The FAIR Research Data Concept

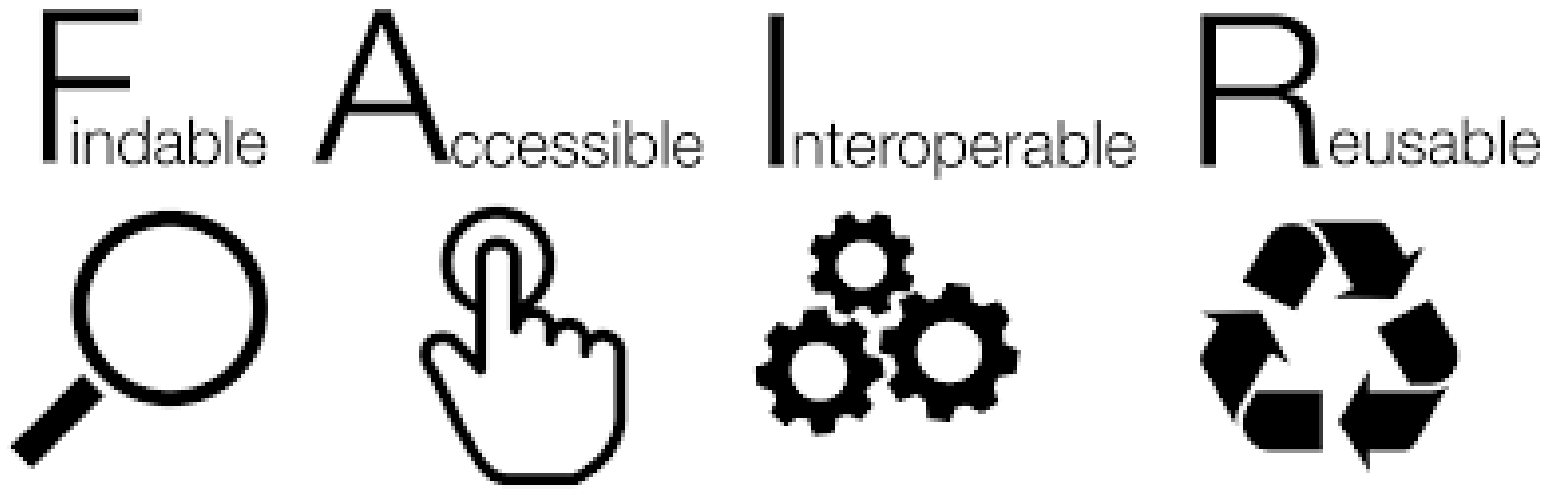
Heinrich Widmann
FifF-Hamburg Treffen
17.04.2019

Outline

- **What is FAIR ?**
- **Data intensive science**
- **The FAIR principles and matrix**
- **Examples for**
  - **FAIR data : Climate simulation output**
  - **FAIRfriendly service : Interdisciplinary discovery by EUDAT-B2FIND**
- **Initiatives and projects driving FAIRisation**
- **Issues and challenges**

What is FAIR ?

A set of principles to ensure that data are shared in a way that enables & enhances reuse, by humans and machines



Motivation : Data intensive science

Data intensive science means

- research in which the capture, curation, and analysis of (usually) **large volumes of data are central to the scientific question**
- research that uses **data sets so large or complex** that they are **hard to process and analyze using traditional approaches** and methods

Challenges and paradigm change :

- **In former times** scientist primarily work on her/his own (+ input from colleagues) to analyze the data
- **In contrast, nowadays** scientists work
  - on **(huge) data resources distributed worldwide in cross-border repositories**
  - with **computer/data scientists and experts in eco-infrastructures** and
  - use **IT-infrastructures in order to make optimal use of contemporary computational tools for integrating, creating, and analyzing data.**

Solution

- Provide researchers with clear guidelines for ‘best practices’ for data management
- Support scientists in focussing on science through IT eco-systems and ‘hide technics’
- Provide data infrastructures allowing scientist to transparently share research output to enable others to correct and reuse them

Origins of FAIR

- **Emerged from a workshop held in Leiden in 2014**
- **Come from life sciences but intended for all (scientific) data**
- **Issued by FORCE11 community**
- **Echo previous principles on open data & curation**

+ OECD Principles (2017)
+ The Royal Society (2012)
+ G8 Science Ministers Statement (2013)

Basic paper in 2016

SCIENTIFIC DATA

Sci Data. 2016; 3: 160018. PMCID: PMC4792175
Published online 2016 Mar 15. doi: 10.1038/sdata.2016.18
Comment

The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson,[1] Michel Dumontier,[2] IJsbrand Jan Aalbersberg,[3] Gabrielle Appleton,[3] Myles Axton,[4] Arie Baak,[5] Niklas Blomberg,[6] Jan-Willem Boiten,[7] Luiz Bonino da Silva Santos,[8] Philip E. Bourne,[9] Jildau Bouwman,[10] Anthony J. Brookes,[11] Tim Clark,[12] Mercè Crosas,[13] Ingrid Dillo,[14] Olivier Dumon,[3] Scott Edmunds,[15] Chris T. Evelo,[16] Richard Finkers,[17] Alejandra Gonzalez-Beltran,[18] Alasdair J.G. Gray,[19] Paul Groth,[3] Carole Goble,[20] Jeffrey S. Grethe,[21] Jaap Heringa,[22] Peter A.C 't Hoen,[23] Rob Hooft,[24] Tobias Kuhn,[25] Ruben Kok,[22] Joost Kok,[26] Scott J. Lusher,[27] Maryann E. Martone,[28] Albert Mons,[29] Abel L. Packer,[30] Bengt Persson,[31] Philippe Rocca-Serra,[18] Marco Roos,[32] Rene van Schaik,[33] Susanna-Assunta Sansone,[18] Erik Schultes,[34] Thierry Sengstag,[35] Ted Slater,[36] George Strawn,[37] Morris A. Swertz,[38] Mark Thompson,[32] Johan van der Lei,[39] Erik van Mulligen,[39] Jan Velterop,[40] Andra Waagmeester,[41] Peter Wittenburg,[42] Katherine Wolstencroft,[43] Jun Zhao,[44] and Barend Mons[a,45,46,47]

Abstract

Go to:

There is an urgent need to improve the infrastructure supporting the reuse of scholarly data. A diverse set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—have come together to design and jointly endorse a concise and measureable set of principles that we refer to as the FAIR Data Principles. The intent is that these may act as a guideline for those wishing to enhance the reusability of their data holdings. Distinct from peer initiatives that focus on the human scholar, the FAIR Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals. This Comment is the first formal publication of the FAIR Principles, and includes the rationale behind them, and some exemplar implementations in the community.

http://doi.org/10.1038/sdata.2016.18

The FAIR principles

According to the FAIR Data Principles, data should be:

1. **Findable** – Easy to find by both humans and computer systems and based on mandatory description of the metadata that allow the discovery of interesting datasets;
2. **Accessible** – Stored for long term such that they can be easily accessed and/or downloaded with well-defined license and access conditions (*Open when possible, closed when necessary*), whether at the level of metadata, or at the level of the actual data content;
3. **Interoperable** – Ready to be combined with other datasets by humans as well as computer systems;
4. **Reusable** – Ready to be used for future research and to be processed further using computational methods.

What FAIR means: 15 principles

Findable (Data Discovery)

F1. (meta)data are assigned a globally unique and eternally **persistent identifier**
F2. data are described with **rich metadata**
F3. (meta)data are registered or **indexed in a searchable resource**
F4. metadata **specifify the data identifier**

Accessible

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

Interoperable

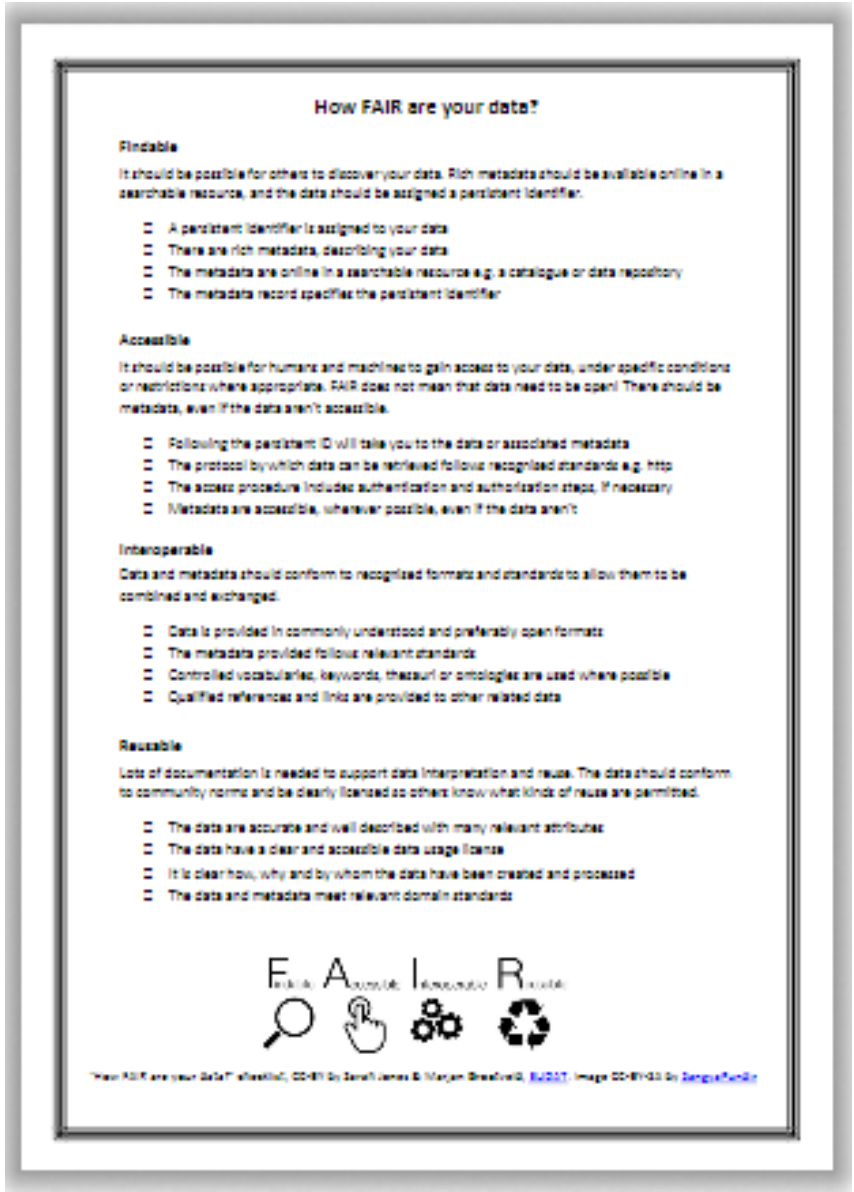
I1. (Meta)data use a formal, accessible, shared, and broadly applicable **language for knowledge representation**.
I2. (Meta)data use **vocabularies** that follow FAIR principles
I3. (Meta)data include **qualified references** to other (meta)data

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 FAIR principles are implemented and FAIR data checklist

- Findable
  - Assign persistent IDs
  - Provide rich metadata online
  - Register in a searchable resource, ...
- Accessible
  - Data online
  - Retrievable by their ID using a standard protocol
  - Metadata remain accessible even if data aren't...
- Interoperable
  - Use formal, broadly applicable languages
  - Use common (open) formats
  - Use standard vocabularies, qualified references...
- Reusable
  - Rich documentation, accurate metadata
  - Clear usage licences
  - Comprehensive provenance
  - Use of community standards...



How FAIR are your data?

Findable

It should be possible for others to discover your data. Rich metadata should be available online in a searchable resource, and the data should be assigned a persistent identifier.

- ☐ A persistent identifier is assigned to your data
- ☐ There are rich metadata, describing your data
- ☐ The metadata are online in a searchable resource e.g. a catalogue or data repository
- ☐ The metadata record specifies the persistent identifier

Accessible

It should be possible for humans and machines to gain access to your data, under specific conditions or restrictions where appropriate. FAIR does not mean that data need to be open! There should be metadata, even if the data aren't accessible.

- ☐ Following the persistent ID will take you to the data or associated metadata
- ☐ The protocol by which data can be retrieved follows recognised standards e.g. http
- ☐ The access procedure includes authentication and authorisation steps, if necessary
- ☐ Metadata are accessible, wherever possible, even if the data aren't

Interoperable

Data and metadata should conform to recognised formats and standards to allow them to be combined and exchanged.

- ☐ Data is provided in commonly understood and preferably open formats
- ☐ The metadata provided follows relevant standards
- ☐ Controlled vocabularies, keywords, thesauri or ontologies are used where possible
- ☐ Qualified references and links are provided to other related data

Reusable

Lots of documentation is needed to support data interpretation and reuse. The data should conform to community norms and be clearly licensed so others know what kinds of reuse are permitted.

- ☐ The data are accurate and well described with many relevant attributes
- ☐ The data have a clear and accessible data usage licence
- ☐ It is clear how, why and by whom the data have been created and processed
- ☐ The data and metadata meet relevant domain standards

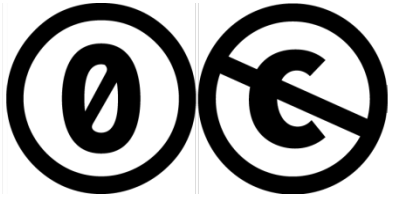
F A I R

Open science, all the way, but ...

Open Source https://github.com/...

Open Content

Open Data



Working in the open

Open roadmap

Open proposals

Participatory development

FAIR vs. Open data

- FAIR does not have to be open
- Data can be shared under restrictions & still be FAIR
- Making data FAIR ensures it can be found, understood and re-used
- Open data is a subset of all data schared

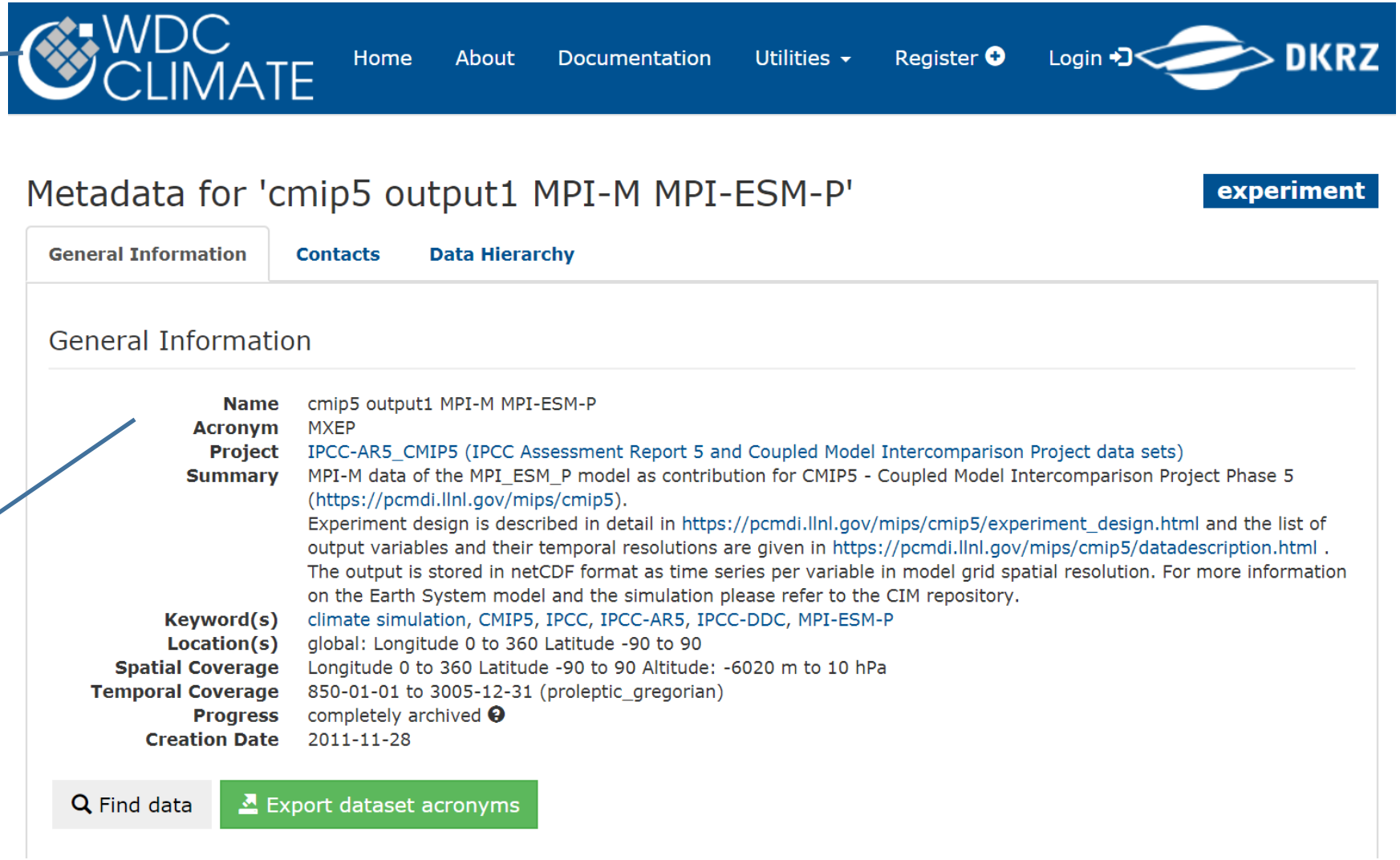
"As open as possiple, as closed as necessary"

Example for FAIR (meta)data : CMIP5 output

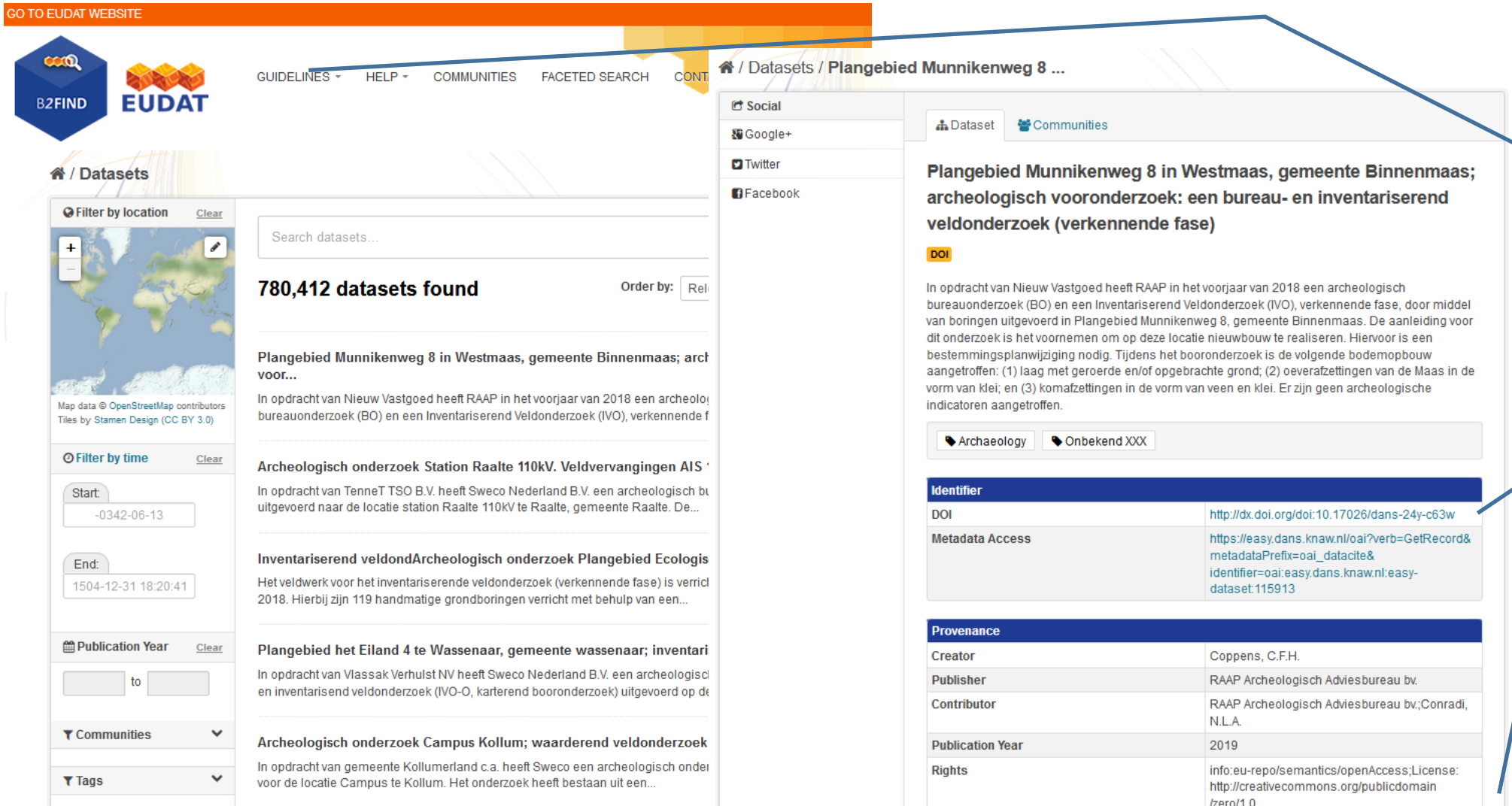
→ https://cera-www.dkrz.de/WDCC/ui/cerasearch/project?acronym=IPCC-AR5_CMIP5

F3 : Metadata indexed in a searchable resource
→ Here in the World Data Cliamte Center (WDCC) with the CERA search GUI

F2 : Data are described by rich MD
→ MD schema comprises 17? generic properties mapped to as completely as possible



Example for a FAIR friendly discovery service : EUDAT-B2FIND → http://b2find.eudat.eu



I1 : MD use a shared and broadly applicable language for knowledge
→ B2FIND MD schema is based on DataCite4.1.DOI OAI-PMH) → http://b2find.eudat.eu/guidelines/mapping.html#b2fmdschema

F3 : Metadata include the identifier of the digital object
→ At least one identifier (preferably a DOI) is mandatory

R1 : Rich MD describe data with relevant attributes
→ Usage Licence and Provenance information is provided

Projects driving FAIR and open data directives

- Research Data Alliance (https://rd-alliance.org)
  - RDA Vision : Vision: Researchers and innovators openly share data across technologies, disciplines, and countries to address the grand challenges of society.
- OpenAire (https://www.openaire.eu/)
  - Services to support FAIR Data from theory to implementation
- GO FAIR (https://www.go-fair.org) :
  - a bottom-up international approach for the practical implementation of the European Open Science Cloud (EOSC) as part of a global Internet of FAIR Data & Services
- FairIsFair (https://www.fairsfair.eu)
  - aims to supply practical solutions for the use of the FAIR data principles throughout the research data life cycle.

Issues and challenges

- FAIR means different things for different research areas
- Which actions are required by different stakeholders (EC, member states, disciplines, international) and when
- Structure and prioritise actions across key topics : Polica, skills, standards, infrastructuer, costs, rewards, social borders, …)
- Use these recommendations to develop a core FAIR Data Action Plan
- FAIR is meanwhile widely known and understood, but how to make own data FAIR is another issue …

Thank you for your attention!

Let me know what you think:

hwidmann@posteo.de