Marjan Grootveld
DANS / OpenAIRE

FAIR data in trustworthy repositories: the basics

International Open Access Week
October 2018

OpenAIRE supports

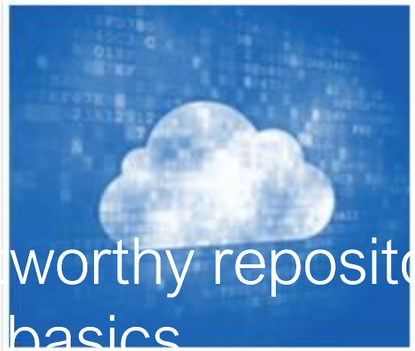
OS policies

- Harmonization for policy makers
- Training
- Support

Infrastructure

- Interoperability
- Setup
- Connectivity
- Repositories

Open Research Data



- FAIR
- Open data
- Tools
- Legal
- Compliance

OA to publications

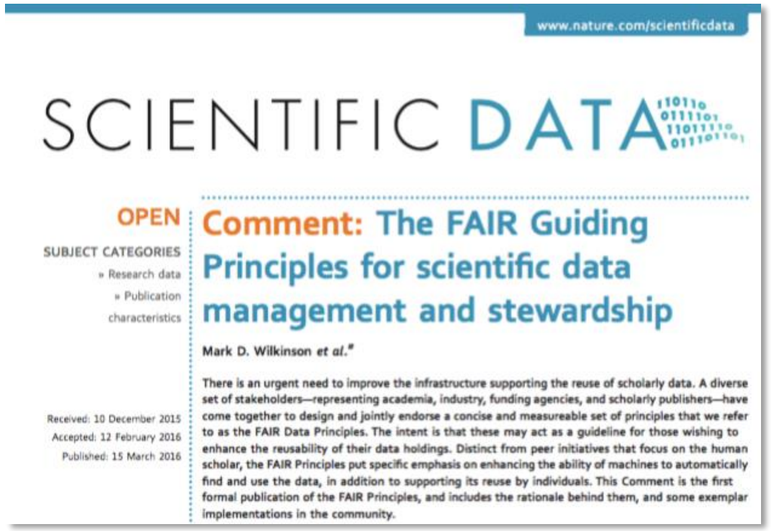
- Guides
- Tools/repositories
- Licenses
- Compliance

FAIR data in t[illegible]sworthy reposito[illegible]
[illegible] basics

https://www.openaire.eu/

FAIR data principles

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 licence and access conditions** (Open Access *when possible*), 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. Re-usable – Ready to be used for **future research** and to be processed further **using computational methods**.



- http://www.dtls.nl/fair-data/
- www.force11.org/group/fairgroup/fairprinciples
- http://www.nature.com/articles/sdata201618

Publish data in your own interest ;-)

arXiv.org > astro-ph > arXiv:1511.02512

Search or Ar

Astrophysics > Instrumentation and Methods for Astrophysics

The data sharing advantage in astrophysics

S. B. F. Dorch, T. M. Drachen, O. Ellegaard

(Submitted on 8 Nov 2015)

We present here evidence for the existence of a citation advantage within astrophysics for papers that link to data. Using simple measures based on publication data from NASA Astrophysics Data System we find a citation advantage for papers with links to data receiving on the average significantly more citations per paper than papers without links to data. Furthermore, using INSPEC and Web of Science databases we investigate whether either papers of an experimental or theoretical nature display different citation behavior.

Comments: 4 pages, 2 figures, Conference proceedings of Focus Meeting 3 on Scholarly Publication in Astronomy, IAU GA 2015, Honolulu

Subjects: **Instrumentation and Methods for Astrophysics (astro-ph.IM)**; Digital Libraries (cs.DL)

Cite as: arXiv:1511.02512 **[astro-ph.IM]**

(or arXiv:1511.02512v1 **[astro-ph.IM]** for this version)

FAIR data High-level Expert Group's recommendations

Rec. 10: Trusted Digital Repositories

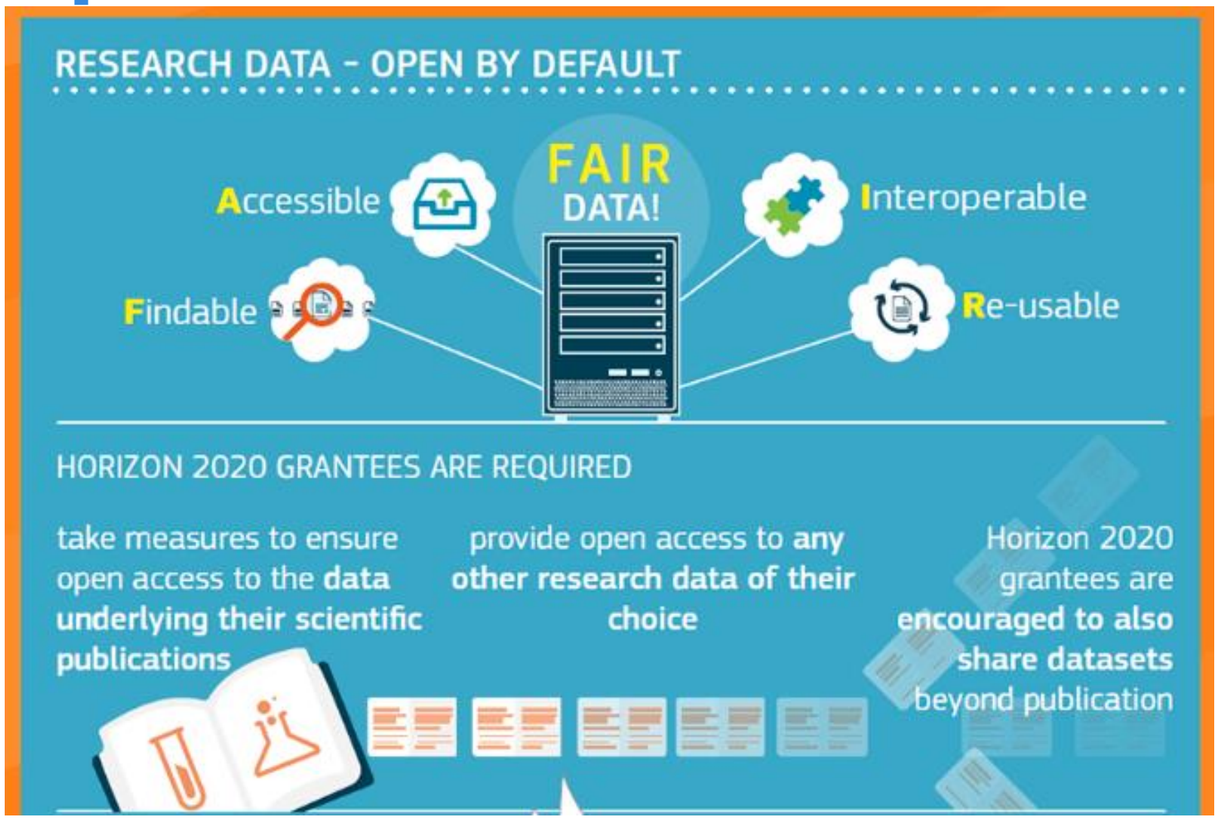
Repositories need to be encouraged and supported to achieve CoreTrustSeal certification. The development of rival repository accreditation schemes, based solely on the FAIR principles, should be discouraged.

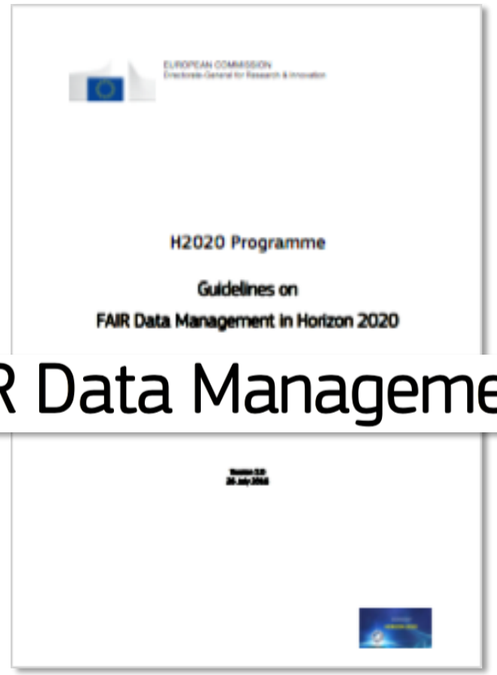
Rec. 29: Implement FAIR metrics

(…) Repositories should publish assessments of the FAIRness of datasets, where practical, based on community review and the judgement of data stewards.

FAIR data Expert Group presentation: https://www.slideshare.net/sjDCC/fair-data-interim-report-and-action-plan

Open and FAIR

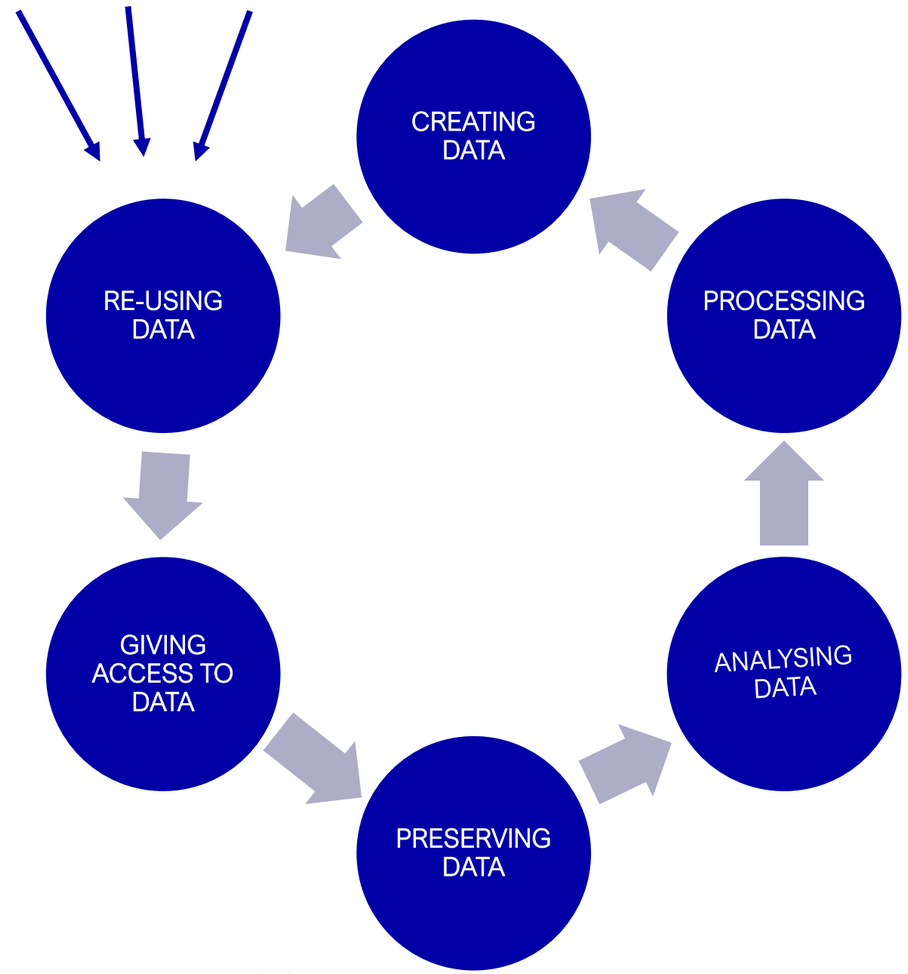




European Commission in the Guidelines: "Where will the data and associated metadata, documentation and code be deposited? Preference should be given to certified repositories which support open access where possible."

http://ec.europa.eu/research/images/infographics/policy/open-data-2016-w920.png
http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

Focus on re-using data



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.

- ☐ pen formats
- ☐ re used where possible
- ☐ l 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 license
- ☐ It is clear how, why and by whom the data have been created and processed
- ☐ The data and metadata meet relevant domain standards

Findable Accessible Interoperable Reusable

'How FAIR are your data?' checklist, CC-BY by Sarah Jones & Marjan Grootveld, EUDAT, https://doi.org/10.5281/zenodo.1065991 Image CC-BY-SA by SangyaPundir

"Lots of documentation is needed"

EUDAT FAIR checklist - CC-BY Sarah Jones & Marjan Grootveld, EUDAT.
https://doi.org/10.5281/zenodo.1065991

Metadata

- Metadata are needed to find the research data and get a first idea of the content.
- Use relevant community standards to enable interoperability.
- Check which standards the long-term repository supports or expects.

FAIRsharing.org
standards, databases, policies

https://fairsharing.org/

RDA
RESEARCH DATA ALLIANCE

DCC

http://rd-alliance.github.io/metadata-directory

https://rdamsc.dcc.ac.uk/
Extra: metadata tools: https://rdamsc.dcc.ac.uk/tool-index

Index of metadata tools

- AgriMetamaker
- ANZ-MEST (Metadata Entry and Search Tool)
- AVM Adobe Metadata Panels
- AVM Web Tool
- Bio-Formats
- CF Compliance Checker
- CIF2Cell
- CIM Comparator Tool
- CIM Questionnaire Generator
- CIM Viewer Tool
- CKAN
- CMOR (Climate Model Output Rewriter)
- Converis
- Darwin Core Archive Assistant
- Darwin Core Archive Validator
- Data Package libraries
- Data Package Validator
- Data Package Viewer
- Data Packagist
- DataCite Metadata Store API

OpenAIRE

European Commission

Documentation?

- Code book explaining the variables
- Study design
- Lab journal
- iPython or Jupyter notebook
- Statistical queries
- Software or instruments to understand or to reproduce the data
- Machine configurations
- Informed consent information
- Data usage licence
- …

In short: document and preserve everything that is needed to replicate the study – ideally following the standard in your discipline

Select a trustworthy repository

For giving (i.e. archiving & sharing) and taking (i.e. reusing) data:

- **Certified** as a 'Trustworthy Digital Repository'
- Matches your data needs regarding **file formats, access, licences**
- Supports **metadata standards**
- Provides a **persistent and globally unique identifier**
- Provides guidance on **data citation**
- Provides clear information about **costs** (if any)

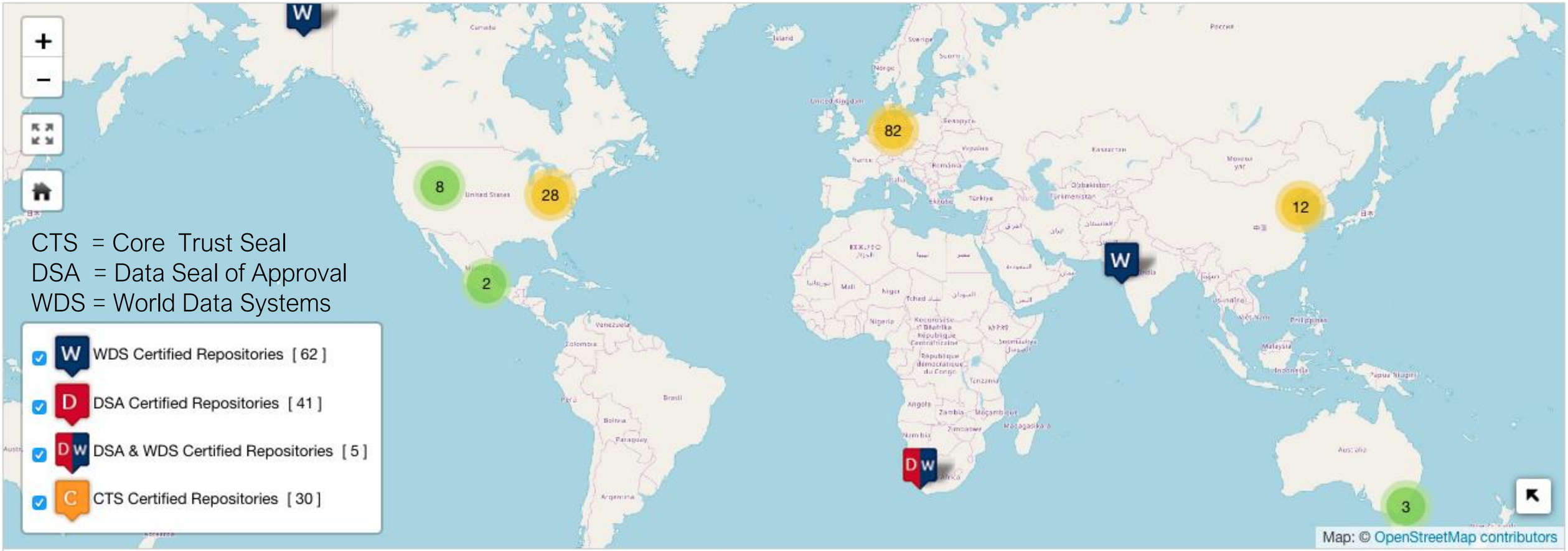
Contact your repository timely and benefit from its FAIR support.

https://www.coretrustseal.org/

Home | About | Certification | Certified Repositories | Apply | Contact

Core Certified Repositories

Home › Why certification › Core Certified Repositories



Part of the CoreTrustSeal requirements

R2. The repository maintains all applicable licenses covering data access and use and monitors compliance.

…

R10. The repository assumes responsibility for long-term preservation and manages this function in a planned and documented way.

…

R13. The repository enables users to discover the data and refer to them in a persistent way through proper citation.

R14. The repository enables reuse of the data over time, ensuring that appropriate metadata are available to support the understanding and use of the data.

…

But there is more, and you – as data producer or data consumer – play a role in it too!

https://www.coretrustseal.org/

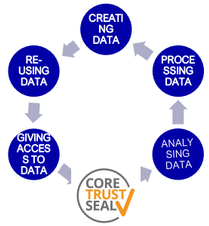
Assessing the FAIRness of data

- **Making data FAIR is essential.**
- **But how? For instance by inspecting how FAIR *existing* datasets are:**
  - Go to a data repository and look at a dataset:
  - Would you feel comfortable with reusing the data?
  - Why? Why not? What would help you to trust these data?

- **How can we measure this in a more structured way?**
  - Prototypes for FAIR data metrics are being developed.
  - Who should do the assessment?
    - Researchers from the data producer's domain?
    - Actual re-users?
    - Data repository staff?
    - A machine?

Photo by Yuri Catalano – CC0 - https://www.pexels.com/photo/city-landscape-sky-people-127420/

It's all about TRUST!



- Aim for a FAIR-aligned research data lifecycle.
- Credit researchers and others who seek value in and add value to existing data.
- Support FAIR and Open data in trustworthy repositories, for instance by
  - sticking to standards for data documentation and file formats ("replication packages");
  - checking how FAIR existing data are and learning from this;
  - teaching early-career researchers to manage data to make them FAIR and as open as possible.

More information:
OpenAIRE/EOSC-hub webinar: https://www.openaire.eu/how-to-manage-your-data-to-make-them-open-and-fair
OpenAIRE webinar on Open Research Data in Horizon 2020: https://www.openaire.eu/open-access-to-publications-in-horizon-2020
EUDAT/OpenAIRE/DANS webinar: https://eudat.eu/events/webinar/fair-data-in-trustworthy-data-repositories-webinar

Thank you!

Marjan Grootveld

marjan.grootveld@dans.knaw.nl