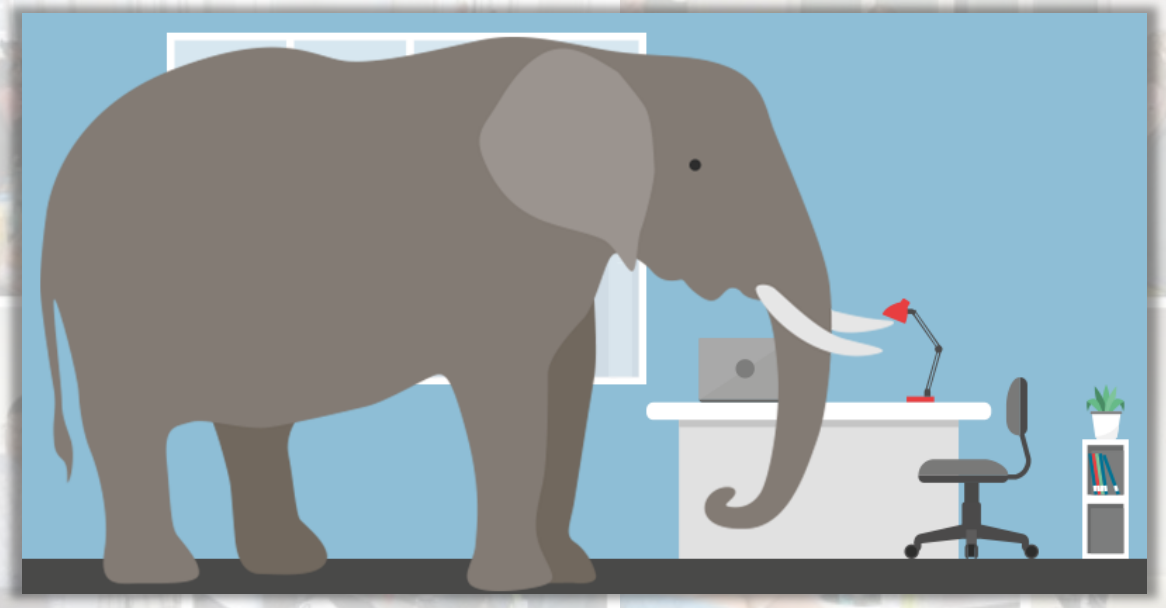
Data Management for Early Career Scientists – How to Tame the Elephant



Laia Comas-Bru [1] and Marcus Schmidt [2]

1. University of Reading, UK
2. University of Goettingen, Germany

Data Management can be overwhelming!

The World Data System (WDS) organised a **3-day EGU-sponsored workshop at Institut de Physique du Globe in Paris, France**, on current achievements and future challenges in Data Management in November 2019.

Purpose

To gain practical skills in data curation and management. Training took the form of l**ectures and group discussions, plus working on individual problems**. Although possible topics were suggested in advance, the final content of the training was tailored to the expectations stated in the registration forms.

Participants

23 Early Career Researchers and Scientists from

14 different countries.

Participants were invited to attend by the WDS Scientific Committee after selection from more than 100 applicants.

What are Research Data?

World Café

Data Management Plans

Research Data Management

Open Science and Data

How to Exchange Data

Copyright and Plagiarism

Ontologies and Knowledge Networks

Data Repositories

Big Data

Cloud Platforms in Science

The World Data System

WDS ECR Network

What are Research Data?

Data

« facts and statistics collected together for reference or analysis »

Notion of usefulness (What to keep?)

Dependant on the project involved but also on the scientific community practises

Research data?

- **Scientific Domain**
- **Origin** (e.g., academic, company, citizen science, mixed origin…)
- **Type:** primary data (raw), secondary data (easy to go back to raw data), data products (with significant processing), non-standard outputs (lab or field protocols, codes, etc)
- **Shape/format:** numerical data (audio-visual) or physical data (samples)
- **“Granularity”:** finite dataset (e.g. campaign) or time-series still running (e.g. observatories)

Research Data Management

To organise your work from data acquisition to publication and beyond
To enhance data re-use, data access, data interoperability

Save time and nerves for an efficient research process
Meet expectations and requirements of research funders and legislation

data management on a day-to-day basis will help you:

- Increase research efficiency: you can find, understand and use data when you need it
- Publish easily your data at the end of the project
- Protect your data against loss, deterioration or privacy and copyright breaches
- Save time and nerves

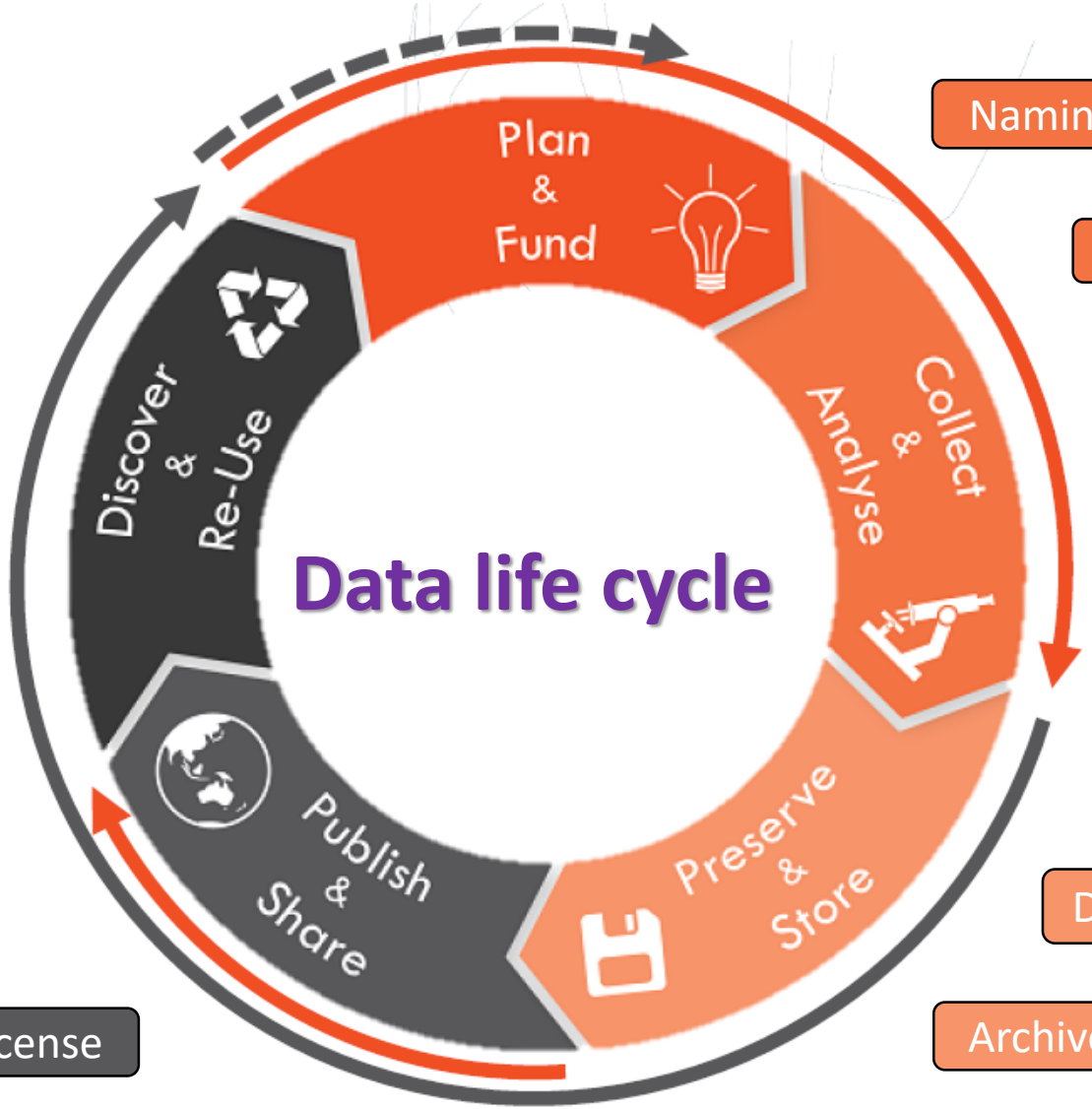
It will also help YOU:

- Find new collaborators, Return on investment
- Increase your reputation: shared data enhances research visibility and increases citations
- Reinforce your scientific integrity: helps to verify research findings over time and avoid accusation of fraud or bad science
- Meet requirements from funders and legislation

But it will also help SCIENCE:

- Others can reuse and build on your data
- Enhance new scientific approach: for education, Big data analysis, …
- Promote innovation and allow research in your field to advance faster

Research Data Management



DMP

Naming convention

Backup & Storage

Versioning

Metadata

Storage

Data formats

Archive

License

Data papers

Repository

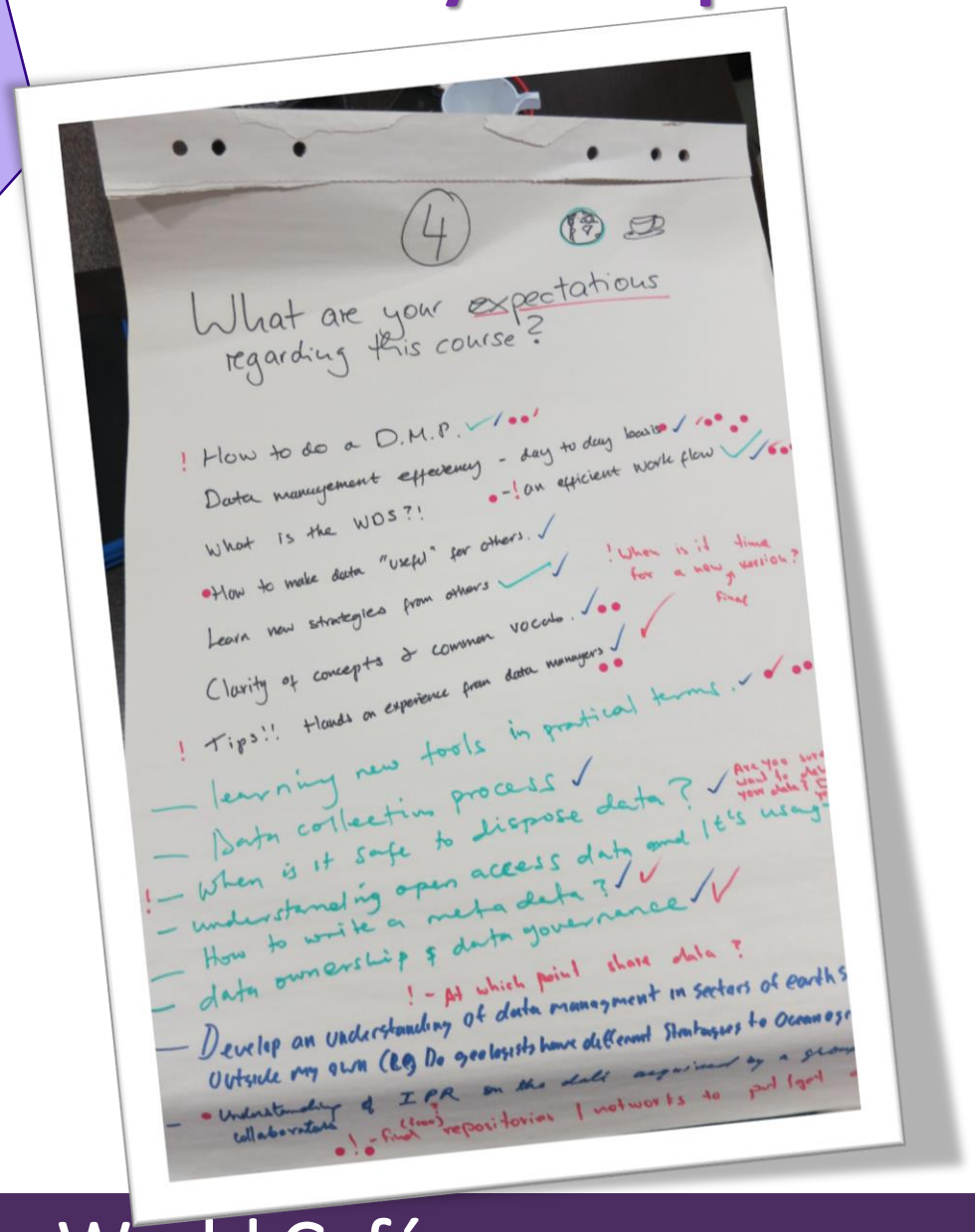
Persistent identifier

Metadata catalogue

Citation

World Café

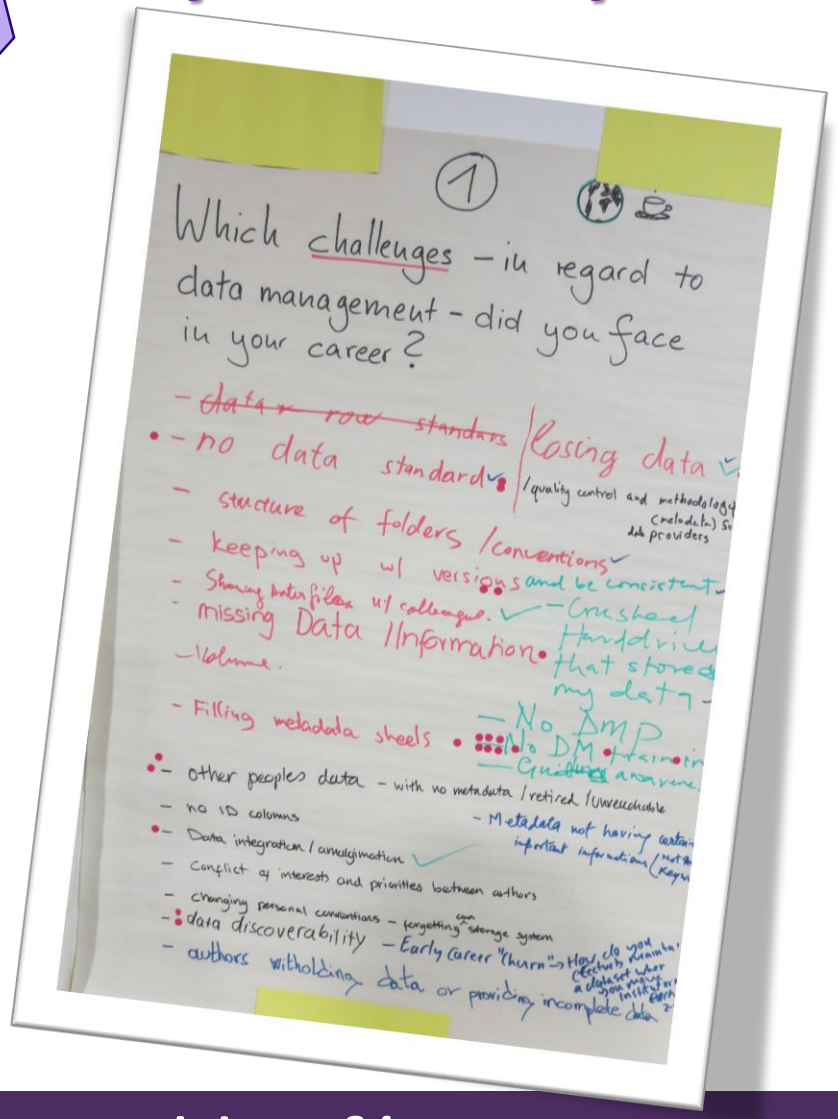
What are your expectations regarding this course?



- What is and how to use a data management plan?
- Gain experience from data managers
- Learn new tools
- Understand open access
- Find data repositories

World Café

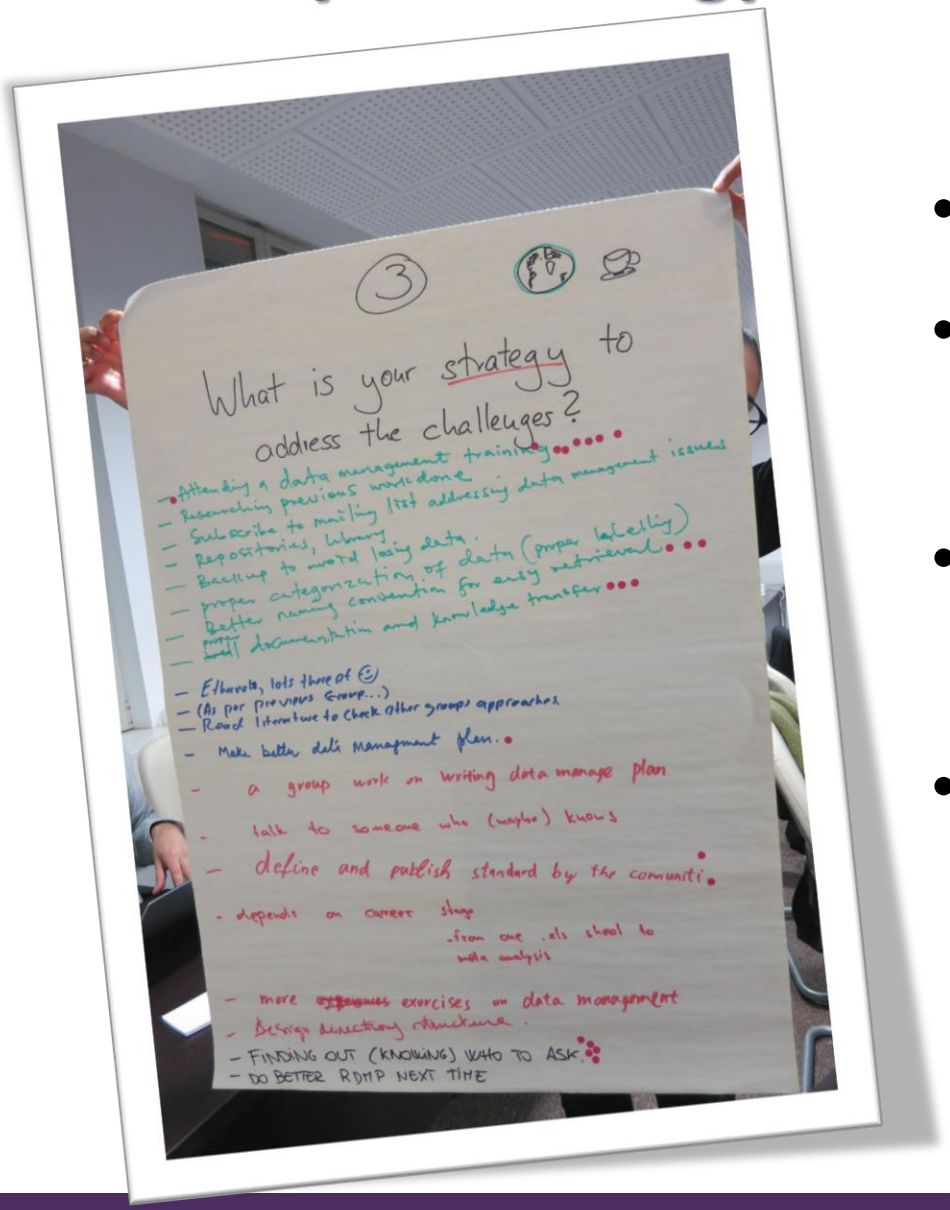
Which challenges – in regard to data management – did you face in your career?



- Losing data
- Keeping up with versions
- Data volume
- Working with data from others
- No access to some data

World Café

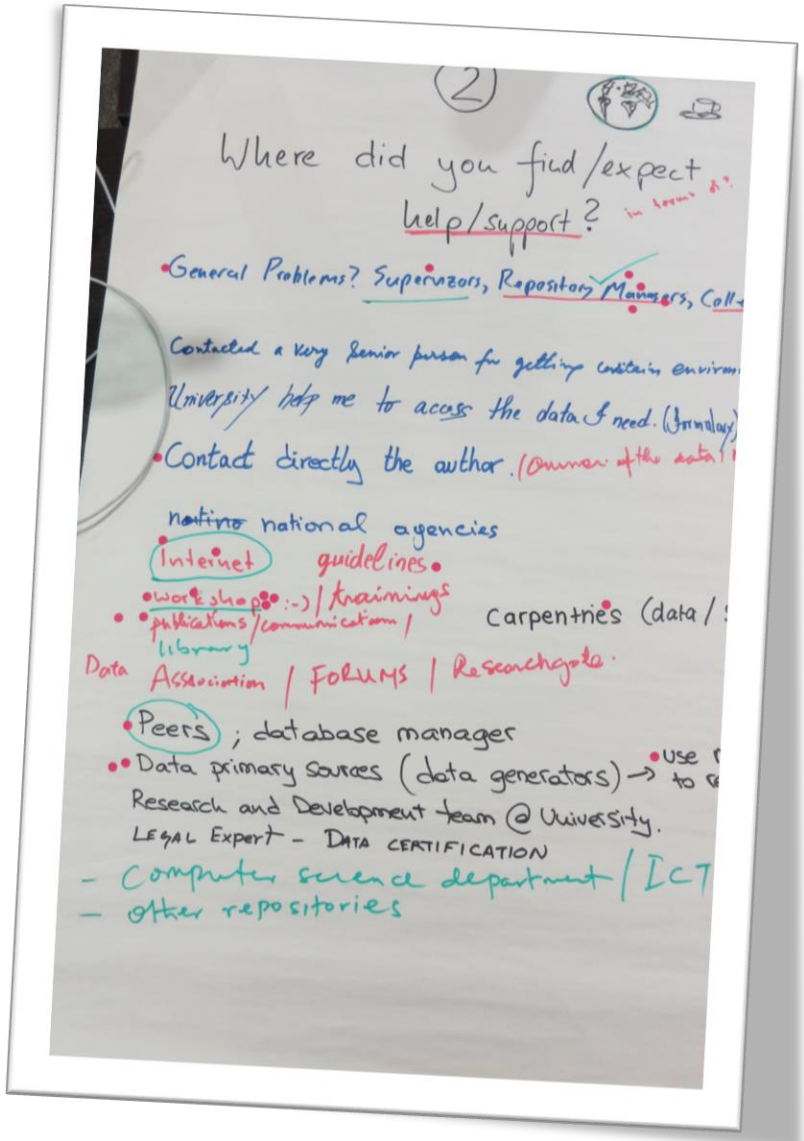
What is your strategy to address the challenges?



- Backing up data
- Documentations and knowledge transfer
- Knowing someone who knows
- Training on data management

World Café

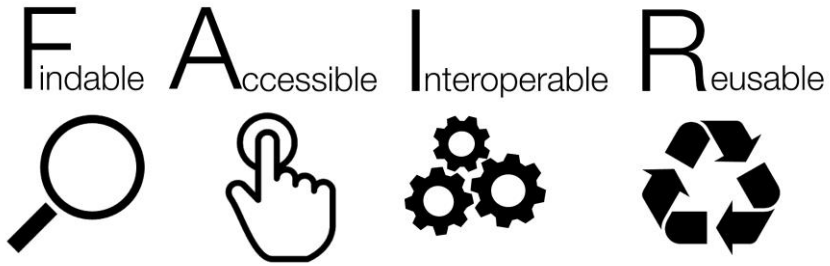
Where did you find/expect help/support?



- Supervisors, colleagues, repository managers
- Internets, workshops, trainings
- Data generators
- Computer Science department

Open Science and Data

Open data/open science is coming and we need to be prepared!



Open data/open science **is not going to be easy** (even the "experts" don't know how to deal with some issues) **and it will require time** (for management, not science)

Further reading: Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3,** 160018 (2016). https://doi.org/10.1038/sdata.2016.18

Adapted from Australian National Data Service website at http://ands.org.au/discovery/opendata.html

Data Management Plans

What is a DMP?

Document that outlines how data are to be handled both **during** a research project, and **after** the project is completed.

Source: http://www.elra.info/en/services-around-lrs/dmp/

A DMP describes:

- **What** data will be created
- **What** policies will apply to the data
- **Who** will own and have access to the data
- **What** data management practices will be used (e.g. storage, security, access controls)
- **Who** is responsible for data management activities (i.e. if you are collaborating with others)

Data Management Plans

5 tips to avoid the most common errors:

- **Describe precisely why data cannot be shared.** It is legitimate not to share certain data, for example, if it is sensitive or personal non-anonymizable data, or if data is subject to copyright restrictions.
- **If necessary, define a reasonable embargo period.** Typically, embargo periods of 6-12 months are acceptable.
- **Use existing metadata standards** to ensure that data are findable. A list of technical metadata standards is available at http://rd-alliance.github.io/metadata-directory/.
- **Keep it simple.** The DMP should be clearly structured.
- **Call for help**

Source: www.snf.ch

Useful links to create DMPs online:

DMP ONLINE
https://dmponline.dcc.ac.uk/

DMPTool
Build your Data Management Plan
https://dmptool.org/

DMP OPIDoR
https://dmp.opidor.fr/

How to Exchange Data

Lasting data formats

Recommendations:

- For text: PDF/A, plain text (*.txt) or XML
- For spreadsheets/tables: CSV (comma-separated values)
- DO NOT use Excel or Word files

Source: https://www.museum.ie/Archaeology/Exhibitions/Current-Exhibitions/Ancient-Egypt/Writing-Art

Tips for long-term storage

- Follow standards that are open and not proprietary
- Convert your files to lasting formats, but keep original files as well
- Check out DROID to identify file formats:

https://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/droid/

Copyright and Plagiarism

Copyrights:

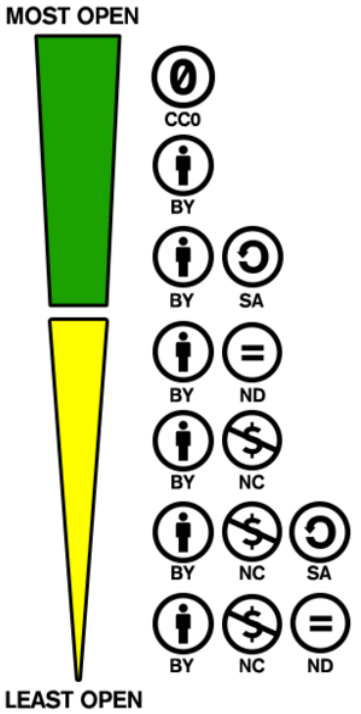
- Protect the expression of an idea, not the idea itself
- Prevent copying, adapting, distributing, performing and broadcasting to the public
- Exclusive right of the author

Licensing:

- Tells precisely what can be done with data
- Encourages reuse
- Creates visibility

Very common: CC-BY

- Data can be reused, but needs to be attributed to the author



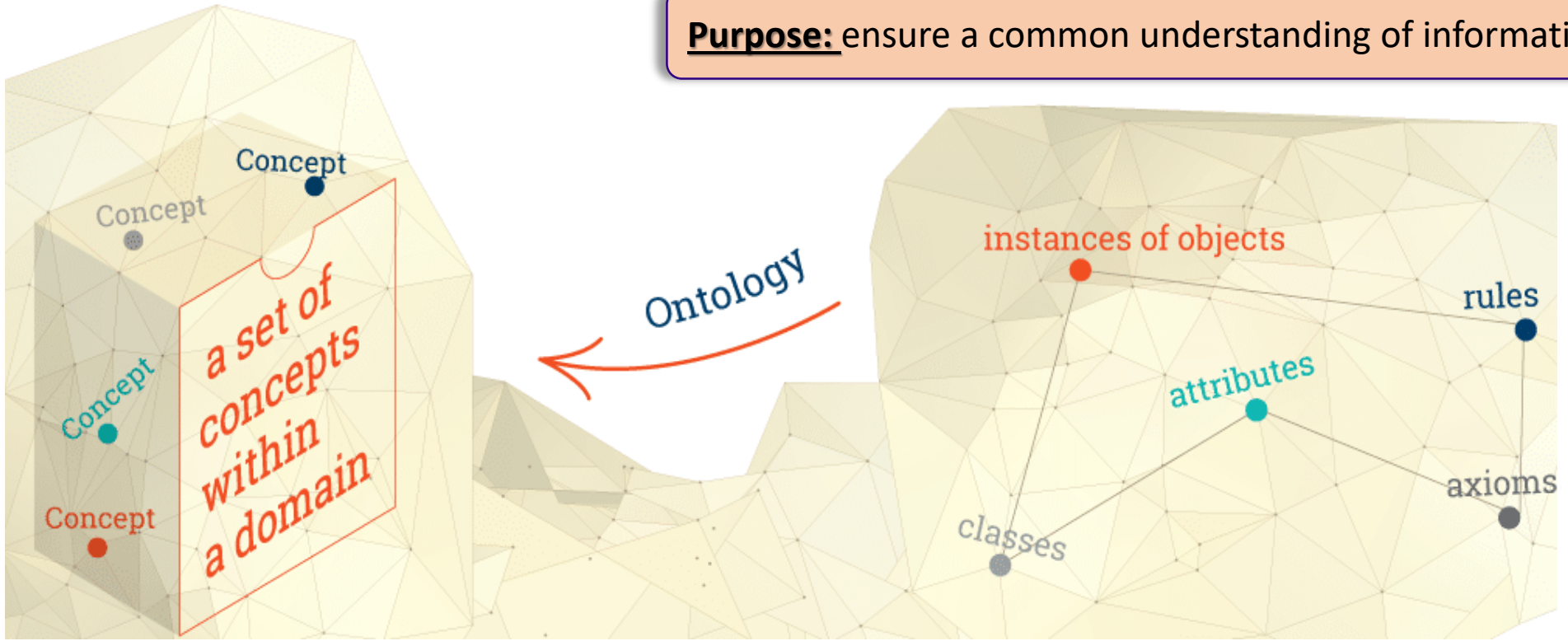
Source: https://en.wikipedia.org/wiki/Talk%3ACreative_Commons_license

Ontologies and Knowledge Networks

What is Ontology?

- A set of concepts and categories in a subject area or domain that shows their properties and the relations between them
- Ontologies introduce a **sharable and reusable knowledge representation** and can also add **new knowledge** about the domain

Purpose: ensure a common understanding of information

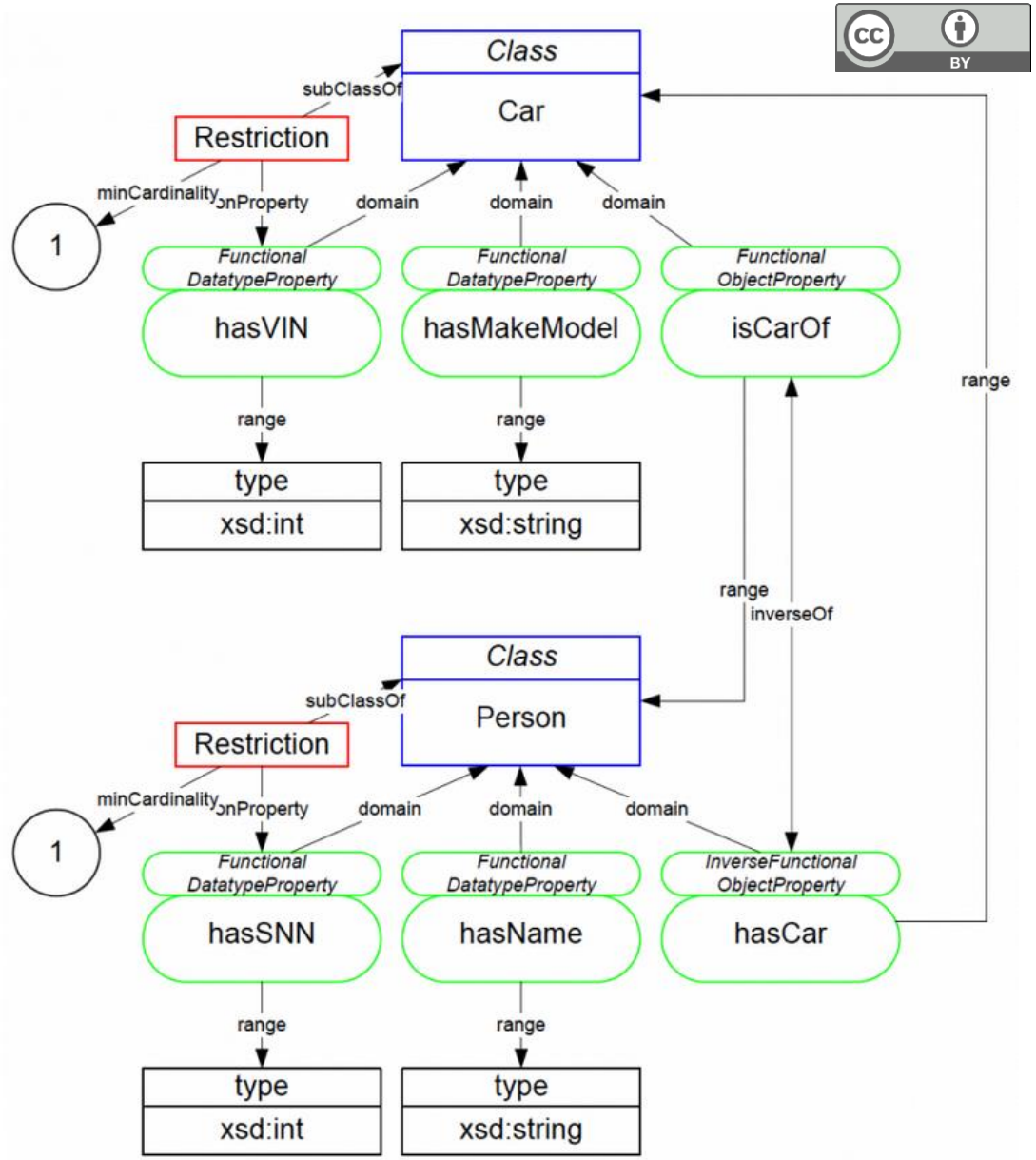


Source: https://www.ontotext.com/knowledgehub/fundamentals/what-are-ontologies

Ontologies and Knowledge Networks

Web Ontology Language (OWL)

- is a **semantic web computational logic-based language**, designed to represent rich and complex knowledge about things and the relations between them
- Many other languages: CASL, Common logic, Cyc, DOGMA, Gellsih, IDEF5, KIF, MOF, Olog, OBO, OntoUML, RIF, SADL, SBVR, TOVE



Source: http://myrosh.com/owl-web-ontology-language-overview/

Ontologies and Knowledge Networks

Further resources:

- What are ontologies? https://www.ontotext.com/knowledgehub/fundamentals/what-are-ontologies/
- Building ontologies: an introduction for engineers: https://www.youtube.com/watch?v=Gh0f2Us0hr0

Ontology libraries: Cupboard, BioPortal, Ontology Design Patterns, TONES, Schema, Biopax, SWEET,

Ontology repositories: https://www.w3.org/wiki/Ontology_repositories

Publications:

- Ashburner, M. et al. “Gene Ontology: Tool for the unification of biology. The Gene Ontology Consortium, 2000. 10.1038/75556
- Carbon, S. et al. “Gene Ontology Hierarchy: Based on the AmiGO, the GO Consortium's annotation and ontology toolkit”, 2009. 10.1093/bioinformatics/btn615
- Du, H. et al. “An Ontology of Soil Properties and Processes”, 2016. 10.1007/978-3-319-46547-0_4
- Ma, X., and Fox, P. “Recent progress on geologic time ontologies and considerations for future works”, 2013. 10.1007/s12145-013-0110-x

Data Repositories

Technical barriers to data sharing include….

- A system does not operate according to its objectives and specifications
- Datasets are not complete or include unintended modifications
- Datasets do not contain what they claim to contain
- Access to data and services is not guaranteed
- Datasets and services are not usable (for whatever reason)

Being unable to trust data from other sources is one of the **major challenges preventing proper data management, preservation, and sharing** (along with data ownership and the fear of being discredited or scooped)

Trustworthy Data Repositories:

- **Certification standards** play an important role in establishing trust, and hence sustaining the opportunities for long-term data sharing

Data Seal of Approval Certification of Trusted Data Repositories

WDS Certification of Regular Members

Research Data Alliance Repository Audit and Certification DSA–WDS Partnership WG

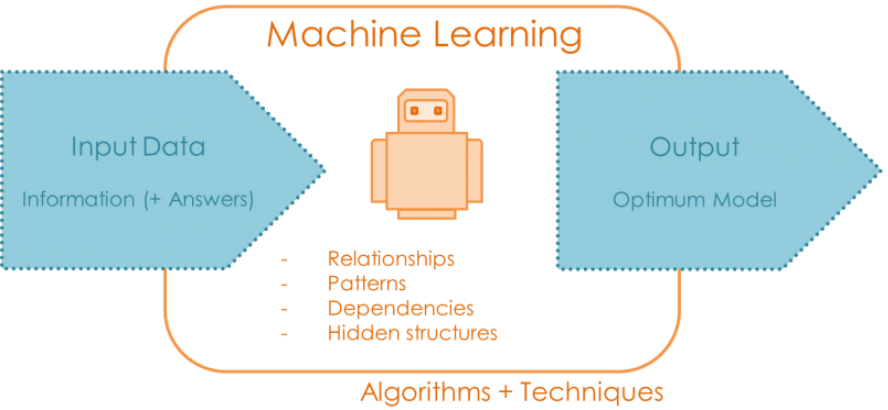
Useful links: re3data.org REGISTRY OF RESEARCH DATA REPOSITORIES

Big Data

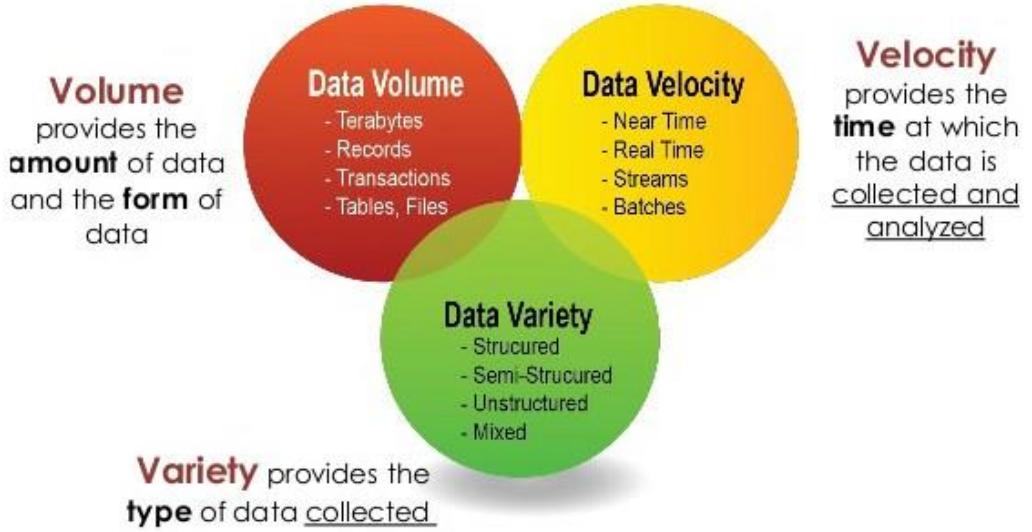
"Big data is what we got when the decision cost of deleting data became greater than the cost of storing it"
George Dyson at Strata London

Data becomes "big" through...

- ... volume and/or
- ... velocity and/or
- ... variety



Source: https://quantdare.com/machine-learning-a-brief-breakdown/



Source: https://vinodsblog.com/2017/12/28/data-an-unbelievable-hidden-treasure/

Machine learning...

- This is the detection of patterns through high computational power
- Training data is used to build models which can then be used for predictions
- Example: Detection of land uses by satellite imaging

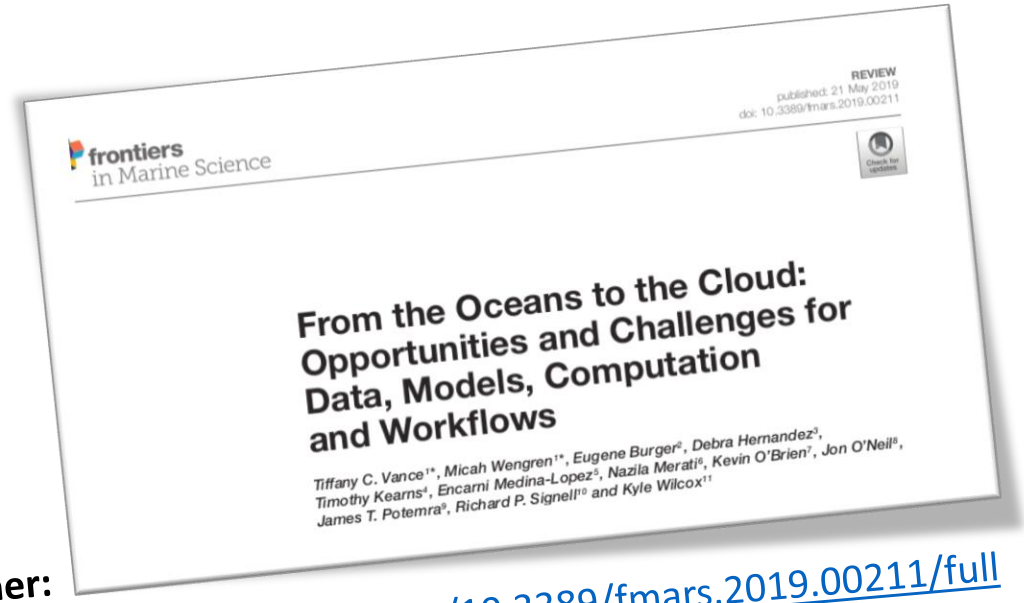
Cloud Platforms in Science

Cloud computing = computational resources hosted on the internet

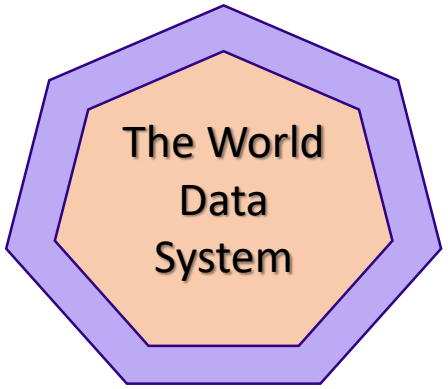
- Can be very practical, but consider the host and data security!
- Doesn't come at a one-time, but at an ongoing cost

- Used to outsource computational intensive tasks
- Any computing service done via internet
- Examples include **search engines, music streaming or web-based email or software** (in fact, one of the authors is using a web-based program to create this presentation right now :)

Source: https://www.explainthatstuff.com/cloud-computing-introduction.html



Read further: https://www.frontiersin.org/articles/10.3389/fmars.2019.00211/full



Role and responsibilities of the WDS:

Facilitates scientific research endeavours by coordinating trustworthy scientific data services for the provision, use, and preservation of relevant datasets.
https://www.icsu-wds.org/

WDS Objectives:

- Enable universal and equitable (full and open) access to quality-assured scientific data, data services, products, and information
- Ensure long-term data stewardship
- Foster compliance to agreed-upon data standards and conventions
- Provide mechanisms to facilitate and improve access to data and data products

WDS Specific targets:

1. Improve the trust in and quality of open Scientific Data Services
2. Nurture active disciplinary and multidisciplinary scientific data services communities
3. Make trusted data services an integral part of international collaborative scientific research

ipo@icsu-wds.org

WDS ECR Network

How does the WDS support ECRs?

The co-Chairs:

Alice Frémand,
UK Polar Data Centre,
British Antarctic Survey

Geophysics data manager

Sabrina Delgado Arias
Science Systems & Applications, Inc.; NASA GSFC

Current representative of the Network at the WDS Scientific Committee

What we do?

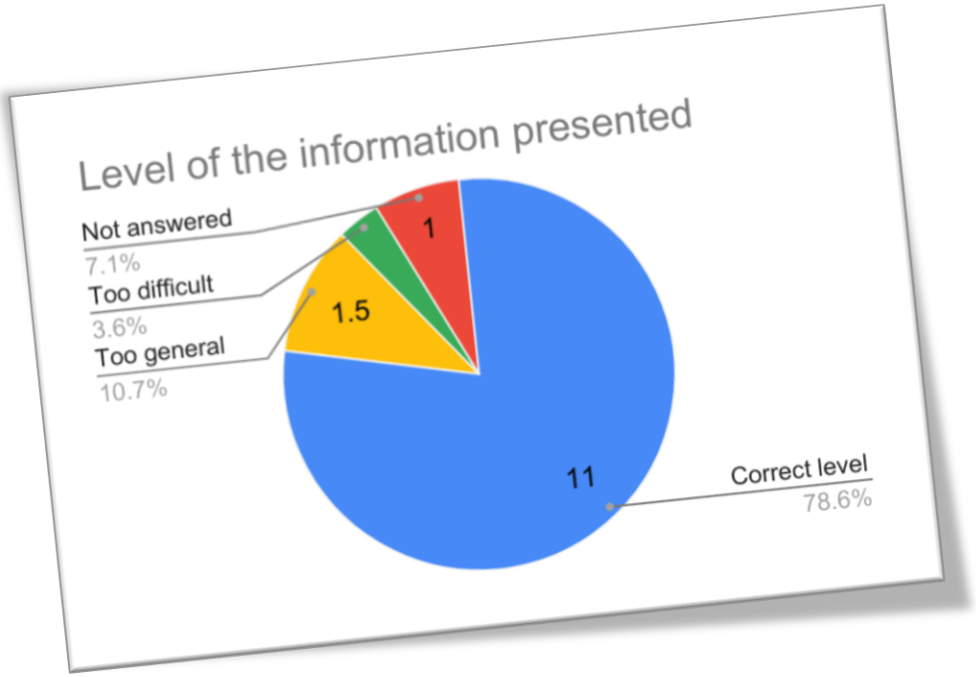
- Foster better communication among ECRs
- Design activities targeting ECR interests and concerns
- Share ideas on how we can best shape our role for future data sharing
- Connect ECRs with new training and job opportunities.
- Share our activities through our newsletter

Want to join?

@wdsECR

ECR-Chairs@icsu-wds.org

https://www.icsu-wds.org/community/ecr-network



Writing a DMP for a research project highlighted many unnoticed aspects of dealing with data...

At the moment I don't have questions. But I'm pretty sure that I will have some in the future and I know which people I can contact.

The part about the roles and responsibilities of a data repository (beyond just storing data) was very informative for me.

Maybe a hands-on session on preparing an example data set for publication and submitting it to a repository would be useful...

Conceptual aspect of copyright and plagiarism became very clear. This is something we always face in research...

Interesting yet challenging to grasp everything...

Trainers

1. **Prof Sandy Harrison** (WDS-SC Chair; Professor of Palaeoclimates and Biogeochemical Cycles, University of Reading, UK)
2. **Prof Aude Chambodut** (WDS-SC; Director of the International Service of Geomagnetic Indices, Ecole et Observatoire des Sciences de la Terre, France)
3. **Prof Elaine Faustman** (WDS-SC; Professor and Director of the Institute for Risk Analyses and Risk Communication at the University of Washington, School of Public Health, US)
4. **Dr Isabelle Gärtner-Roer** (WDS-SC; Science Officer of the World Glacier Monitoring Service, Senior Researcher in the Glaciology and Geomorphodynamics Group, and Coordinator of the Zurich Graduate School in Geography at the Department of Geography at the University of Zurich, Switzerland)
5. **Dr Ioana Popescu** (WDS-SC; Associate Professor of Hydroinformatics at IHE-Delft Institute for Water Education, The Netherlands)
6. **Ms Alice Frémand** (WDS Early Career Researcher Network; Scientific Data Manager, UK Polar Data Centre, UK)
7. **Dr Rorie Edmunds** (Acting Executive Director, WDS International Programme Office)
8. **Dr Karen Payne** (Associate Director, WDS International Technology Office)

All participants were presented with a certificate of attendance at the end of the training.

WDS: *"We strongly believe that it was highly successful. The ECRs in attendance found it not only was a good experience, but more importantly was relevant and meaning to their careers moving forward."*