
A1.2	The protocol allows for an authentication and authorization procedure, where necessary.	The protocol allows for an authentication and authorization procedure, where necessary.	Remain the same
A2	Metadata are accessible, even when the data are no longer available.	Software metadata are accessible, even when the software is no longer available.	Rephrased
I1	(Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.	Software and its associated metadata use a formal, accessible, shared and broadly applicable language to facilitate machine readability and data exchange.	Rephrased and extended
I2	(Meta)data use vocabularies that follow FAIR principles.	–	Reinterpreted, extended and split
I2S.1	–	Software and its associated metadata are formally described using controlled vocabularies that follow the FAIR principles.	Reinterpreted, extended and split
I2S.2	–	Software use and produce data in types and formats that are formally described using controlled vocabularies that follow the FAIR principles.	Reinterpreted, extended and split
I3	(Meta)data include qualified references to other (meta)data.	–	Discarded
I4S	–	Software dependencies are documented and mechanisms to access them exist.	Newly proposed
R1	(Meta)data are richly described with a plurality of accurate and relevant attributes.	Software and its associated metadata are richly described with a plurality of accurate and relevant attributes.	Rephrased
R1.1	(Meta)data are released with a clear and accessible data usage license.	Software and its associated metadata have independent, clear and accessible usage licenses compatible with the software dependencies.	Rephrased and extended
R1.2	(Meta)data are associated with detailed provenance.	Software metadata include detailed provenance, detail level should be community agreed.	Rephrased
R1.3	(Meta)data meet domain-relevant community standards.	Software metadata and documentation meet domain-relevant community standards.	Rephrased

How to FAIR

Implementation Challenges for Researchers

- How to plan
- How to create suitable structures to host data, ensure the access, the privacy and licensing of the data and the versioning
- How to allocate time and resources for it

Practices for data curation:

Data description and staging

- Write a clear documentation for the data interpretation and reuse
- Use community standards for data and metadata when available
- Use repositories for staging the data
- Automatize the data processing and curation
- Assign PIDs to datasets, software and cite them in your publication

FAIR implementation Profile

Researcher and data provider

- Curate data description
- Use standard formats
- Follow the FAIR guidelines and contribute to the services optimization
- Cite datasets using DOI

Institutional data manager

- Raise awareness of community standards
- Support researchers with tailored services
- Ensure standard protocols to data access and the compliance to data policy
- Associate an ID to the data
- Enable the data visibility
- Promote good practices for research data management
- Scale down the complexity

HMC and reserchers community work together for a GO BUILD

Thanks.

luigia.cristiano@helmholtz-berlin.de
And the Hub Matter Team

HMC Hub Matter