# E-Commons RDM "workshop"

Jeremy Azzopardi | Chalmers Data Office | 2024-04-09

Agenda

Main aim: overview of the RDM and its context – why and how is it useful?

- Overview of RDM, OS, RDM lifecycle
- Data reusability – the red thread
- Infrastructure for reuse – storage, archives, repositories, data catalogues
- FAIR – in theory and practice
- Making all of this somehow usable for researchers – opportunities and challenges
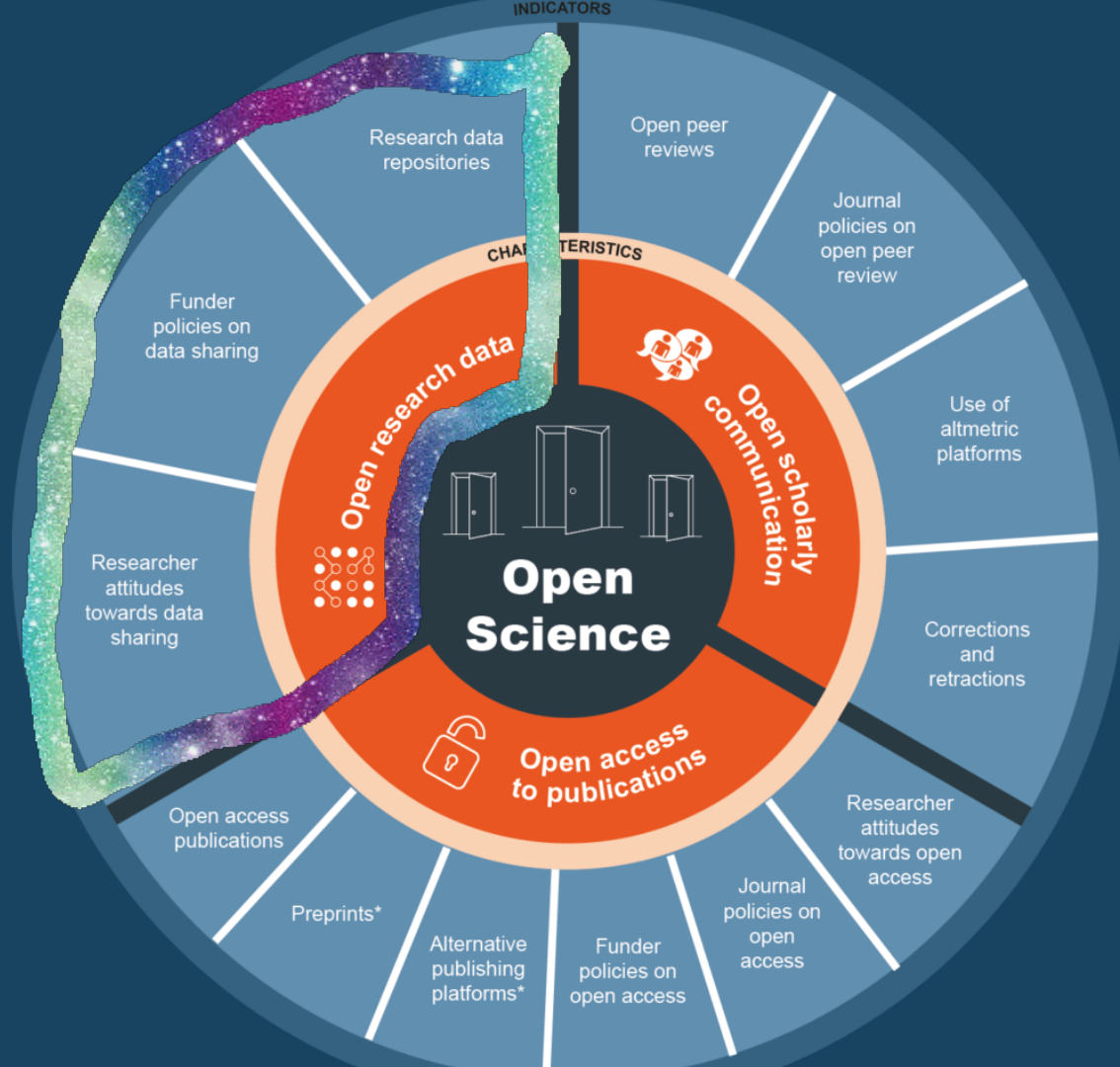- Questions
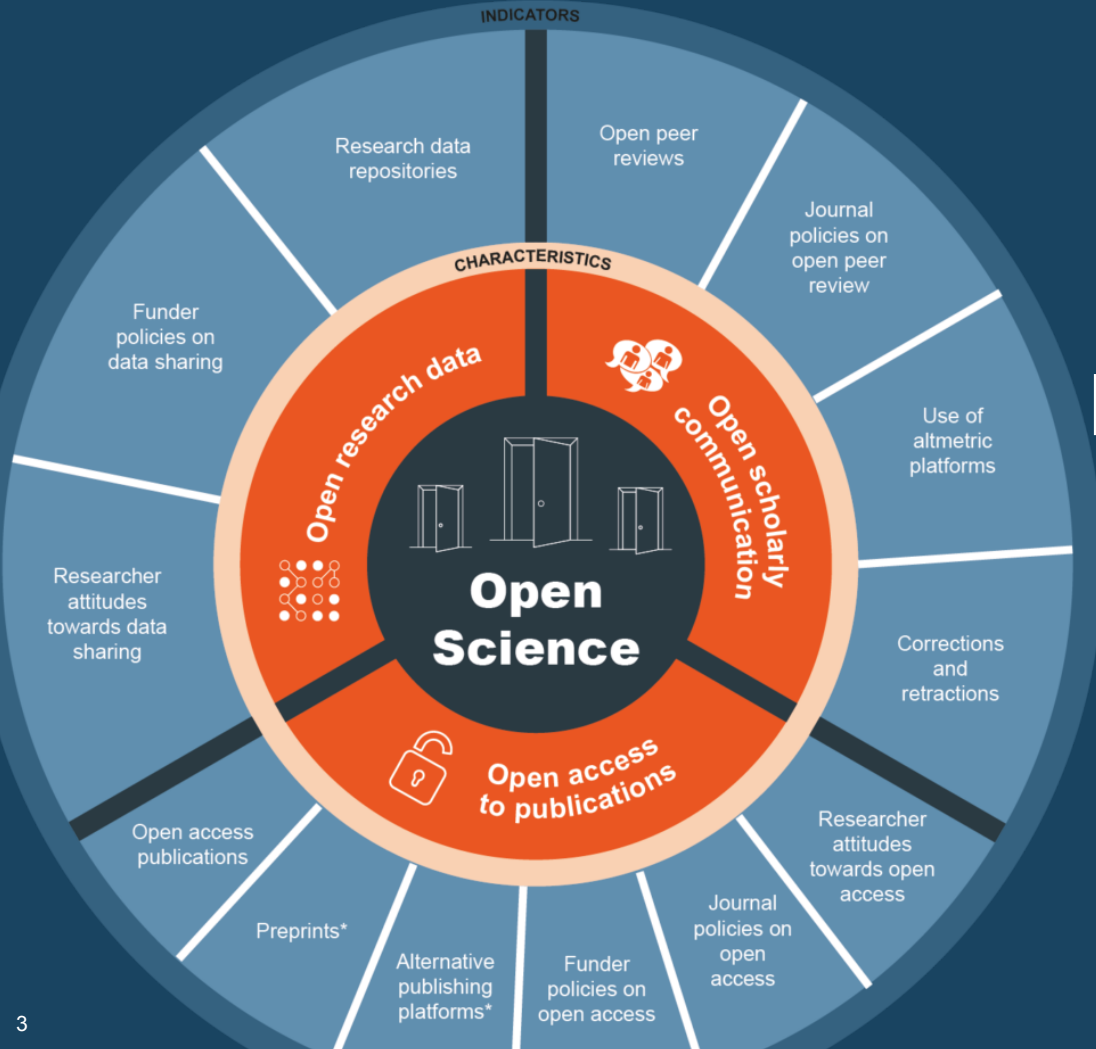
For context - OS

- For researchers' (+ colleagues) own sake
- Because you have to (funding requirements, local policies, laws, etc.)
- To make research & knowledge sharing more effective, collaborative, open, equitable.

Data, metadata, documentation

INFORMATION TYPES

Data: Material collected or produced for scientific purposes, and forming the basis for publications (Swedish Research Council)

Documentation: information to interpret and use data. E.g. provenance, file/folder/database structure and naming, versioning.

Metadata: data about data – structured information, in this context linked to making data available externally via databases, repositories, etc.

PLAN & DESIGN → COLLECT & CAPTURE → COLLABORATE & ANALYSE ↓ MANAGE, STORE & PRESERVE ← SHARE & PUBLISH ← DISCOVER, REUSE & CITE ↑

2024-04-08
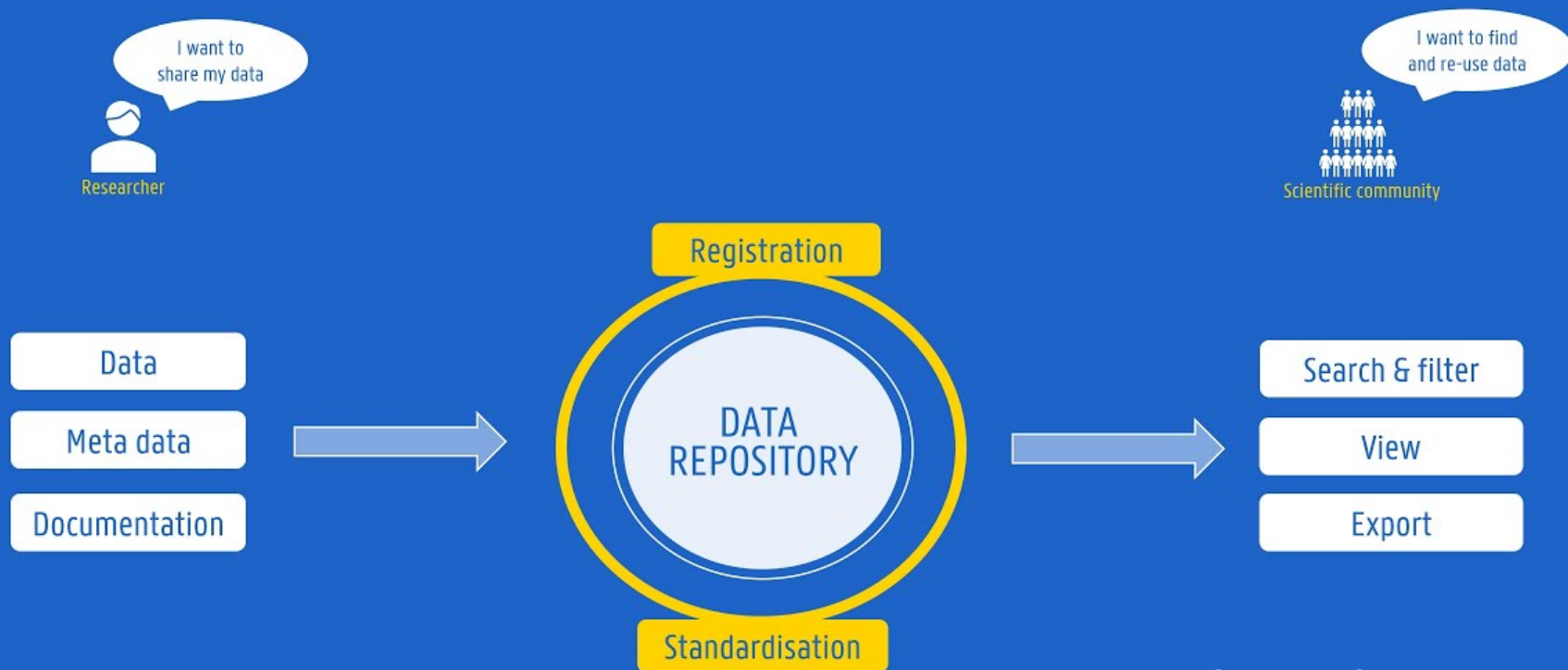
Data reuse needs infrastructure

- (Research) Data Repositories
- Data Catalogs
- Archives
- Storage and sharing
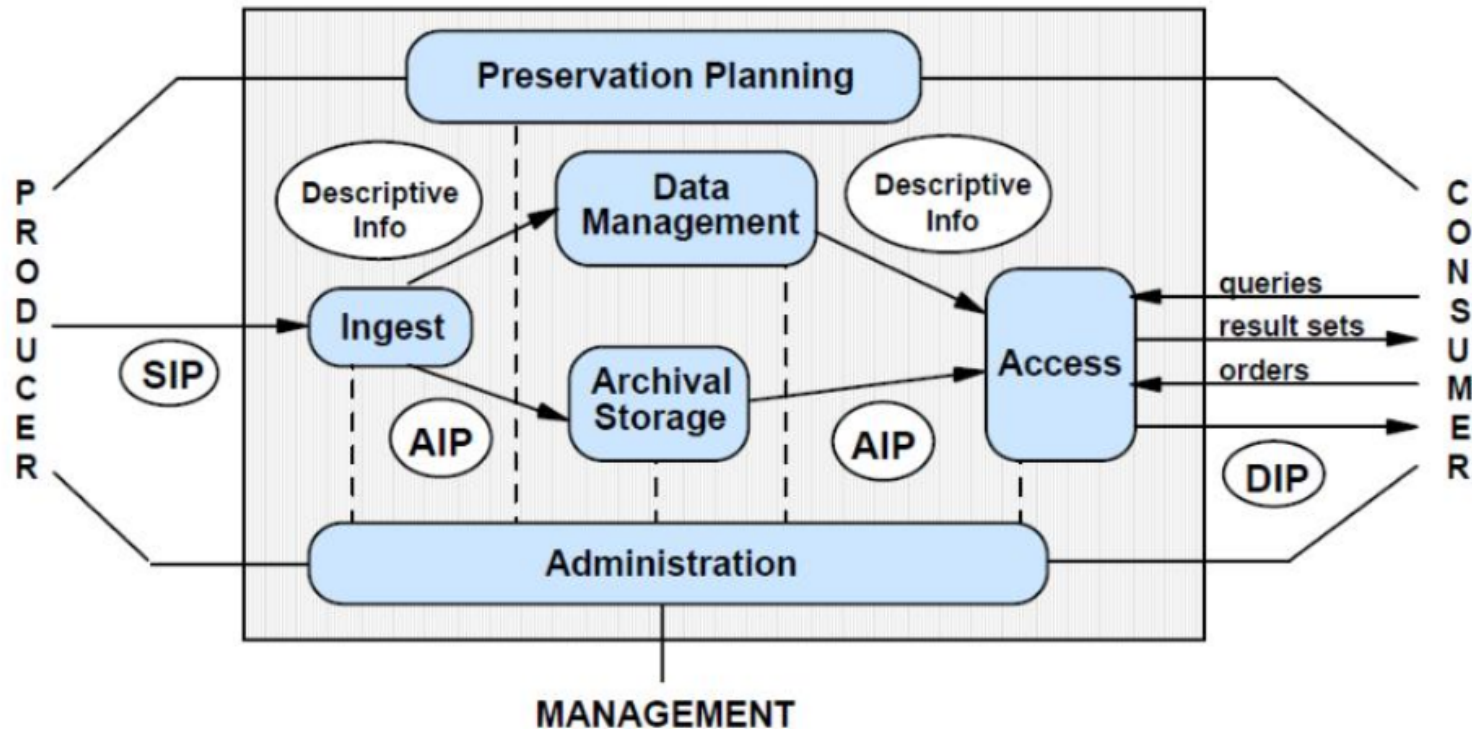- Tools (DMP, ELN, etc.)
- Support + training

WHAT IS A DATA REPOSITORY?

A database infrastructure that collects, manages and stores data for preservation and sharing


Source: Ghent University

The OAIS Basic Model

CoreTrustSeal

- Most popular certification for resaerch data repositories
- 16 requirements – 3 categories:
  - Governance + Organization (mission, funding, confidentiality,e expertise, continuity)
  - Digital Object Management (workflows, procedures, documentation, models, in-house expertise, etc.)
  - Technology (tech infra + security)

Data Catalogues

- Often a layer on top of repositories, but can be separate (pointing via DOI), especially when metadata is propagated

- Uses common metadata standards, controlled vocabularies, ontologies, etc. to structure metadata for easy search, sharing, retrieval.

- Offer API:s for metadata

2024-04-08

FAIR – getting data from creation to reuse
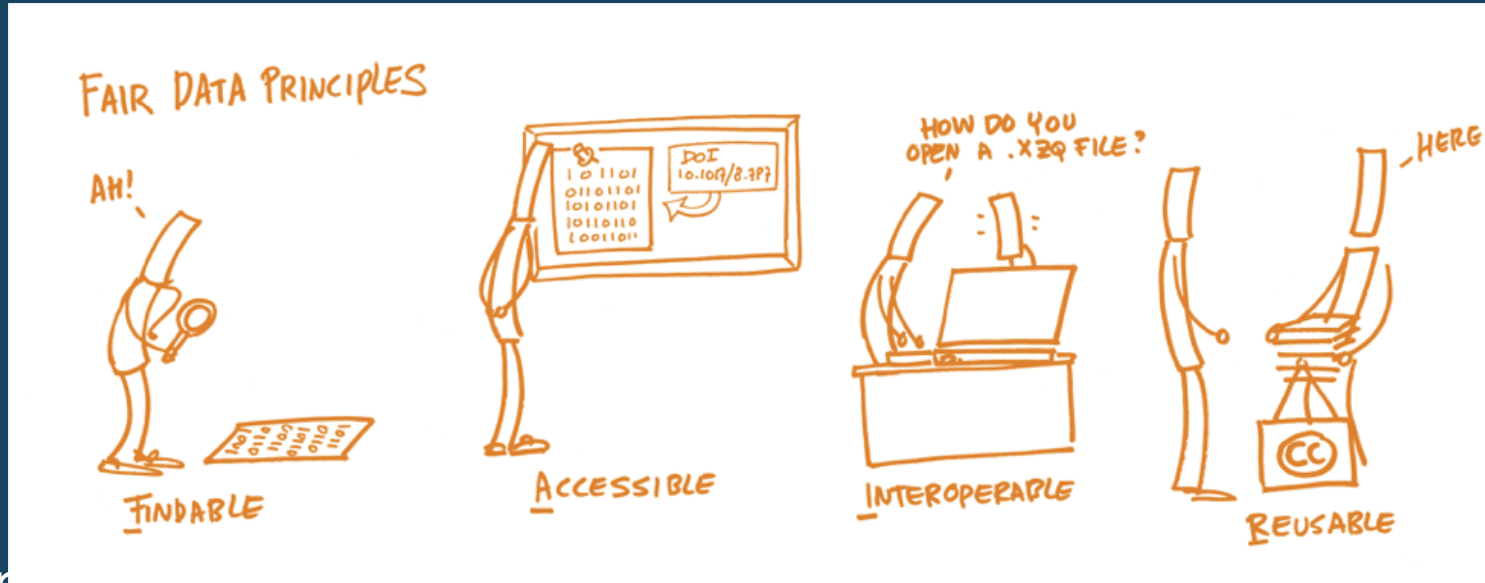
*Aim: machine-actionable data

In 2016, the '**FAIR Guiding Principles for scientific data management and stewardship**' were published in *Scientific Data*. The authors intended to provide guidelines to improve the **F**indability, **A**ccessibility, **I**nteroperability, and **R**euse of digital assets. The principles emphasise machine-actionability (i.e., the capacity of computational systems to find, access, interoperate, and reuse data with none or minimal human intervention) because humans increasingly rely on computational support to deal with data as a result of the increase in volume, complexity, and creation speed of data.

What FAIR isn't

- FAIR is *not* a standard
- FAIR is *not* equal to 'Open' or 'Free'
- Data are often Open but not FAIR
- Data could be Closed, yet perfectly FAIR
- FAIR is *not* equal to Linked Data, Semantic Web or RDF
- FAIR is *not* assuming that only humans can find and re-use data
- FAIR is *not* for humans only but for machines as well
- Data that are *not* FAIR are pretty *'Re-useless'*…..

M. Wilkinson et al. (2018): *Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud,* http://doi.org/10.3233/ISU-170824 and GO-FAIR (2018): *FAIR Data Stewardship Awareness Course.*

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

Technical infrastructure: Repository
Data/metadata: biomedical, geological, historical...

Box 2 | The FAIR Guiding Principles

https://www.nature.com/articles/sdata201618

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

- Credit: Erik Schultes go-fair.org shorturl.at/aijDQ

2024-04-08

Technical infrastructure: Repository
Data/metadata: biomedical, geological, historical...

Box 2 | The FAIR Guiding Principles   https://www.nature.com/articles/sdata201618

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

Partly via repo!

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

Partly via repo!

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

Partly via repo!

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

Partly via repo!

2024-04-08

Steps how to make data FAIR

Findable
- select a data repository at an early stage and check out its data format and metadata requirements
- make sure the data can get a persistent identifier so that it can be cited
- select a catalogue to make your data more findable, especially if the repository is more generic in nature

Accessible
- guarantee longevity of the data (i.e., by submitting it to a repository that has a certification like e.g. ISO)
- check and describe the legal conditions under which the data can be made available
- establish an embargo period if necessary
- make sure your ICT infrastructure will keep the data available even in case of equipment failure or human error

Interoperable
- select commonly used data formats
- select commonly used vocabularies for data items

Reusable
- make sure you keep proper provenance information (i.e., details about how and where the data was generated, incl. machine settings, details about processing steps: the software tools with their versions and parameters)
- select the right minimal metadata standard and collect the necessary metadata (many minimal metadata standards are included in ELIXIR's biosharing.org repository)
- select a license for the data (preferably an open license) and the associated software tools
- make sure the important conclusions of your study will not only be available in a paper in a narrated form, but also in a digital file (e.g., a nanopublication)

- Source: https://www.dtls.nl/fair-data/fair-data-knowledge-expertise/

One (similar) implementation – Linked Open Data

★ make your stuff available on the Web (whatever format) under an open license[1]

★★ make it available as structured data (e.g., Excel instead of image scan of a table)[2]

★★★ make it available in a non-proprietary open format (e.g., CSV instead of Excel)[3]

★★★★ use URIs to denote things, so that people can point at your stuff[4]

★★★★★ link your data to other data to provide context[5]

Using a (trusted) data repository is a big step towards FAIR

- Infrastructure fulfils FAIR

- Forces FAIRness in metadata

- Can facilitate FAIRness in data

Trusted…?

- Provision of **Persistent and Unique identifiers** (e.g. DOI) for data and metadata
- Enables **searching, citing and retrieving** data
- Supports **data versioning**
- Provides **metadata**, openly, based on accepted standards, even for removed/non-accessible data
- **clear licence and provenance** information, and authentication if necessary
- **transparent** about mission, scope, preservation policies, retention periods, continuity etc.

Science Europe has a good guide: https://www.scienceeurope.org/media/ffkb51ei/se-rdm-template-2-criteria-for-the-selection-of-trustworthy-repositories.docx
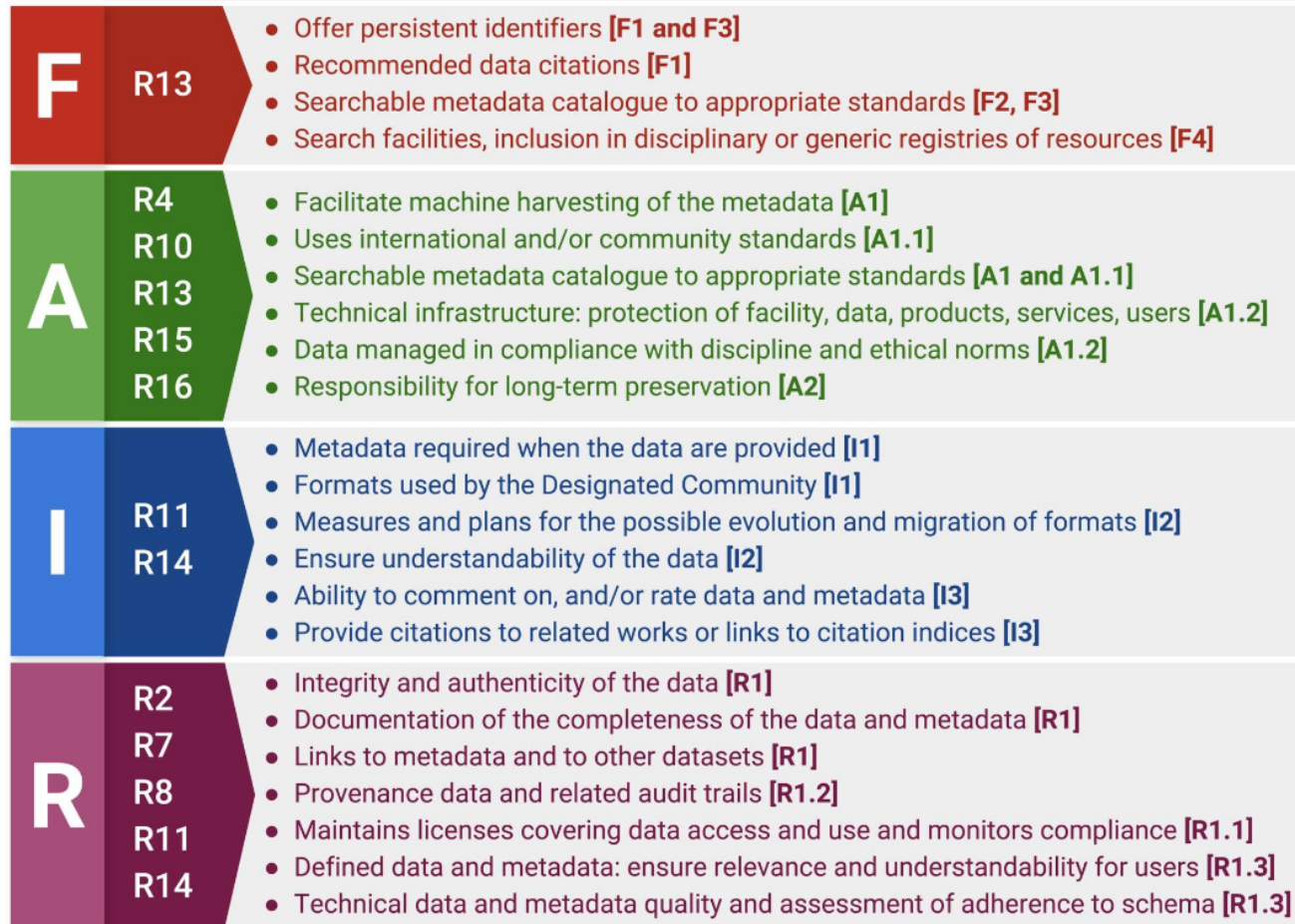
Figure 1 Mapping between FAIR and CoreTrustSeal.

Legal, funder, policy requirements on data management

- Offentlighetsprincipen / Principle of public access to official records
- Arkivlagen (archival law)
- GDPR
- Ethics law
- Export control
- Funder requirements
- University policies on security, OS, etc.
- Etc. etc. etc.

2024-04-08

Chalmers Guiding principles for good research practice

- Chalmers policy document. Registration number: C 2023-0615
- Three sections: Good management of Research Data, Open access, Responsible internationalisation
- Good management of Research Data: 3 Rules

Rule 1: All research activities at Chalmers shall establish and maintain a data management plan.

Chalmers: Guiding principles for good research practice (C 2023-0615)

1. Good management of research data
2. Open Access to scientific publications
3. Responsible internationalisation

Good RDM policy: Three rules

- Rule 1: All research activities at Chalmers shall establish and maintain a data management plan
- Rule 2: The right to decide on the use of research data shall be retained within Chalmers and data shall as far as possible be stored under Chalmers control
- Rule 3: All research data that is considered important for long-term documentation of the research should be as FAIR as possible

Data management plan

documentation and metadata

safety storage and backup

data sharing

project continuity

ehtics and legal compliance

project integrity

reusable data

responsibilities and resources

Data Stewardship Wizard
dsw.chalmers.se (Intranet >E-Services)

View   Comments   TODOs   Version history

Current Phase

Before Submitting the DMP

Chapters

I. Disclaimer - Read this first!	✓
II. About the research project	✓
III. Description of existing data	✓

Are there any data sets available that are relevant to...

| IV. Description of new data (creating and/or co... | 17 |
| V. Processing data, storage and backup | 3 |

1 **Are there any data sets available that are relevant to this project?**

Is there any pre-existing data any where in the world that is relevant to your research?

Search for data sets at re3data

○ **a. No**

◉ **b. Yes**

↺ Clear answer

Answered 5 months ago by Jeremy Azzopardi.

1.b.1 **What existing data have you considered re-using?**

Much of todays data is used in comparison with reference data. You may be comparing your own data with a "standard