Data Archiving and Networked Services

DANS

Landelijk Coördinatiepunt
Research Data Management

Research Data Management: Data reuse

Ellen Leenarts	DANS	@EllenLeen	
Marjan Grootveld	DANS	@MarjanGrootveld	
Margriet Miedema	LCRDM	@lcrdm	
Maaike Verburg	DANS	@MaaikeVerburg	#FAIRAwareTool

SURF Bootcamp, 15 November 2021

Data reuse in six steps

1. 13.40 - 14.00 Reusing data: a data user's perspective of FAIRness - Ellen
2. 14.00 - 14.20 From depositing to discovering data - Marjan

14.20 - 14.30 Short break

3. 14.30 - 15.00 Find a repository - Ellen
4. 15.00 - 15.15 How repositories help to make and keep your data FAIR - Marjan

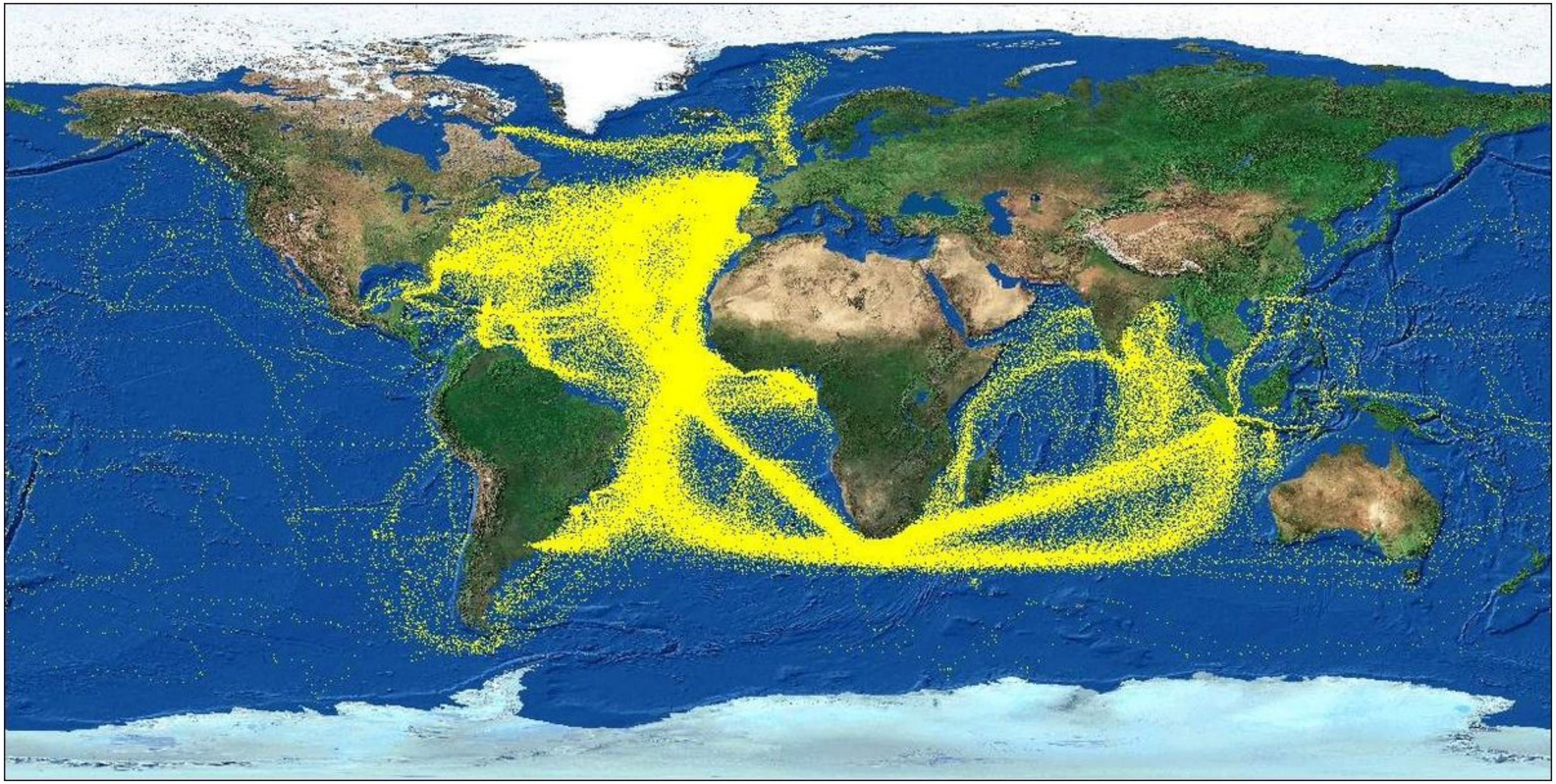
15.15 - 15.40 Break

5. 15.40 - 16.00 Measuring FAIR adoption, how do you do this? - Margriet
6. 16.00 - 16.40 Work with a FAIR tool - Maaike

Hands-on explorations: steps 1, 3, 6

This slideset contains slides step 1, 2, 3 and 4, the other slides (steps 5, 6) are provided separately

From “real life” to research data: CLIWOC - climatological database for the world’s oceans



Every yellow dot represents a ship report.

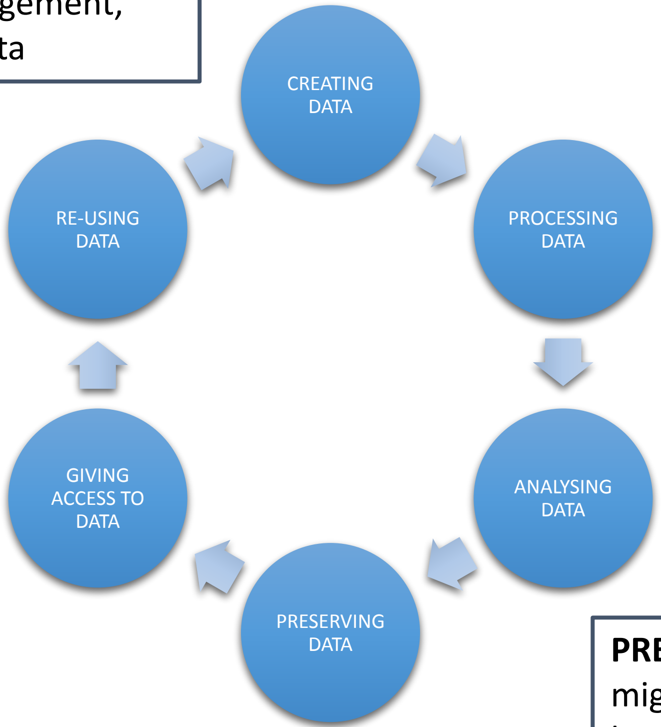
Image copied from https://www.knmi.nl/kennis-en-datacentrum/achtergrond/cliwoc

Project web site: http://pendientedemigracion.ucm.es/info/cliwoc/

Simplified research data lifecycle

CREATING DATA: designing research, DMPs, planning consent, locate existing data, data collection and management, capturing and creating metadata

RE-USING DATA: follow-up research, new research, undertake research reviews, scrutinising findings, teaching & learning



PROCESSING DATA: entering, transcribing, checking, validating and cleaning data, anonymising data, describing data, manage, store, back-up data

ANALYSING DATA: interpreting, & deriving data, producing outputs, authoring publications, preparing for sharing

ACCESS TO DATA: distributing data, sharing data, controlling access, establishing copyright, promoting data

PRESERVING DATA: data archiving, migrating to best format & medium for long term, creating metadata and documentation

Based on UK Data Archive lifecycle: https://www.ukdataservice.ac.uk/manage-data/lifecycle
Used in OpenAIRE RDM briefing paper: https://www.openaire.eu/briefpaper-rdm-infonoads

Hands on Reuse data (individual exercise)

Imagine you are considering the reuse of a dataset for your research.

Look here for a list of datasets and pick **one** and consider this dataset for reuse based on a couple of questions

You have **10 minutes** to investigate the dataset for reuse

Note your issues here

Thereafter we have reserved **5 minutes**

to share your insights

Examples of resources on reuse of data

Blog series (OpenAIRE project): https://www.openaire.eu/data-reuse-use-cases

Webinars (OpenAIRE project) on the reuse of data and legal aspects, 2020: https://www.openaire.eu/item/openaire-legal-policy-webinars

Webinars organised by UKDS (CESSDA partner) on the reuse of data, for example “Key issues in reusing data” (UK Data Service, 2020)

Event: organised by AUSSDA (CESSDA partner) this week: https://www.cessda.eu/Training/Event-Calendar/Finding-data-using-it-and-creating-new-knowledge-COVID-19-studies-at-AUSSDA

Past event organised by CESSDA but resources available: Challenges with the reuse of data on circular economy: Experience of researchers at the InnoRenew CoE https://doi.org/10.5281/zenodo.5609348

Dataset: Gregory, K.M. (2020): Data Discovery and Reuse Practices in Research. DANS. https://doi.org/10.17026/dans-xsw-kkeq

From depositing to discovering data

Most slides by Kathleen Gregory and Charlotte Glas (DANS)

SURF Bootcamp “Re-use”

15 November 2021

Core issue in research: Trust

Trust is a central element in the research world:

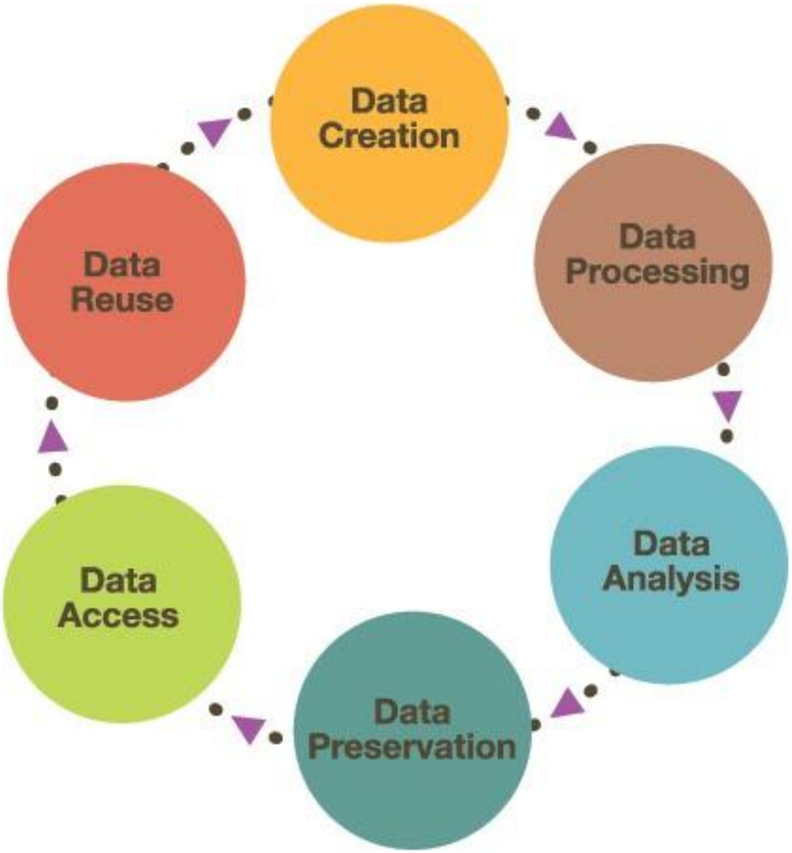
- How can I be sure that “they” interpret and use my data* in the right way?
- Where do these data come from? What has happened to them along the way?

* and other research output, such as software, workflows, … - although not discussed today

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

Data lifecycle

Data producer and data consumer meet at the repository.



Assumption: FAIRness helps to create trust

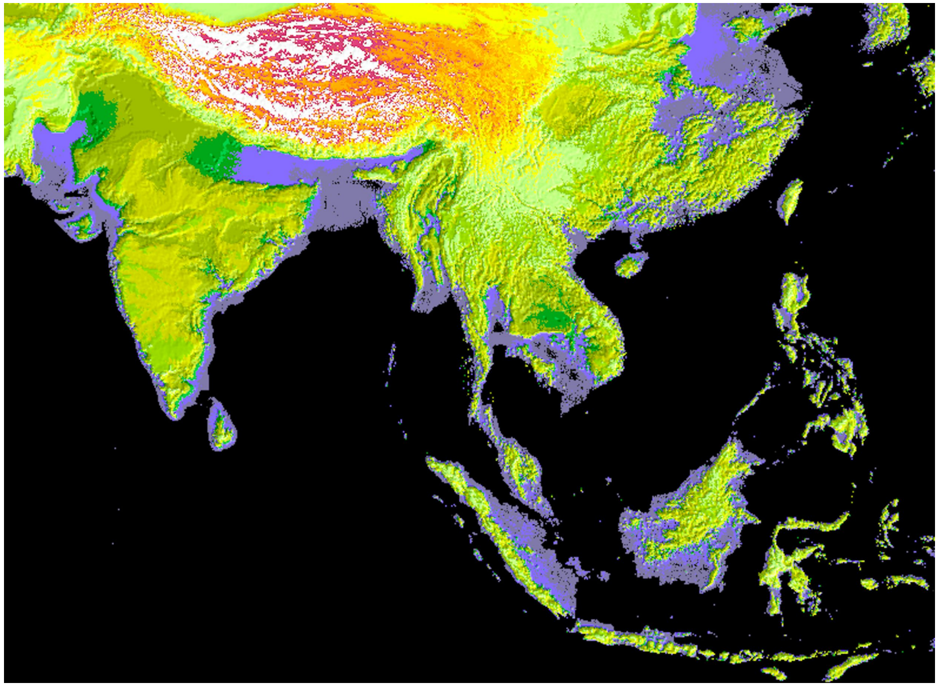
- When the data comes with documentation and information,
- in a usable and preferably open format,
- with a licence that makes explicit what you are allowed to do

… a potential re-user may become an actual re-user.

Data stewardship is about maximising the value of and trust in research data.

Findability

Findable?



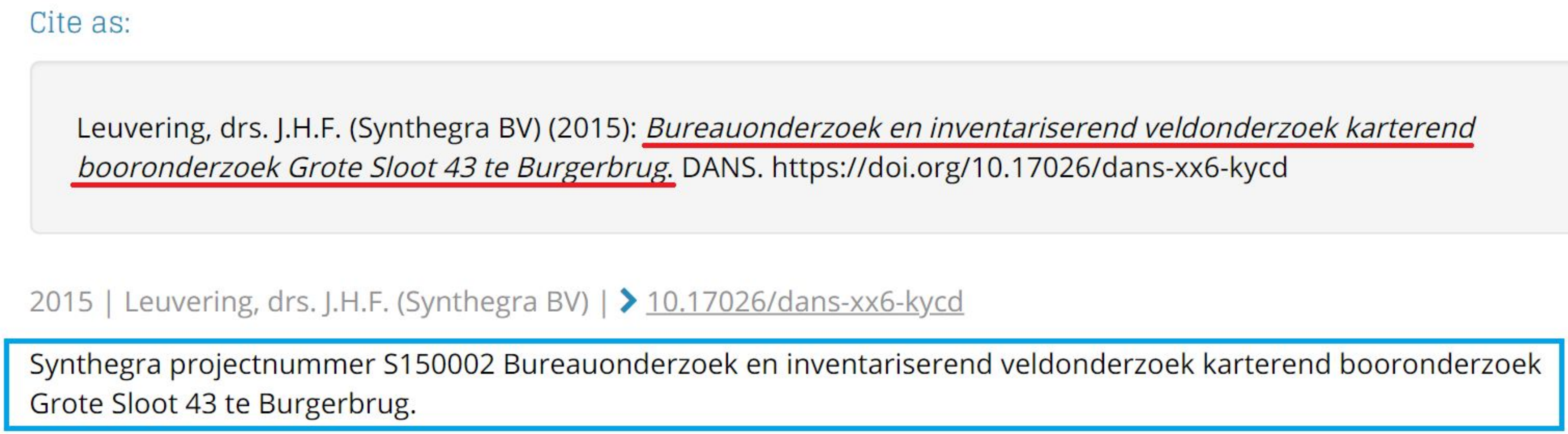
Not all data exist or are findable.

Not all data are described or made available in a standard way.

Image: NASA. Screenshot of map of Aceh area. https://search.earthdata.nasa.gov/search

Findable: minimal description

Description = almost identical to the title



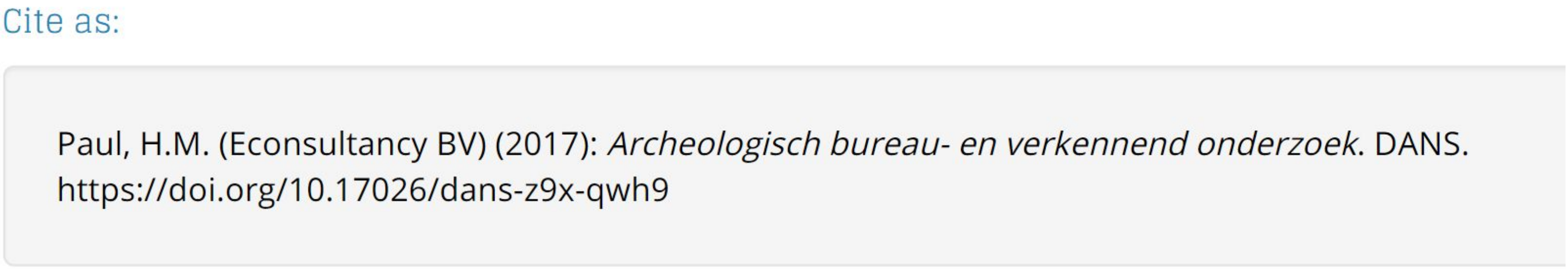
Cite as:

Leuvering, drs. J.H.F. (Synthegra BV) (2015): *Bureauonderzoek en inventariserend veldonderzoek karterend booronderzoek Grote Sloot 43 te Burgerbrug*. DANS. https://doi.org/10.17026/dans-xx6-kycd

2015 | Leuvering, drs. J.H.F. (Synthegra BV) | › 10.17026/dans-xx6-kycd

Synthegra projectnummer S150002 Bureauonderzoek en inventariserend veldonderzoek karterend booronderzoek Grote Sloot 43 te Burgerbrug.

Findable: short title

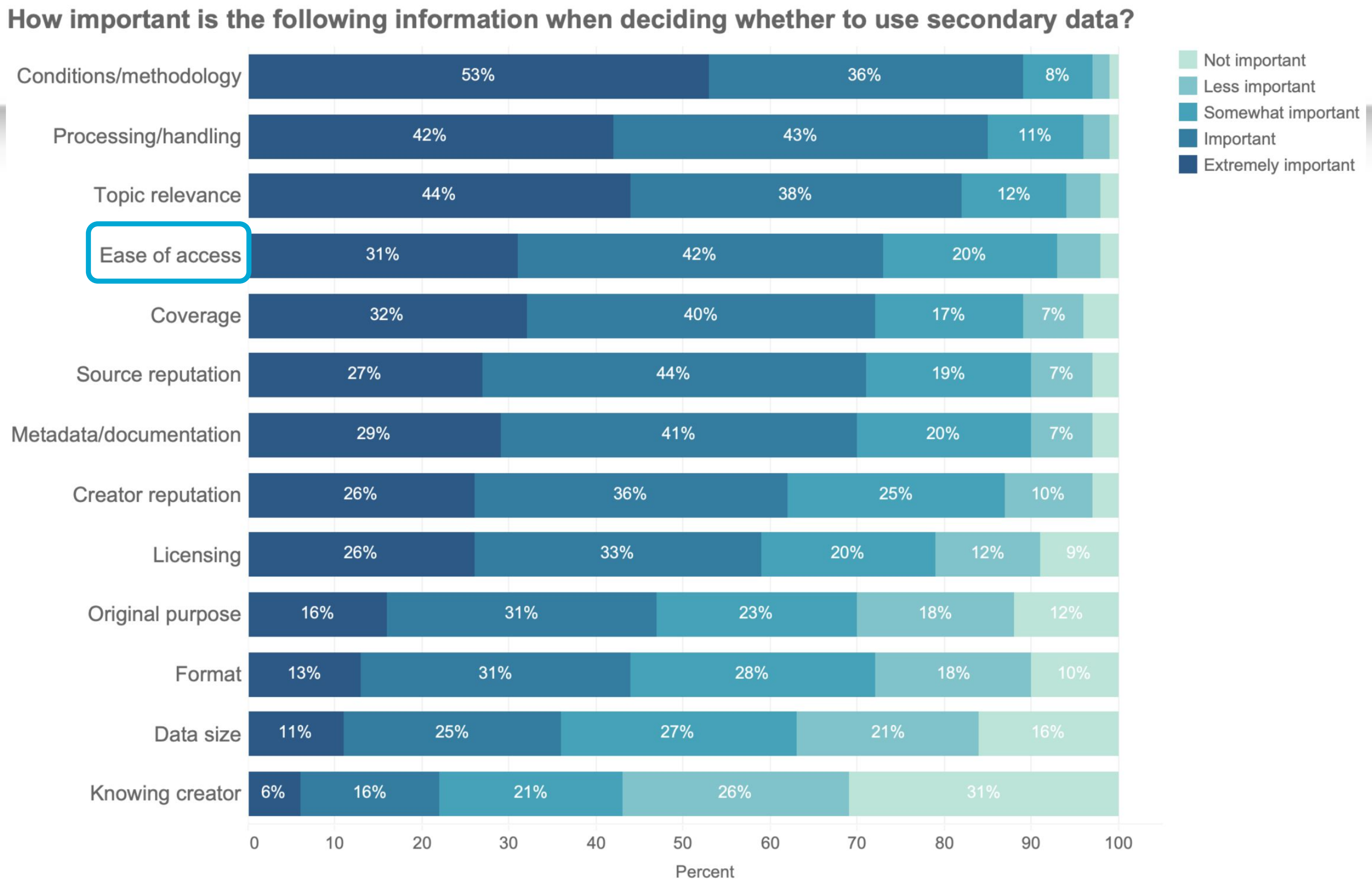


Cite as:

Paul, H.M. (Econsultancy BV) (2017): *Archeologisch bureau- en verkennend onderzoek*. DANS. https://doi.org/10.17026/dans-z9x-qwh9

Archeologisch bureau- en verkennend onderzoek -> too broad, there are thousands of this type of datasets within the DANS archive

Accessibility



Information used in evaluating data for reuse (n=1677). Percent denotes percentage of respondents.

Gregory, K., Groth, P. Scharnhorst, A., Wyatt, S. (2020). Lost or found? Discovering data needed for research. *Harvard Data Science Review*. https://doi.org/10.1162/99608f92.e38165eb

Accessible?

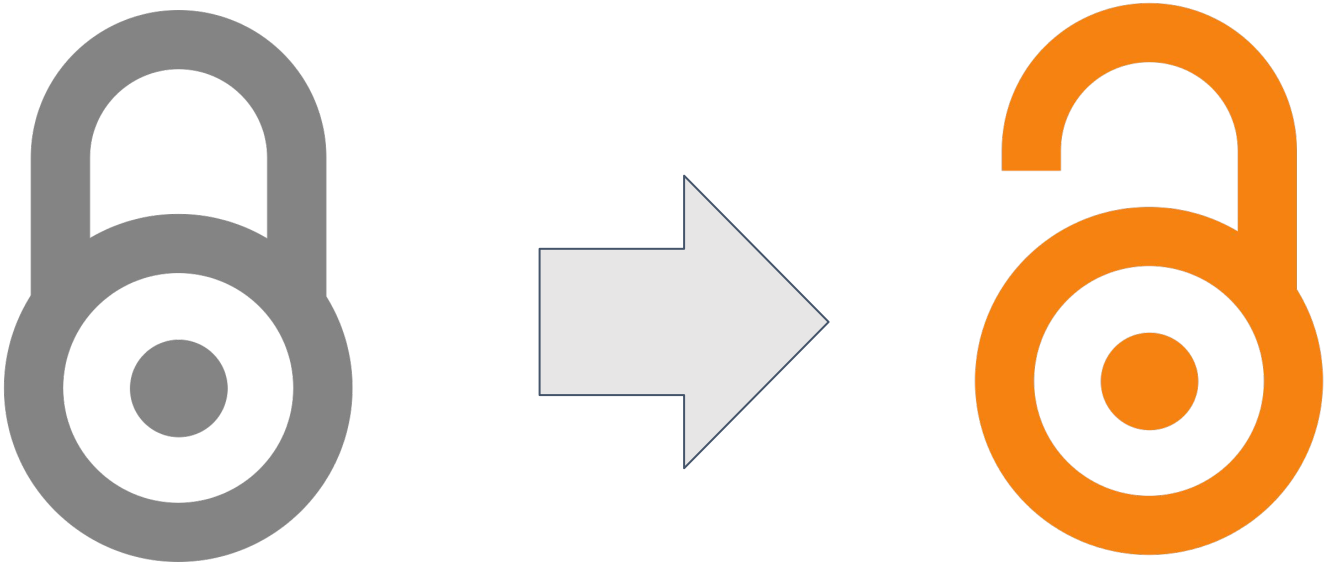
Accessing data is important in sensemaking, developing trust and understanding quality.

Accessing needed data is challenging.

Accessing metadata can be a starting point, even if data are not available.

Image: The fossils from the Cretaceous age found in Lebanon by Brocken Inaglory. Wikimedia Commons under CC BY-SA 3.0

Accessible - why restricted?



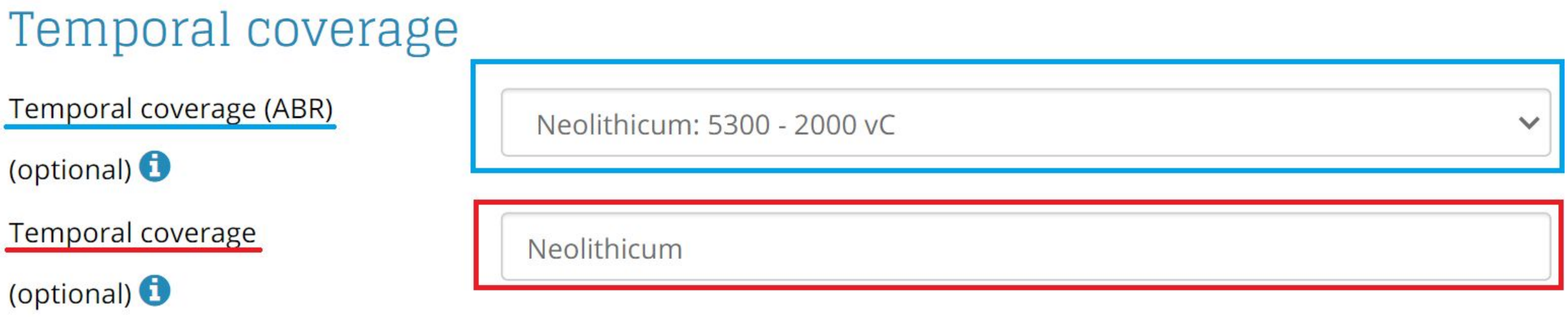
Interoperability

Interoperable?

Different sources for different data

For astronomical data, we have used mainly online virtual observatories. The general data (for teaching purposes), we've used digital/institutional repositories *(Respondent ID 305)*

Interoperable: controlled vocabulary



Temporal coverage

Temporal coverage (ABR)

(optional) ℹ

Neolithicum: 5300 - 2000 vC

Temporal coverage

(optional) ℹ

Neolithicum

Rather this than that

Reusability

Reusable?

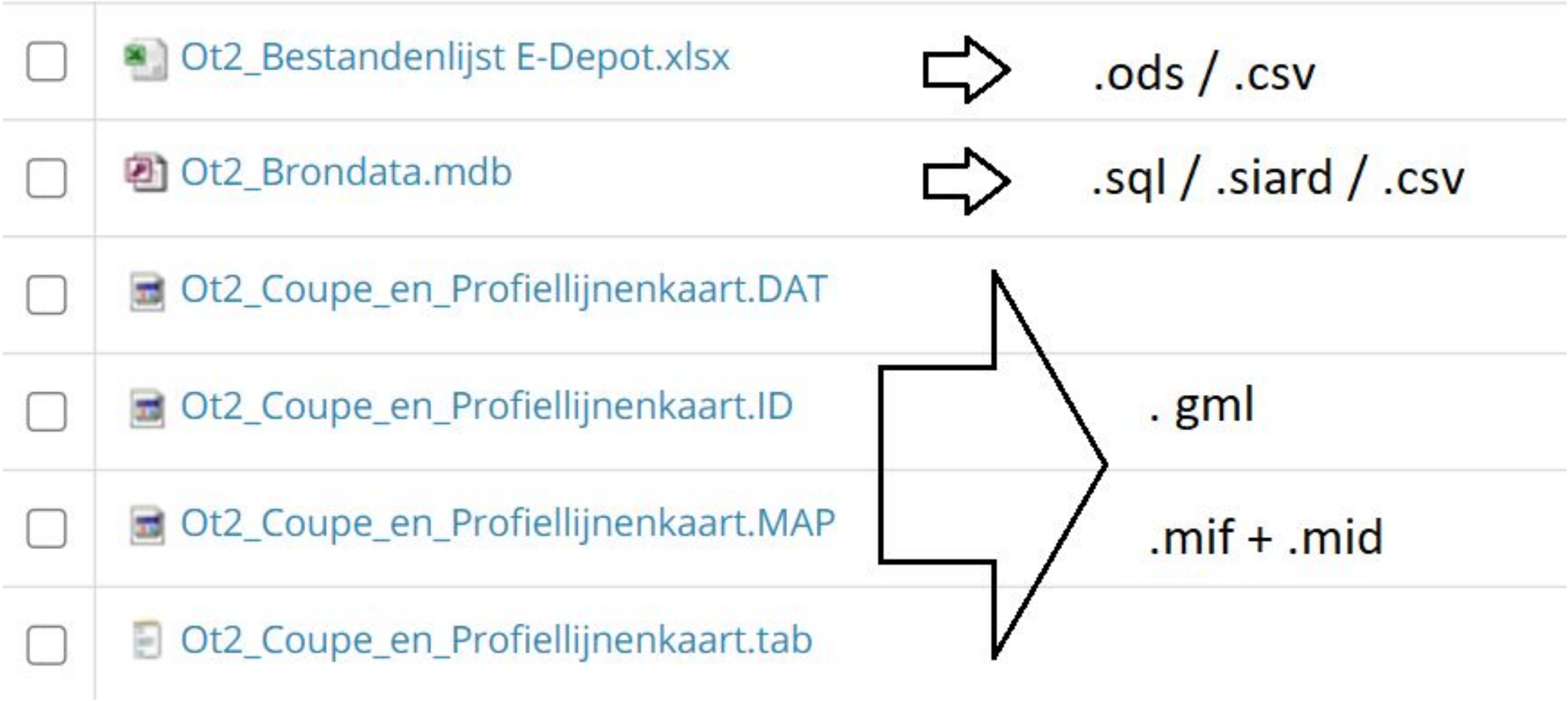
I am used to working with experts from different areas of knowledge. For me it is usual to have partners with different expertise: biology, agronomy, economy…I know the language of LCA (life cycle assessment), not of electronics or agricultural biology. ***My limit is not the data that I cannot find, but people that can work with these data*** *(Interview 16)***.**

Reusable?

Using open formats can help with reuse…

...and also facilitate collaboration.

Reusable: file formats



□ Ot2_Bestandenlijst E-Depot.xlsx → .ods / .csv

□ Ot2_Brondata.mdb → .sql / .siard / .csv

□ Ot2_Coupe_en_Profiellijnenkaart.DAT

□ Ot2_Coupe_en_Profiellijnenkaart.ID

□ Ot2_Coupe_en_Profiellijnenkaart.MAP

□ Ot2_Coupe_en_Profiellijnenkaart.tab

→ . gml

.mif + .mid

References

Koesten, L.*, Gregory, K.***,** Groth, P., Simperl, E. (2020). Talking datasets: Understanding data sensemaking behaviors. *International Journal of Human-Computer Studies*. https://doi.org/10.1016/j.ijhcs.2020.102562. ***Equal contributions**.

Gregory, K., Groth, P. Scharnhorst, A., Wyatt, S. (2020). Lost or found? Discovering data needed for research. *Harvard Data Science Review*. https://doi.org/10.1162/99608f92.e38165eb

Gregory, K., Cousijn, H., Groth, P., Scharnhorst, A., Wyatt, S. (2019). Understanding data search as a socio-technical practice. *Journal of Information Science*. https://doi.org/10.1177/0165551519837182

Gregory, K., Groth, P., Cousijn, H., Scharnhorst, A., Wyatt, S. (2019). Searching data: A review of observational data retrieval practices in selected disciplines. *Journal of the Association of Information Science and Technology*. https://doi.org/10.1002/asi.24165

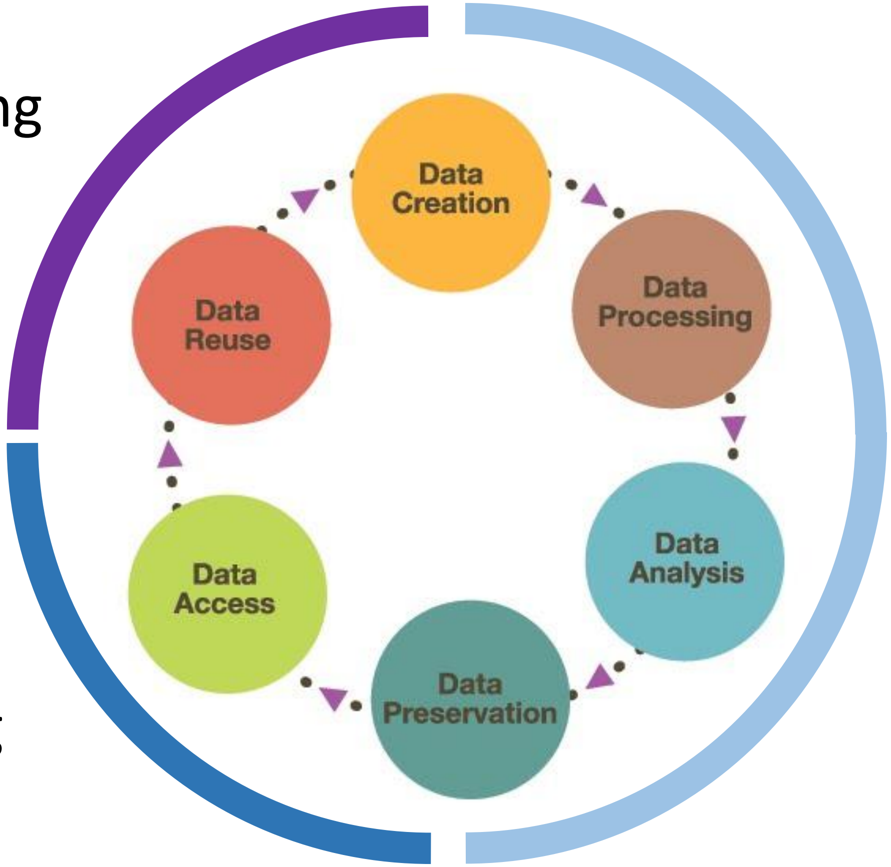
Gregory, K. (2021). *Findable and reusable? Data discovery practices in research.* PhD thesis. https://doi.org/10.26481/dis.20210302kg

DANS long-term data archive deposit information: https://dans.knaw.nl/en/depositing-data-manual/

Images without references were downloaded under a Pixabay license (free for reuse, without attribution) from pixabay.com

Short break
till 2.30 PM

Introduction to the hands on 'Find a repository'
Terminology … sharing versus archiving versus publishing



publishing

archiving

sharing

- Local systems
- Dropbox
- SURFdrive
- B2SHARE – EUDAT
- etc.

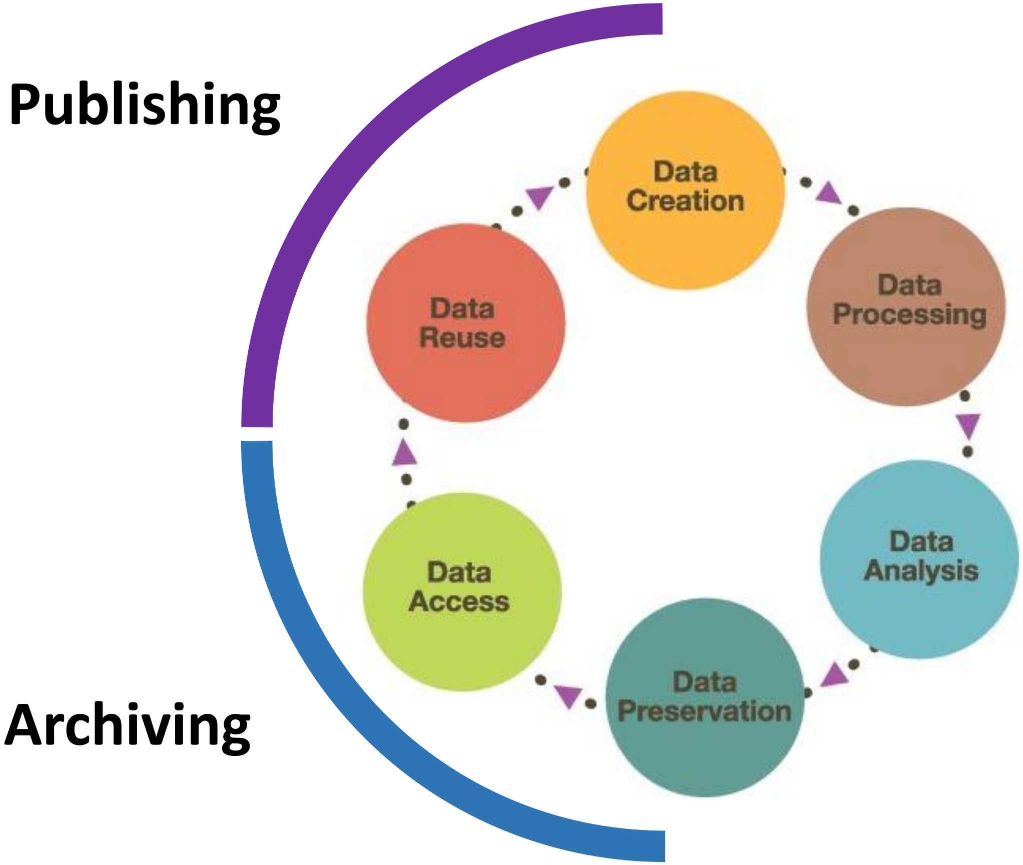
Owner in command
Flexible
But security is an issue...

The slides are collaborative work, for these on terminology, thanks Cees Hof

The difference …

Focus on visibility and accessibility of data and information.

Focus on long term preservation and retrievable data and information.



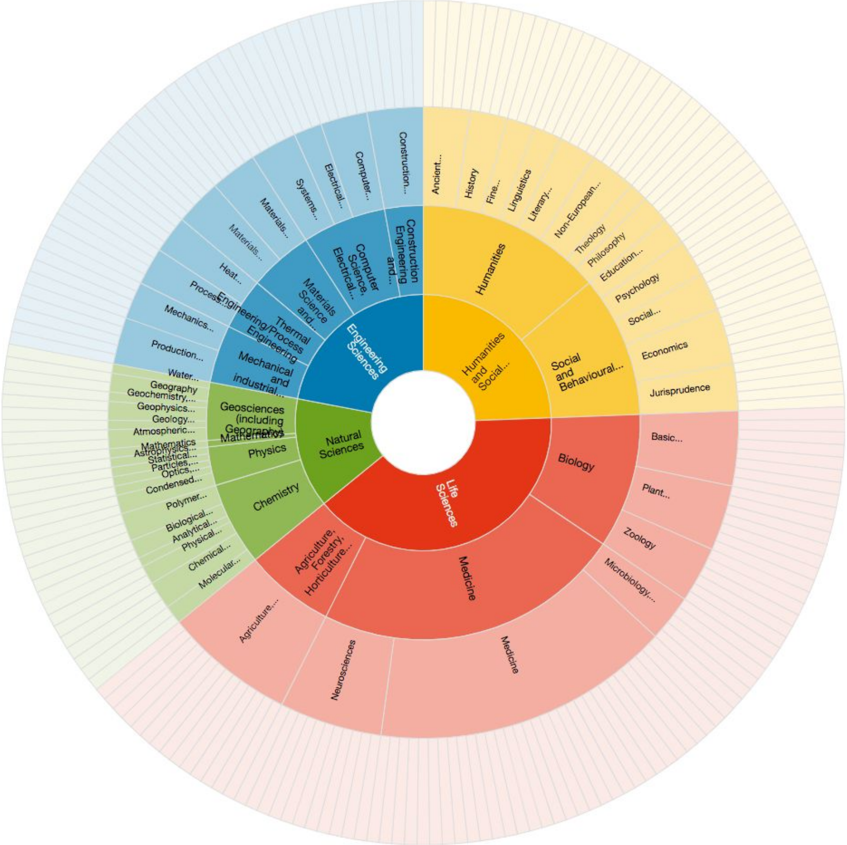
Let’s get to work….

Many Repositories available, see: **re3data.org**

https://www.re3data.org

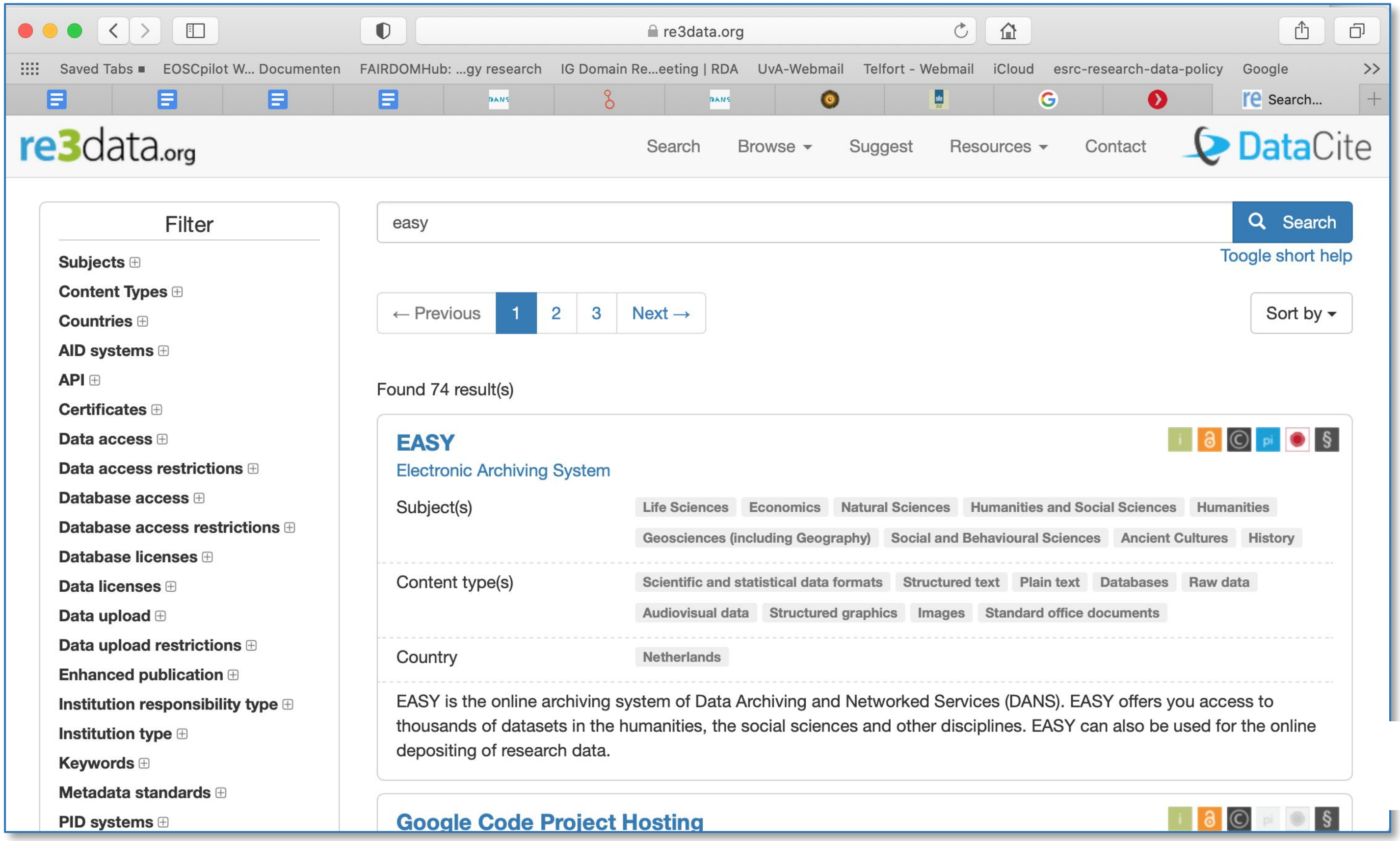
Over 2600 international repositories for a wide spread of domains

Plentiful filters to find appropriate repositories



Data Archiving and Networked Services

Let’s get to work…. re3data.org



easy

Found 74 result(s)

EASY

Electronic Archiving System

Subject(s)	Life Sciences Economics Natural Sciences Humanities and Social Sciences Humanities Geosciences (including Geography) Social and Behavioural Sciences Ancient Cultures History
Content type(s)	Scientific and statistical data formats Structured text Plain text Databases Raw data Audiovisual data Structured graphics Images Standard office documents
Country	Netherlands

EASY is the online archiving system of Data Archiving and Networked Services (DANS). EASY offers you access to thousands of datasets in the humanities, the social sciences and other disciplines. EASY can also be used for the online depositing of research data.

Google Code Project Hosting

For example, the DANS EASY repository in re3data.org

Icons and Tags provide info on characteristics of repositories.

Hands on: Use re3data to find a repository

First take 10 minutes to get to know http://www.re3data.org/ and familiarise yourself with a sample project that is described in the exercise instructions sheet

Follow the instructions and discuss your results in the break out groups. (15 minutes)

Discuss what you would do if there isn't an appropriate repository?

How to select a repository?

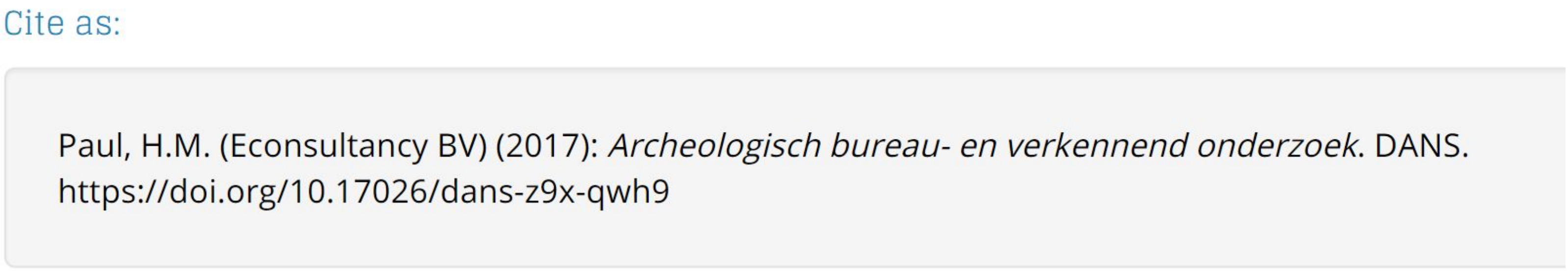
For giving (i.e. archiving & publishing) and for taking (i.e. reusing) data:

- Matches your particular data needs:
  - e.g. file formats accepted;
  - mixture of open and restricted access;
  - usage licences
- Gives your submitted dataset a persistent and globally unique identifier for sustainable (unique) citations and to link back to particular researchers and grants
- Provides guidance on how to cite the deposited data
- Certification as a ‘Trustworthy Data Repository’ with an explicit ambition to keep the data available for the long term

How repositories help to make data FAIR

Ellen’s slide: a good repository...

- Matches your particular data needs:
  - e.g. file formats accepted;
  - mixture of open and restricted access;
  - usage licences
- Gives your submitted dataset a persistent and globally unique identifier for sustainable citations and to link back to particular researchers and grants
- Provides guidance on how to cite the deposited data



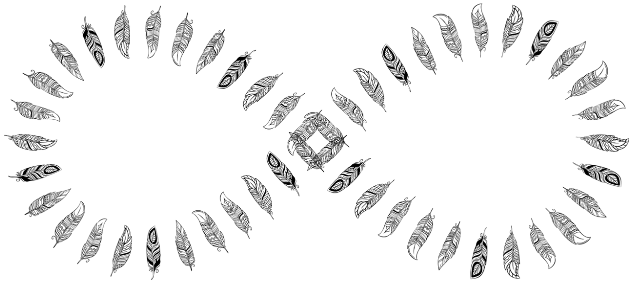
Cite as:

Paul, H.M. (Econsultancy BV) (2017): *Archeologisch bureau- en verkennend onderzoek*. DANS. https://doi.org/10.17026/dans-z9x-qwh9

But trust requires FAIRness over time

Recall the shipping log example: data should *remain* FAIR.

The FAIR principles don’t mention “long time” or “data preservation”, but the initial paper does.



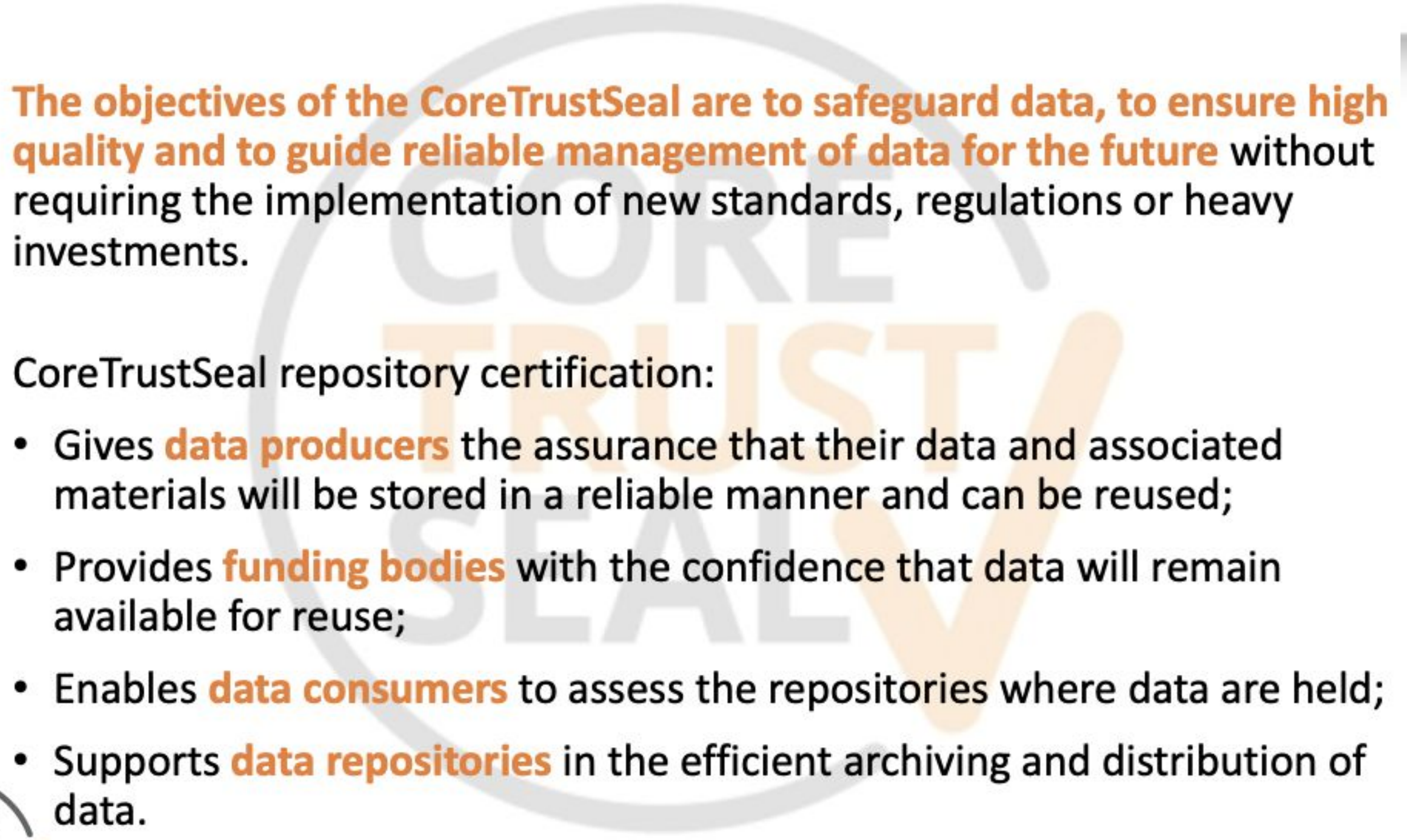
Trustworthy data repositories (TDRs) have a long-time mission.

How to recognise trustworthiness? By CoreTrustSeal
https://www.coretrustseal.org/

- ~ 175 certified repositories worldwide
- 18 in The Netherlands

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).

CoreTrustSeal



The objectives of the CoreTrustSeal are to safeguard data, to ensure high quality and to guide reliable management of data for the future without requiring the implementation of new standards, regulations or heavy investments.

CoreTrustSeal repository certification:

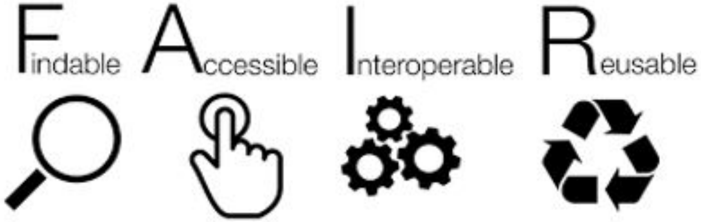
- Gives **data producers** the assurance that their data and associated materials will be stored in a reliable manner and can be reused;
- Provides **funding bodies** with the confidence that data will remain available for reuse;
- Enables **data consumers** to assess the repositories where data are held;
- Supports **data repositories** in the efficient archiving and distribution of data.

Requirements: background

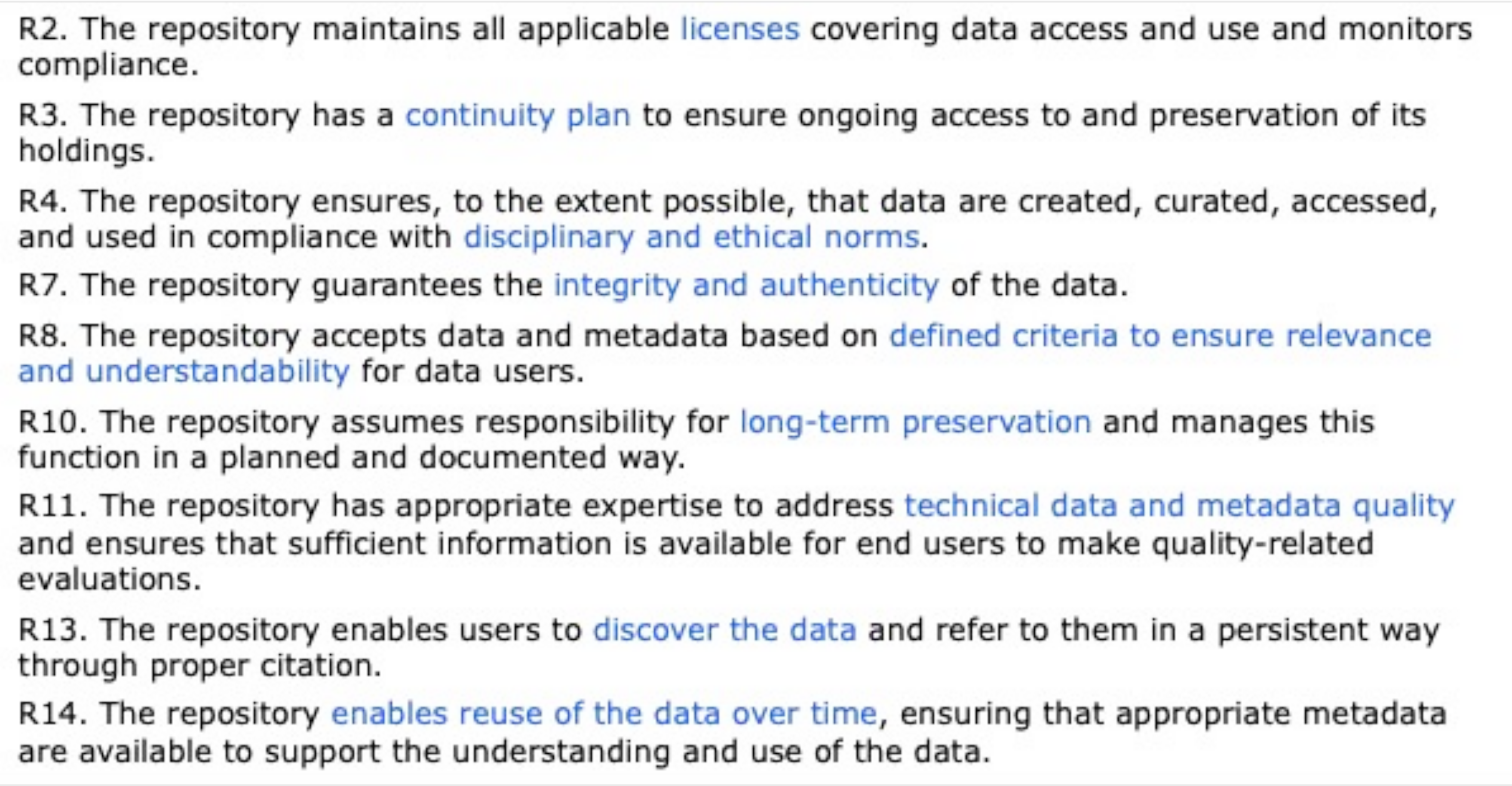
Fundamental to the requirements are five criteria that together determine whether or not the digital data may be considered as sustainably archived:

- The data can be found on the Internet;
- The data are accessible, while taking into account relevant legislation with regard to personal information and intellectual property;
- The data are available in a usable format;
- The data are reliable;
- The data can be referred to (persistent identifiers).

→ Strong link with:



Main CoreTrustSeal requirements



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

R3. The repository has a continuity plan to ensure ongoing access to and preservation of its holdings.

R4. The repository ensures, to the extent possible, that data are created, curated, accessed, and used in compliance with disciplinary and ethical norms.

R7. The repository guarantees the integrity and authenticity of the data.

R8. The repository accepts data and metadata based on defined criteria to ensure relevance and understandability for data users.

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

R11. The repository has appropriate expertise to address technical data and metadata quality and ensures that sufficient information is available for end users to make quality-related evaluations.

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.

CORETRUSTSEAL— FAIR ALIGNMENT

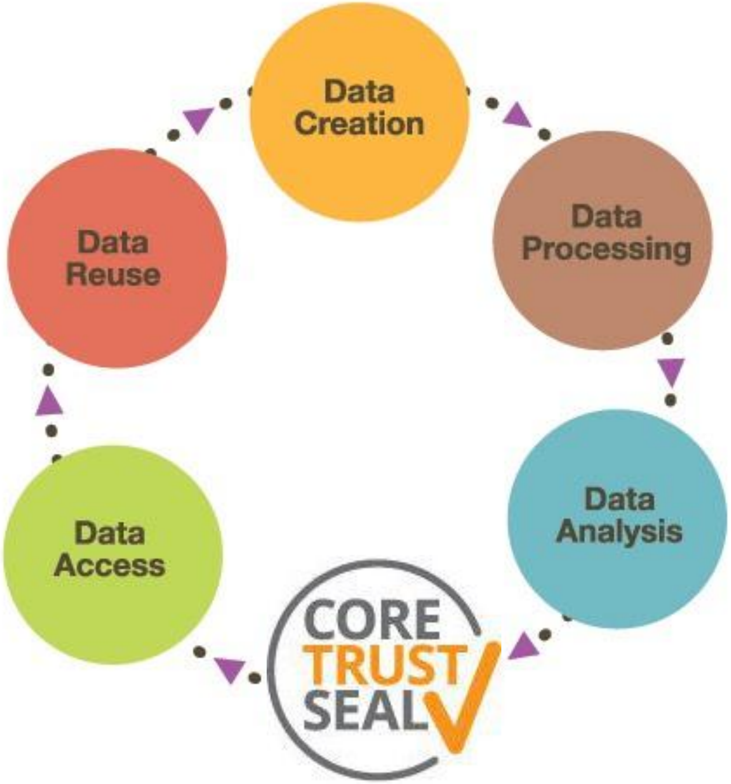
F	R13	• Offer persistent identifiers **[F1 and F3]** • Recommended data citations **[F1]** • Searchable metadata catalogue to appropriate standards **[F2, F3]** • Search facilities, inclusion in disciplinary or generic registries of resources **[F4]**
A	R4 R10 R13 R15 R16	• Facilitate machine harvesting of the metadata **[A1]** • Uses international and/or community standards **[A1.1]** • Searchable metadata catalogue to appropriate standards **[A1 and A1.1]** • Technical infrastructure: protection of facility, data, products, services, users **[A1.2]** • Data managed in compliance with discipline and ethical norms **[A1.2]** • Responsibility for long-term preservation **[A2]**
I	R14 R11	• Metadata required when the data are provided **[I1]** • Formats used by the Designated Community **[I1]** • Measures and plans for the possible evolution and migration of formats **[I2]** • Ensure understandability of the data **[I2]** • Ability to comment on, and/or rate data and metadata **[I3]** • Provide citations to related works or links to citation indices **[I3]**
R	R2 R7 R8 R11	• Integrity and authenticity of the data **[R1]** • Documentation of the completeness of the data and metadata **[R1]** • Links to metadata and to other datasets **[R1]** • Provenance data and related audit trails **[R1.2]** • Maintains licenses covering data access and use and monitors compliance **[R1.1]** • Defined data and metadata: ensure relevance and understandability for users **[R1.3]** • Technical data and metadata quality and assessment of adherence to schema **[R1.3]**

And other repositories?

Repositories without a trustworthy, long-term mission may have a simpler process for depositing and preserving data:

- typically, they don’t ask for preferred file formats – because they won’t convert or migrate the data to new formats in future (mere “bit preservation”);
- they may be less demanding (or helpful!) regarding metadata, and
- they won’t remind data producers to add documentation – which probably diminishes the interpretability and reusability of the data;
- they may not have long-term budget, qualified staff, appropriate technical infrastructure nor a continuity plan, should the organisation or the budget fail.

So…



- FAIR data live in Trustworthy Data Repositories

- CoreTrustSeal requirements are FAIR aligned

Data Archiving and Networked Services

DANS

Thank you!

Landelijk Coördinatiepunt
Research Data Management

Marjan Grootveld	marjan.grootveld@dans.knaw.nl
Ellen Leenarts	ellen.leenarts@dans.knaw.nl
Margriet Miedema	margriet.miedema@surf.nl
Maaike Verburg	maaike.verburg@dans.knaw.nl
	fair-aware@dans.knaw.nl