The What, Why and How of Research Data Management

Iryna Kuchma, EIFL

EaPConnect's OpenAIRE Training| Kyiv | June 5th, 2019

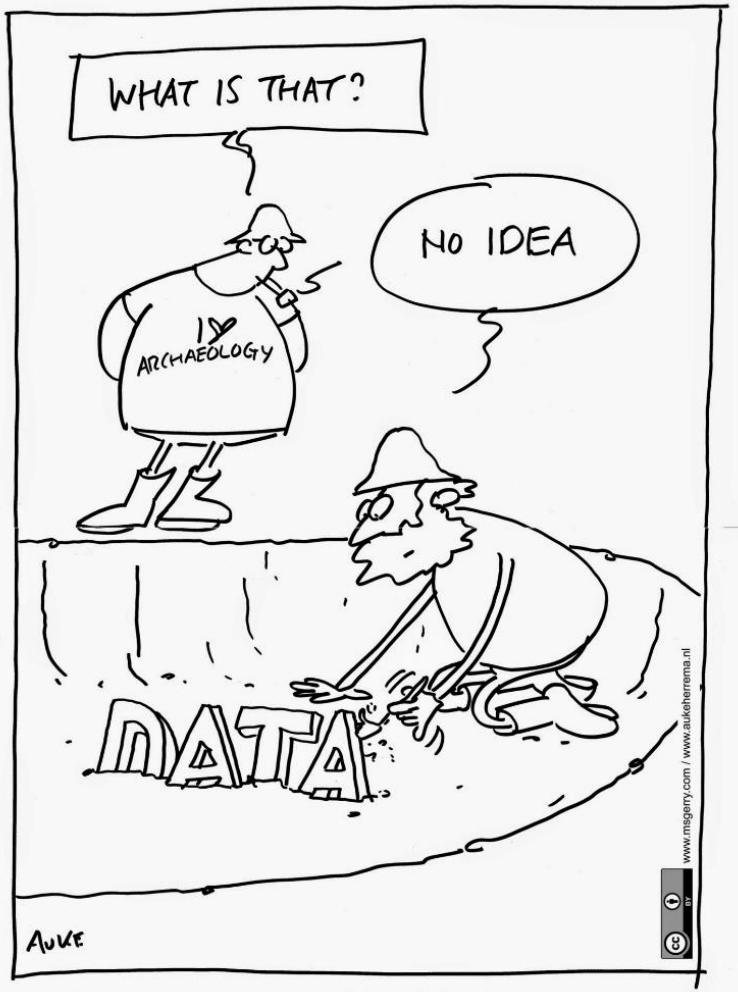
”If we wait 5 years for (Arctic) data to be released, the Arctic is going to be a very different place”

Parsons, Arctic Research Scientist

Source: http://www.nature.com/nature/journal/v461/n7261/index.html

FOSTER

PUBLICATIONS AND DATA



DATA FOR FUTURE GENERATIONS

Levels of openness

Open data - the Open Data Institute (ODI) defines Open Data as those that anyone can access, use and share. According to the ODI, open data must be licensed to make clear that anyone can use the data in any way they want, including transforming, combining, and sharing it with others, even for commercial purposes. The ODI provides a great introduction to all aspects of Open Data in their Open Data Essentials course. We highly recommend reviewing these modules.

Shared data - similar to Open data, shared data may be made widely accessible but could have some conditions such as non-commercial reuse or reuse with attribution. It is important to note that not all shared data has to be available to anyone. Sometimes shared data is only made available to specific groups such as peers from another university.

Closed data - if researchers are dealing with highly sensitive data - such as sensitive personal data or commercially sensitive data - it may not be possible to share the data at all. However, even in such cases a metadata description of the research data should be shared. Sharing of sensitive data can also be supported by making use of safe havens where only authorised users are given controlled access.

Tip – use 5 Star Open Data Model to explain FAIR

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

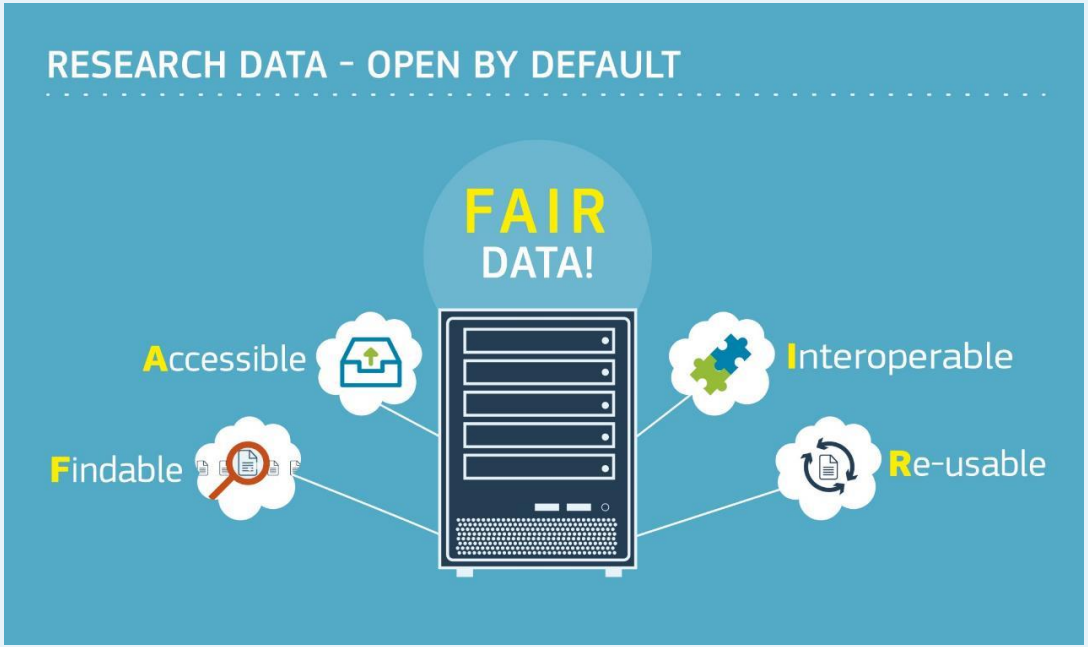
★★ make it available as structured data (e.g., Excel instead of image scan of a table)

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

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

★★★★★ link your data to other data to provide context

Tim Berners-Lee's proposal for five star open data - http://5stardata.info


RESEARCH DATA - OPEN BY DEFAULT
FAIR
DATA!
Accessible
Interoperable
Findable
Re-usable

Making data FAIR

Findable – Assign persistent IDs, provide rich metadata, register in a searchable resource,...

Accessible - Retrievable by their ID using a standard protocol, metadata remain accessible even if data aren't...

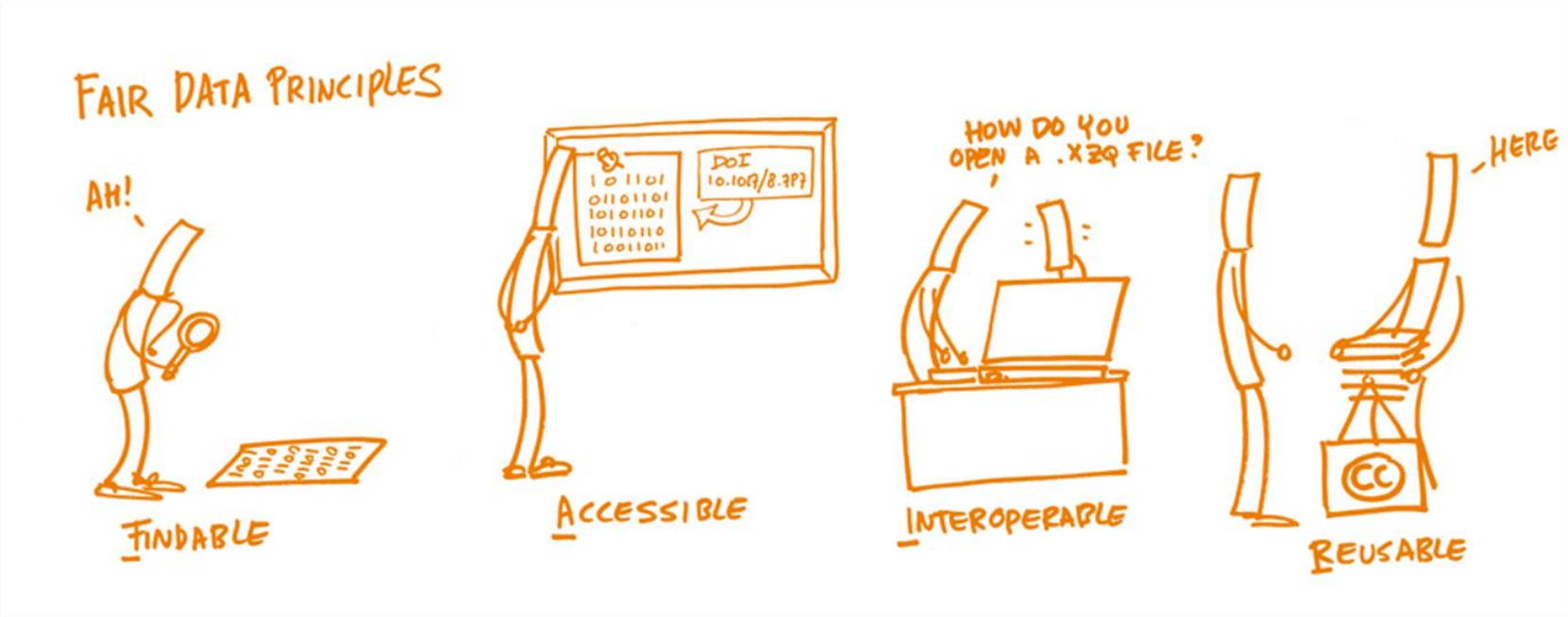
Interoperable - Use formal, broadly applicable languages, use standard vocabularies, qualified references...

Reusable - Rich, accurate metadata, clear licences, provenance, use of community standards

www.force11.org/group/fairgroup/fairprinciples

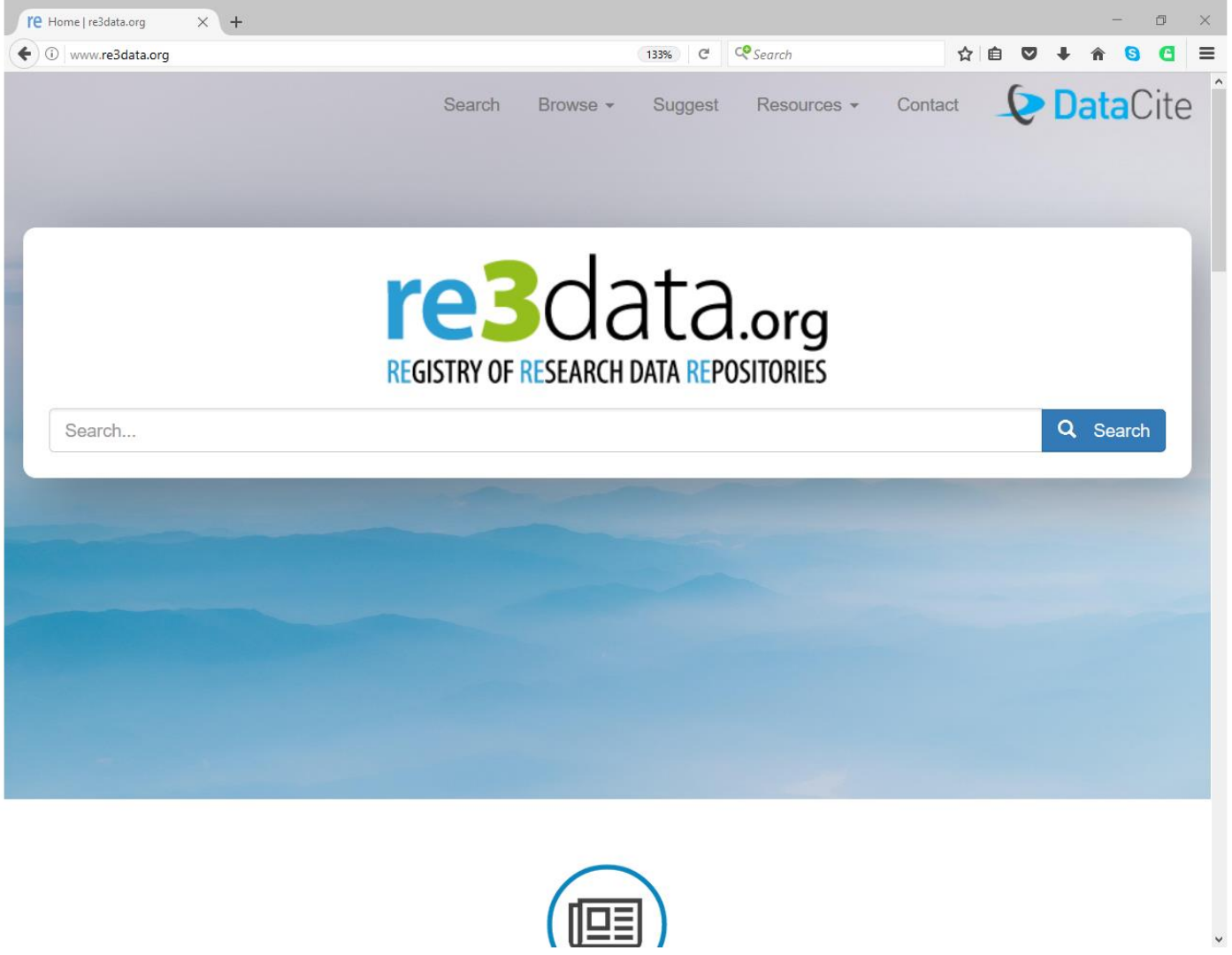
Misconception #1:

My web page is a FAIR way to share my data.



Better options for open data

- Domain repository (first choice)
- General repository (Figshare, Zenodo)
- Institutional repository
- Data journal
- Journal supplementary material


Home | re3data.org
www.re3data.org
133%
Search
Search
Browse
Suggest
Resources
Contact
DataCite
re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES
Search...
Search

Browse by subject | re3data.

www.re3data.org/browse/by-subject/

133%

Search

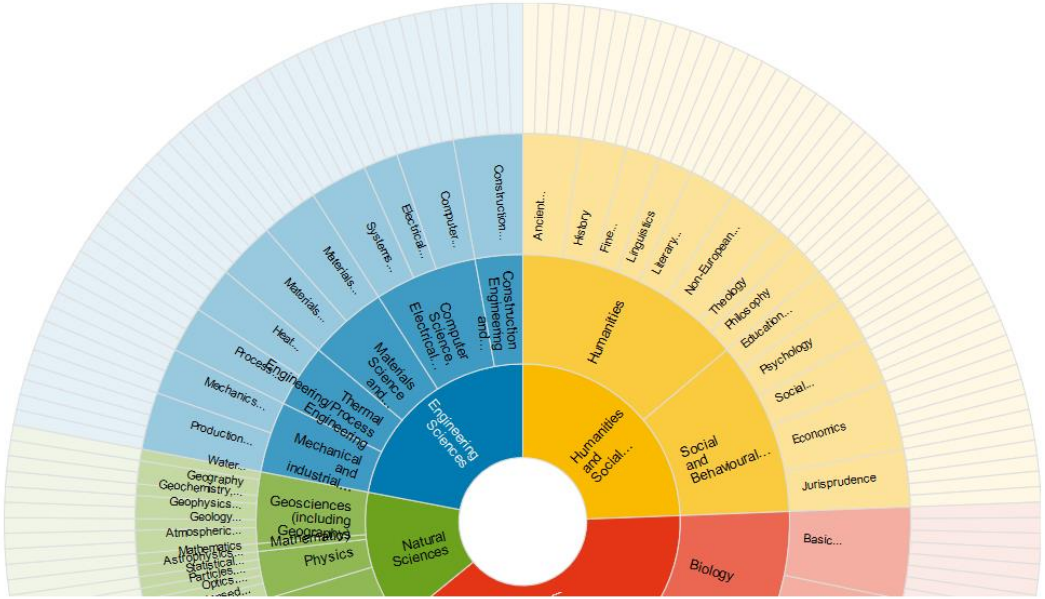
re3data.org

Search Browse Suggest Resources Contact DataCite

Browse by subject

Graphical Text

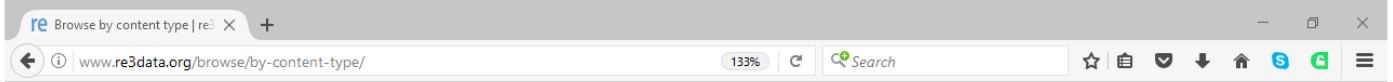
click to zoom into subjects or to select a bottommost subject in the hierachy as filter for the re3data search page
ctrl + click on a top subject to select it as filter



Browse by subject

Graphical | Text

A. Humanities and Social Sciences
- a. Humanities
  - I. Ancient Cultures
    1. Prehistory
    2. Classical Philology
    3. Ancient History
    4. Classical Archaeology
    5. Egyptology and Ancient Near Eastern Studies
  - II. History
    1. Medieval History
    2. Early Modern History
    3. Modern and Current History
    4. History of Science
  - III. Fine Arts, Music, Theatre and Media Studies
    1. Art History
    2. Musicology
    3. Theatre and Media Studies
  - IV. Linguistics
    1. General and Applied Linguistics
    2. Individual Linguistics
    3. Typology, Non-European Languages, Historical Linguistics
  - V. Literary Studies
    1. Medieval German Literature
    2. Modern German Literature
    3. European and American Literature
    4. General and Comparative Literature and Cultural Studies



Browse by content type

- Archived data
- Audiovisual data
- Configuration data
- Databases
- Images
- Networkbased data
- Plain text
- Raw data
- Scientific and statistical data formats
- Software applications
- Source code
- Standard office documents
- Structured graphics
- Structured text
- other

Browse by country | re3data

www.re3data.org/browse/by-country/

133%

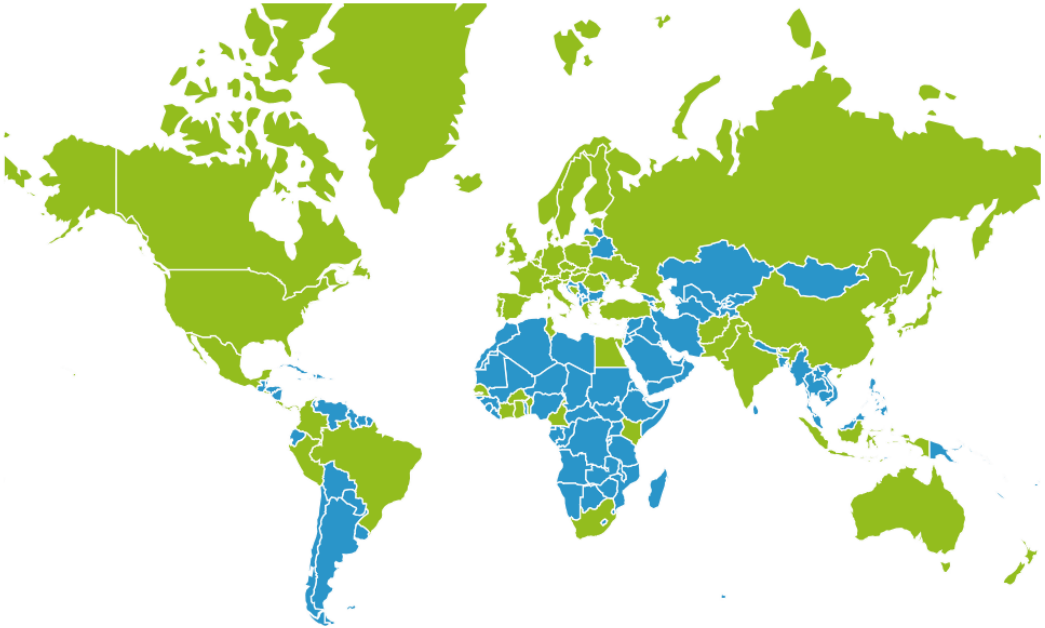
Search

re3data.org

Search Browse Suggest Resources Contact DataCite

Browse by country

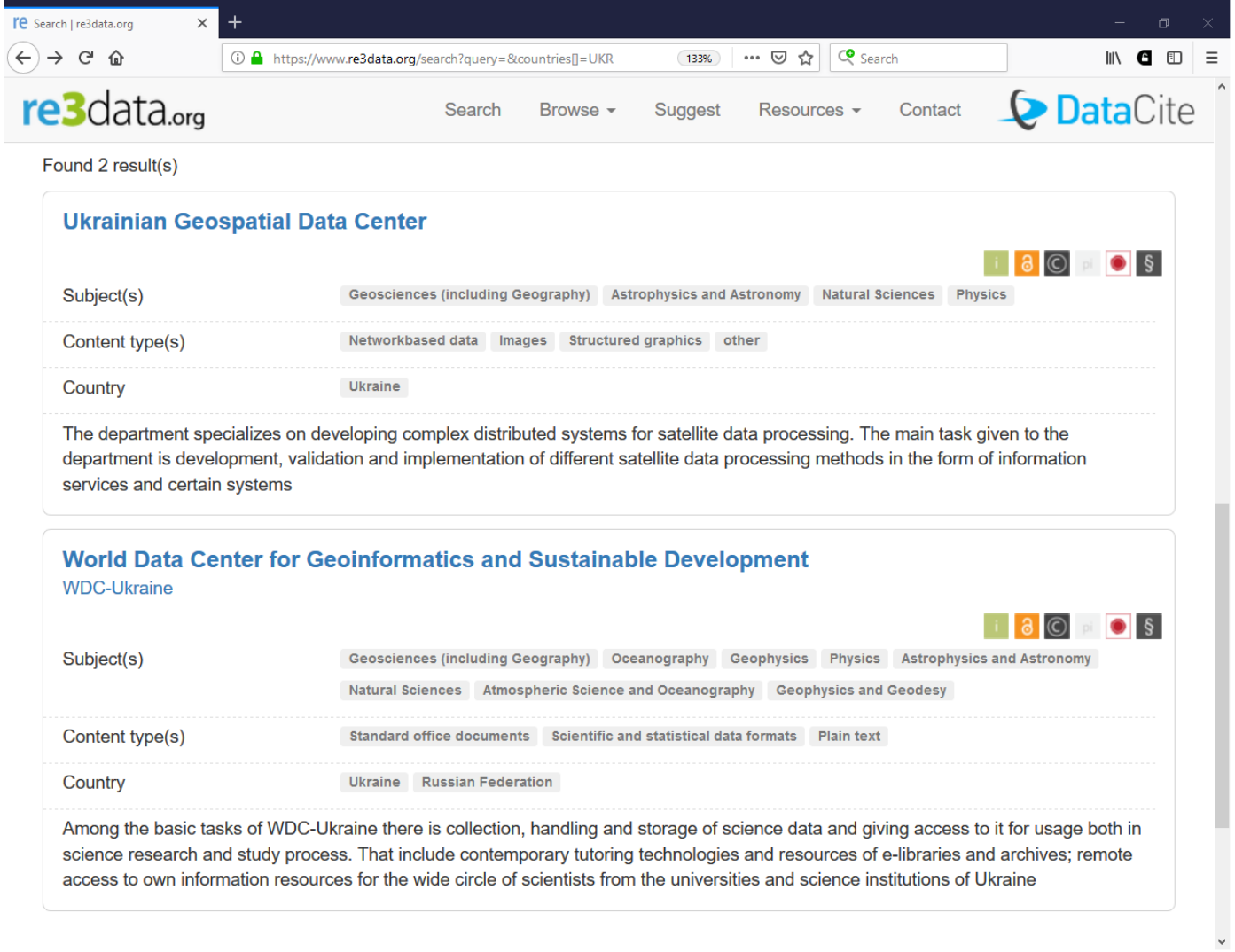
Graphical Text



Browse by country

Graphical | Text

- International
- Afghanistan
- Australia
- Austria
- Azerbaijan
- Belgium
- Benin
- Burkina Faso
- Bosnia and Herzegovina
- Brazil
- Canada
- Switzerland
- China
- Cote d'Ivoire
- Cameroon
- Colombia
- Costa Rica
- Cyprus
- Czech Republic
- Germany
- Denmark
- European Union
- Egypt
- Spain
- Estonia

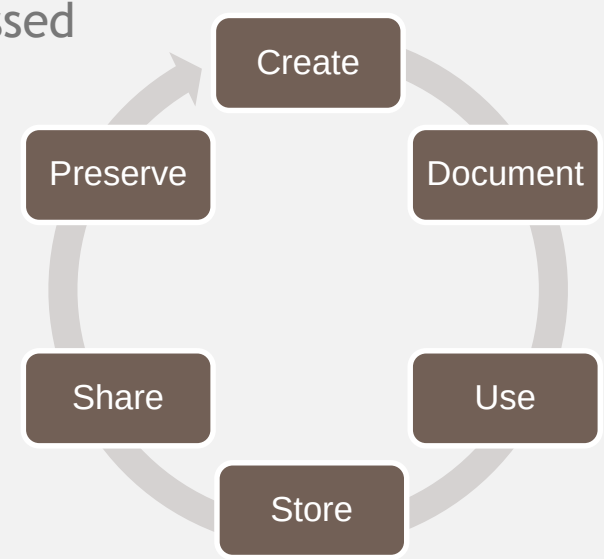

Search | re3data.org
https://www.re3data.org/search?query=&countries[]=UKR
133%
Search
re3data.org
Search
Browse
Suggest
Resources
Contact
DataCite
Found 2 result(s)
Ukrainian Geospatial Data Center
Subject(s)
Geosciences (including Geography)
Astrophysics and Astronomy
Natural Sciences
Physics
Content type(s)
Networkbased data
Images
Structured graphics
other
Country
Ukraine
The department specializes on developing complex distributed systems for satellite data processing. The main task given to the department is development, validation and implementation of different satellite data processing methods in the form of information services and certain systems
World Data Center for Geoinformatics and Sustainable Development
WDC-Ukraine
Subject(s)
Geosciences (including Geography)
Oceanography
Geophysics
Physics
Astrophysics and Astronomy
Natural Sciences
Atmospheric Science and Oceanography
Geophysics and Geodesy
Content type(s)
Standard office documents
Scientific and statistical data formats
Plain text
Country
Ukraine
Russian Federation
Among the basic tasks of WDC-Ukraine there is collection, handling and storage of science data and giving access to it for usage both in science research and study process. That include contemporary tutoring technologies and resources of e-libraries and archives; remote access to own information resources for the wide circle of scientists from the universities and science institutions of Ukraine

Misconception #2:

I don’t need to decide now if I want to share. I can wait and see what I want to do at the end of my project.

Open Data doesn’t just happen - data management planning helps!

- What data will be created (format, types, volume...)
- Standards and methodologies to be used, documentation
- How ethics and Intellectual Property will be addressed
- Plans for storage and back-up
- Plans for data sharing and access
- Strategy for long-term preservation



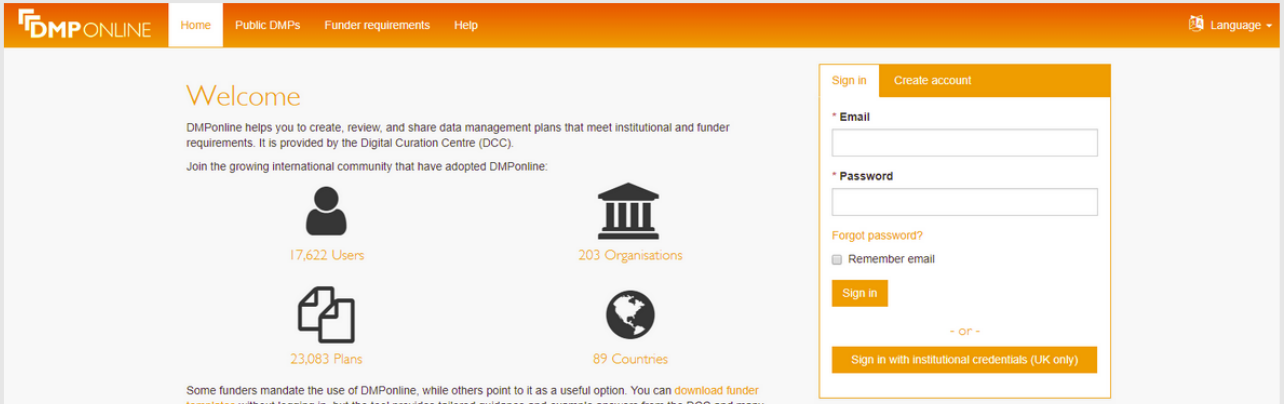
Tip – use existing tools and guidance to help write their plans

https://dmponline.dcc.ac.uk

Data management planning tools - DMPonline

DMPonline is a freely available tool that helps research teams to write data management plans that meet funding body requirements. DMPonline was jointly developed by the Digital Curation Centre (DCC) and the University of California Curation Center (UC3). The tool contains a number of templates that represent the requirements of different funding bodies across Europe. Users are asked three questions at the outset to determine the appropriate template to display (e.g. the Economic and Social Research Council (ESRC) template when applying for an ESRC grant). Using tools like DMPonline takes the guesswork out of writing your data management plan by providing you with the specific set of questions that individual funding bodies want you to answer. The tool also provides users with general guidance - and where provided, institutional guidance - to make sure that your answers are realistic and implementable.

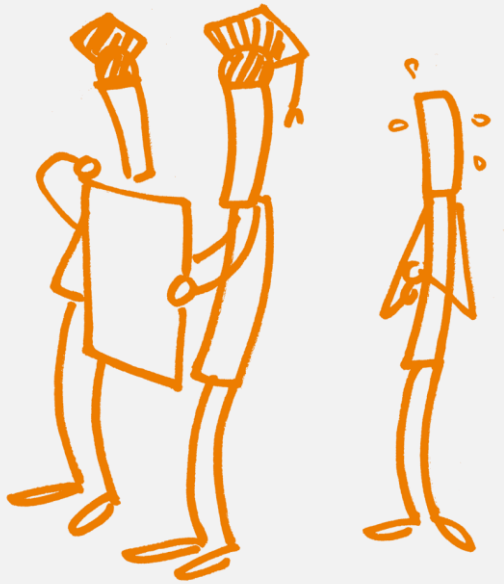
For more information on data management plans and tips on writing them, check out the DCC website.



https://www.fosteropenscience.eu/learning/managing-and-sharing-research-data

Misconception #3:

If I share my data early, I'll be scooped!



Pre-registration timestamps your work

Register Your Project

A registration on OSF creates a frozen, time-stamped version of a project that cannot be edited or deleted. The *original project* can still be edited, while the registered version cannot. You might create a registration to capture a snapshot of your project at certain points in time - such as right before data collection begins, when you submit a manuscript for peer review, or upon completion of a project.

Registrations can be made public immediately or embargoed for up to 4 years. Registrations cannot be deleted, but they can be withdrawn. Withdrawing a registration removes the content of the registration but leaves behind basic metadata, like registration title, contributors, and a reason for the withdrawal (not required).

http://help.osf.io/m/registrations/l/524205-register-your-project

Tips – share preprints too

- Early feedback on methods and initial findings
- Time to correct and mistakes before publishing
- Recognition for your ideas by peers

Misconception #4:

I have to keep and share everything.

Image: ‘Balancing rocks’ by Viewminder CC-BY-SA-ND www.flickr.com/photos/light_seeker/7780857224

Deciding which data need to be kept after the project ends

Five steps to follow

1. **Could** this data be re-used
2. **Must** it be kept as evidence or for legal reasons
3. **Should** it be kept for its potential value
4. **Consider costs** – do benefits outweigh cost?
5. **Evaluate criteria** to decide what to keep

5 steps to decide what data to keep

www.dcc.ac.uk/resources/how-guides/five-steps-decide-what-data-keep

What should be preserved and shared?

- The data needed to validate results in scientific publications (minimally!).
- The associated metadata: the dataset's creator, title, year of publication, repository, identifier etc.
  - Follow a metadata standard in your line of work, or a generic standard, e.g. Dublin Core or DataCite, and be FAIR.
  - The repository will assign a persistent ID to the dataset: important for discovering and citing the data.

What should be preserved and shared? (2)

- Documentation: code books, lab journals, informed consent forms – domain-dependent, and important for understanding the data and combining them with other data sources.
- Software, hardware, tools, syntax queries, machine configurations – domain-dependent, and important for using the data. (Alternative: information about the software etc.)

Basically, everything that is needed to replicate a study should be available. Plus everything that is potentially useful for others.

Tip – link data to other outputs for context (reuse)

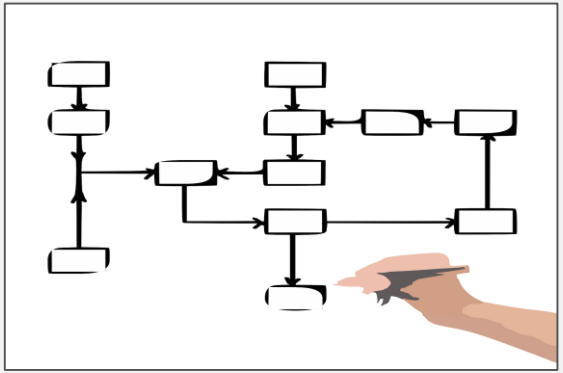
Open Data

To support validation and facilitate reuse

Open Code

Software created to analyse and/or visualise the data

Open Workflows



What steps were taken and in what order?

Consider who else has a say about sharing data

- Collaborators
- Research participants
- Commercial partners
- Data repository
- Publishers
- Institutions, funders

How to make data open?

https://okfn.org

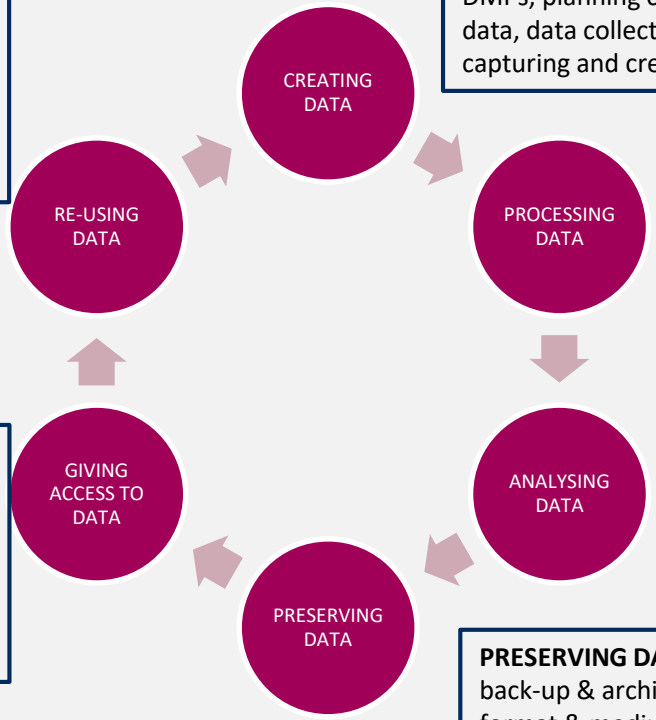
1. Choose your dataset(s)
   - What can you open? You may need to revisit this step if you encounter problems later.
2. Apply an open license
   - Determine what IP exists. Apply a suitable licence e.g. CC-BY
3. Make the data available
   - Provide the data in a suitable format. Use repositories.
4. Make it discoverable
   - Post on the web, register in catalogues…

FOSTER

Research data lifecycle

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

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



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

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

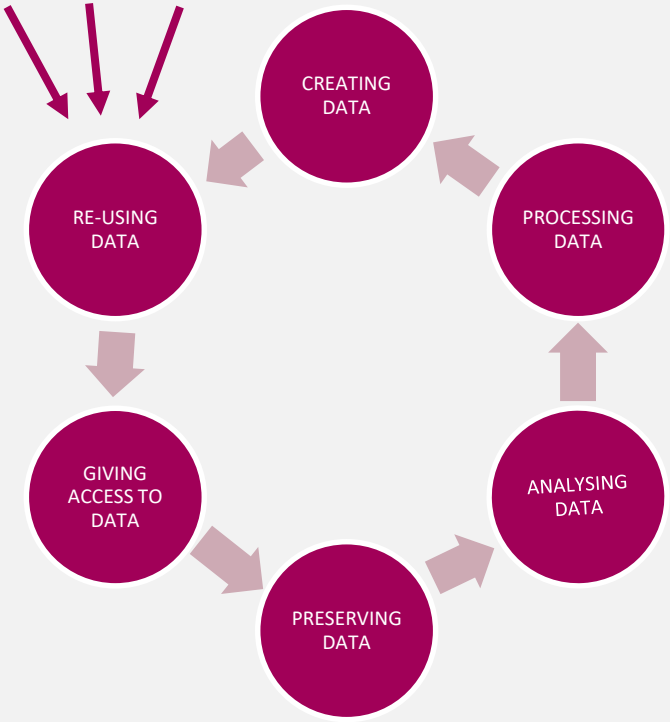
PRESERVING DATA: data storage, back-up & archiving, migrating to best format & medium, creating metadata and documentation

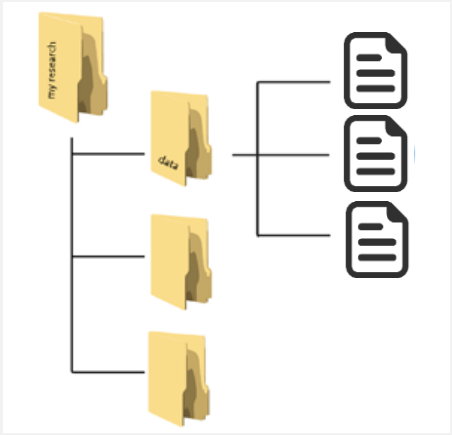
ACCESS TO DATA: distributing data, sharing data, controlling access, establishing copyright, promoting data

Ref: UK Data Archive: http://www.data-archive.ac.uk/create-manage/life-cycle

Planning trick 1: think backwards

What data organisation would a re-user like?





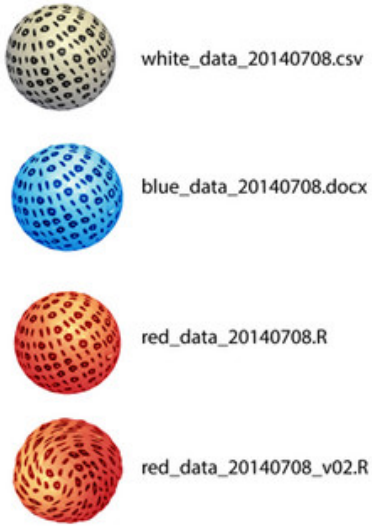
Data organisation

Meaningful file names

Below are tips on meaningful and consistent file names. Read more in '*Choosing a file name*'. (2)

- Make sure to use consistent file names. When you use a date in the file name, choose a notation (for instance, YYYYMMDD of yymmdd).
- Do not use strange characters like ?\!@*%{[<> in the file name.
- Use traceable file names, such as Project_Instrument_locatie_YYYYMMDD.ext.
- Make sure to only use each file once in the folder structure. If you store a file in more than one place, several versions of the same file can unwillingly be created.
- See also version management.

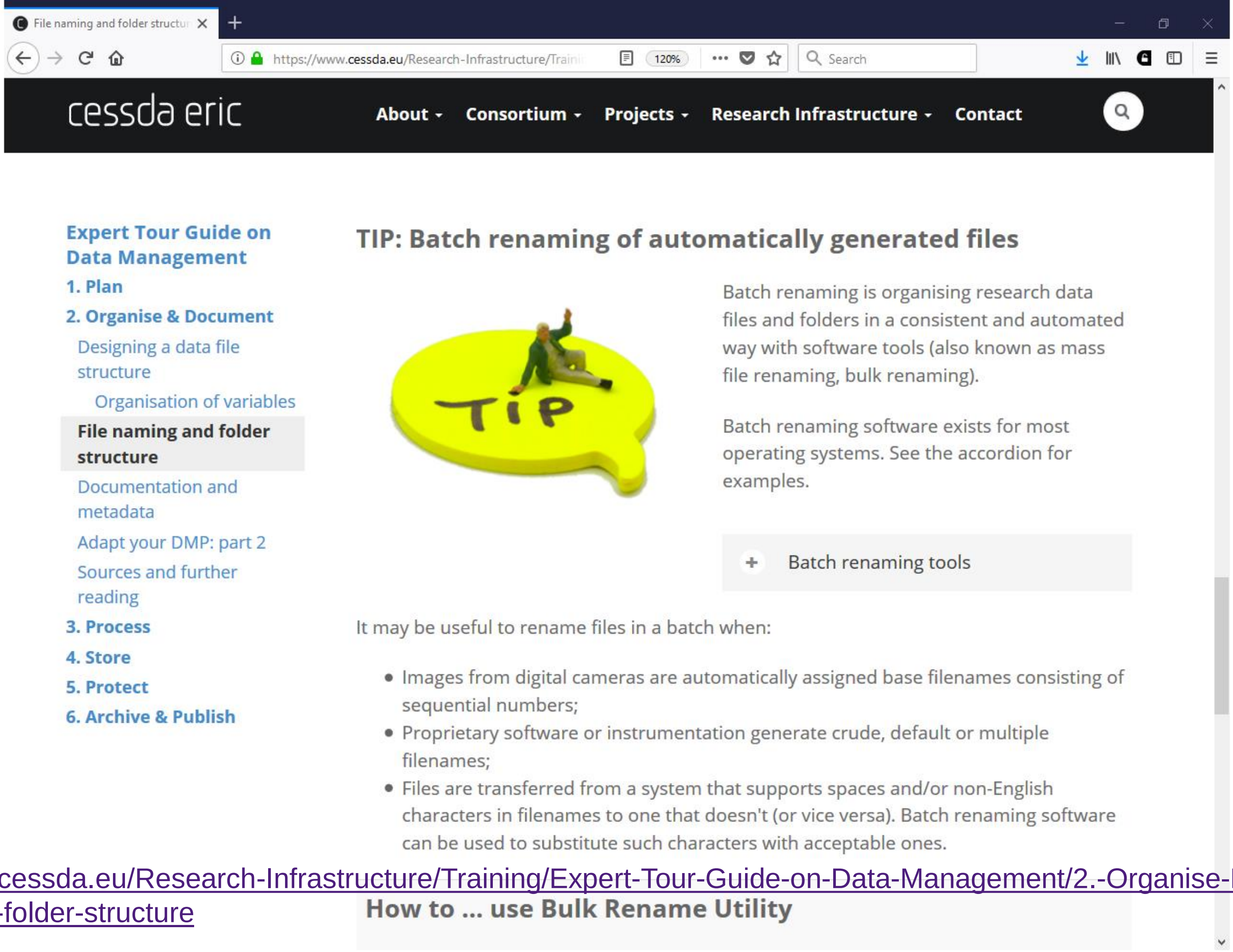
It is good practice to note the file naming and its meaning in a readme.txt.



File naming and version management

Even if a researcher is well underway with his project consistent file naming is still an option by using a *bulk file rename utility*. (3) It is important, however, to check if this bulk renamer delivers on its promises.

http://datasupport.researchdata.nl/en/start-de-cursus/iii-onderzoeksfase/organising-data



File naming and folder structur

https://www.cessda.eu/Research-Infrastructure/Traini

cessda eric

About · Consortium · Projects · Research Infrastructure · Contact

Expert Tour Guide on Data Management

- **1. Plan**
- **2. Organise & Document**
  - Designing a data file structure
    - Organisation of variables
  - **File naming and folder structure**
  - Documentation and metadata
  - Adapt your DMP: part 2
  - Sources and further reading
- **3. Process**
- **4. Store**
- **5. Protect**
- **6. Archive & Publish**

TIP: Batch renaming of automatically generated files

TIP

Batch renaming is organising research data files and folders in a consistent and automated way with software tools (also known as mass file renaming, bulk renaming).

Batch renaming software exists for most operating systems. See the accordion for examples.

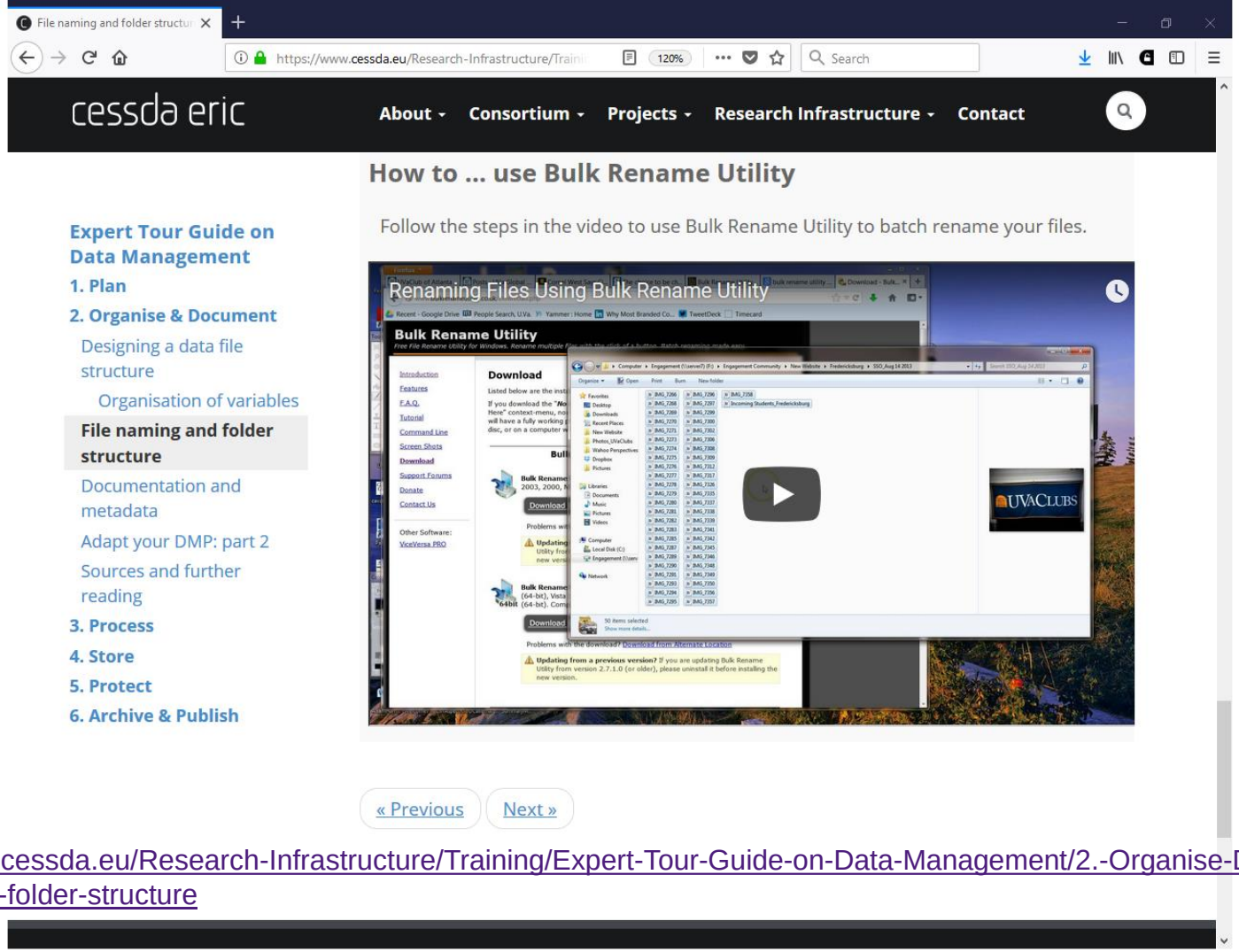
+ Batch renaming tools

It may be useful to rename files in a batch when:

- Images from digital cameras are automatically assigned base filenames consisting of sequential numbers;
- Proprietary software or instrumentation generate crude, default or multiple filenames;
- Files are transferred from a system that supports spaces and/or non-English characters in filenames to one that doesn't (or vice versa). Batch renaming software can be used to substitute such characters with acceptable ones.

How to ... use Bulk Rename Utility

https://www.cessda.eu/Research-Infrastructure/Training/Expert-Tour-Guide-on-Data-Management/2.-Organise-Document/File-naming-and-folder-structure



https://www.cessda.eu/Research-Infrastructure/Training/Expert-Tour-Guide-on-Data-Management/2.-Organise-Document/File-naming-and-folder-structure

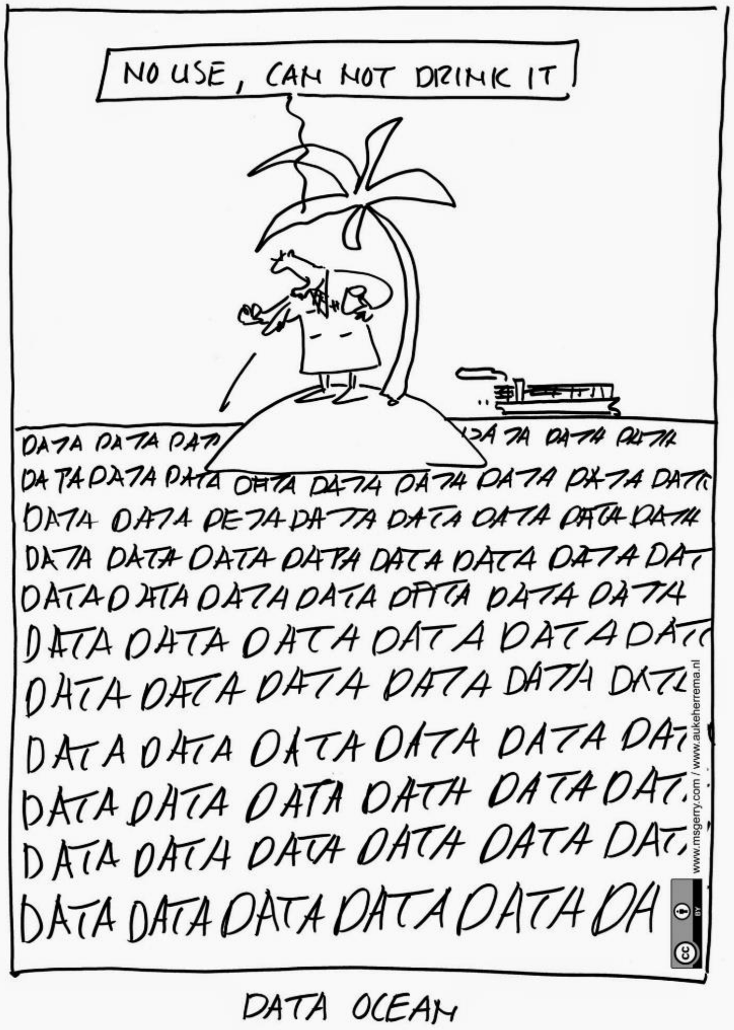
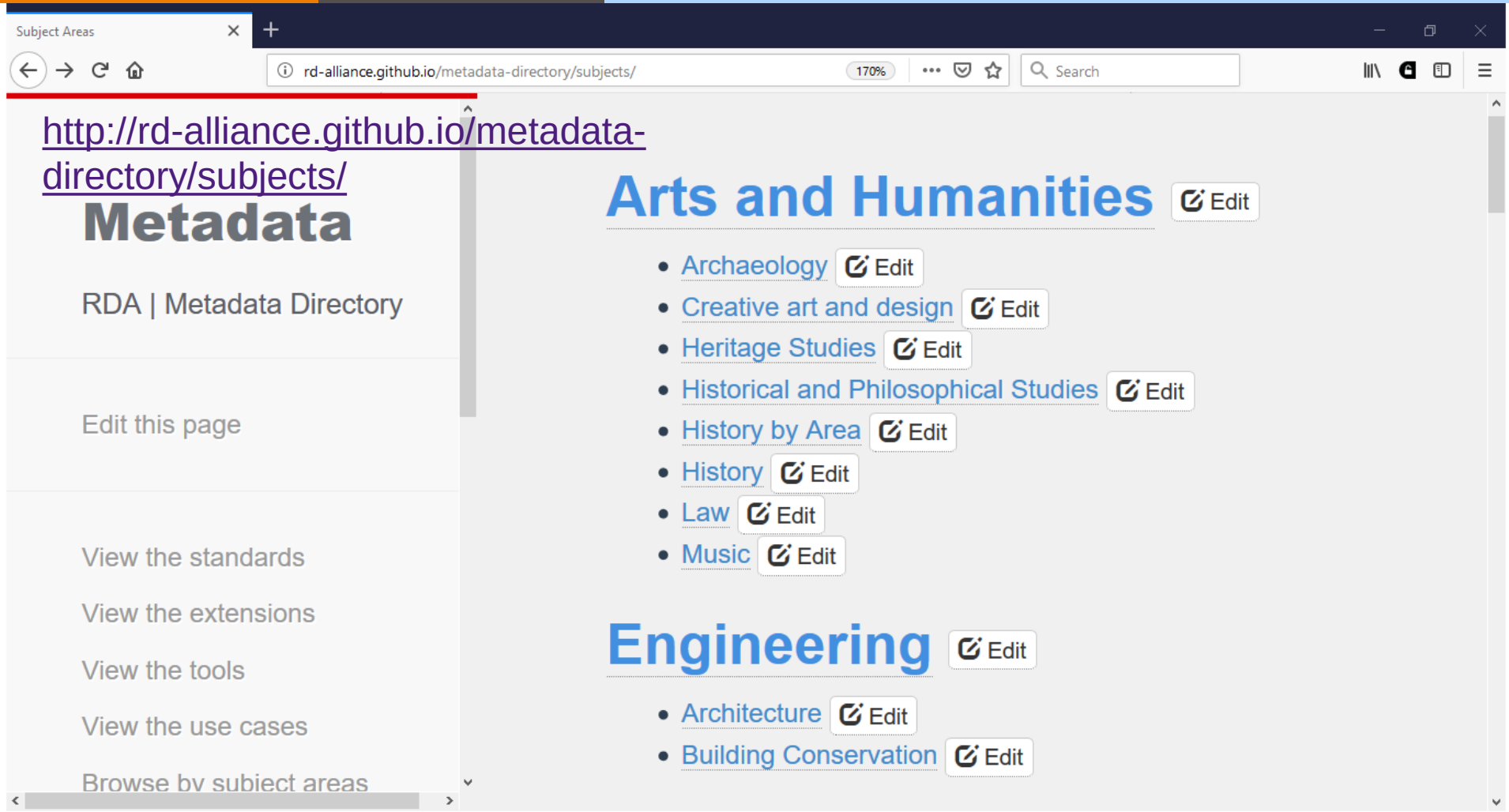


Image courtesy of http://aukeherrema.nl CC-BY

REUSABLE DATA


Subject Areas
rd-alliance.github.io/metadata-directory/subjects/
170%
Search
http://rd-alliance.github.io/metadata-directory/subjects/
Metadata
RDA | Metadata Directory
Edit this page
View the standards
View the extensions
View the tools
View the use cases
Browse by subject areas
Arts and Humanities
Edit
Archaeology
Creative art and design
Heritage Studies
Historical and Philosophical Studies
History by Area
History
Law
Music
Engineering
Architecture
Building Conservation

Archiving, repositories, ehm?

Select a data repository that will preserve your data, metadata and possibly tools in the long term.

It is advisable to contact the repository of your choice when writing the first version of your DMP.

Repositories may offer guidelines for sustainable data formats and metadata standards, as well as support for dealing with sensitive data and licensing.

Where to find a repository?

More information: https://www.openaire.eu/opendatapilot-repository

Zenodo: http://www.zenodo.org

Re3data.org: http://www.re3data.org

How to select a repository?

Main criteria for choosing a data repository:

Certification as a ‘Trustworthy Digital Repository’, with an explicit ambition to keep the data available in the long term.

- Three common certification standards for TDRs:

Data Seal of Approval: http://datasealofapproval.org/en
nestor seal: http://www.langzeitarchivierung.de/Subsites/nestor/EN/nestor-Siegel/siegel_node.html
ISO 16363: http://www.iso16363.org

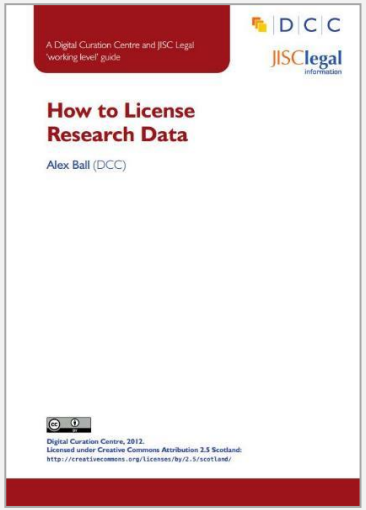
How to select a repository? (2)

- Matches your particular data needs: e.g. formats accepted; mixture of Open and Restricted Access.

- Provides guidance on how to cite the data that has been deposited.

- Gives your submitted dataset a persistent and globally unique identifier: for sustainable citations – both for data and publications – and to link back to particular researchers and grants.

Licensing research data



Horizon 2020 Open Access guidelines point to:

CC ZERO or CC BY

This DCC guide outlines the pros and cons of each approach and gives practical advice on how to implement your licence

CREATIVE COMMONS LIMITATIONS

NC Non-Commercial
What counts as commercial?

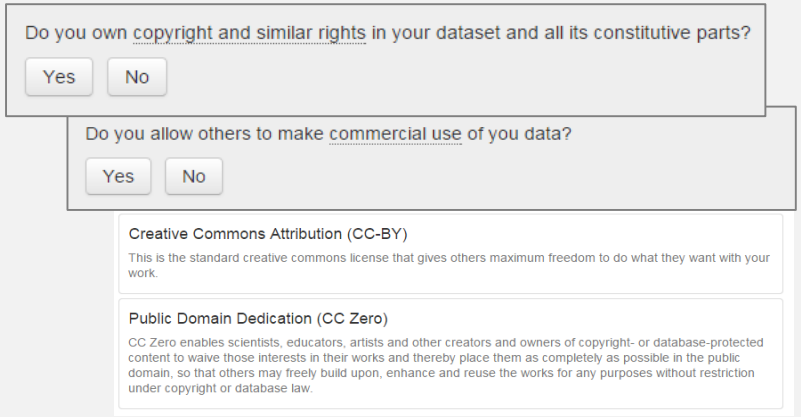
ND No Derivatives
Severely restricts use

These clauses are not open licenses

www.dcc.ac.uk/resources/how-guides/license-research-data

EUDAT licensing tool

Answer questions to determine which licence(s) are appropriate to use



http://ufal.github.io/public-license-selector

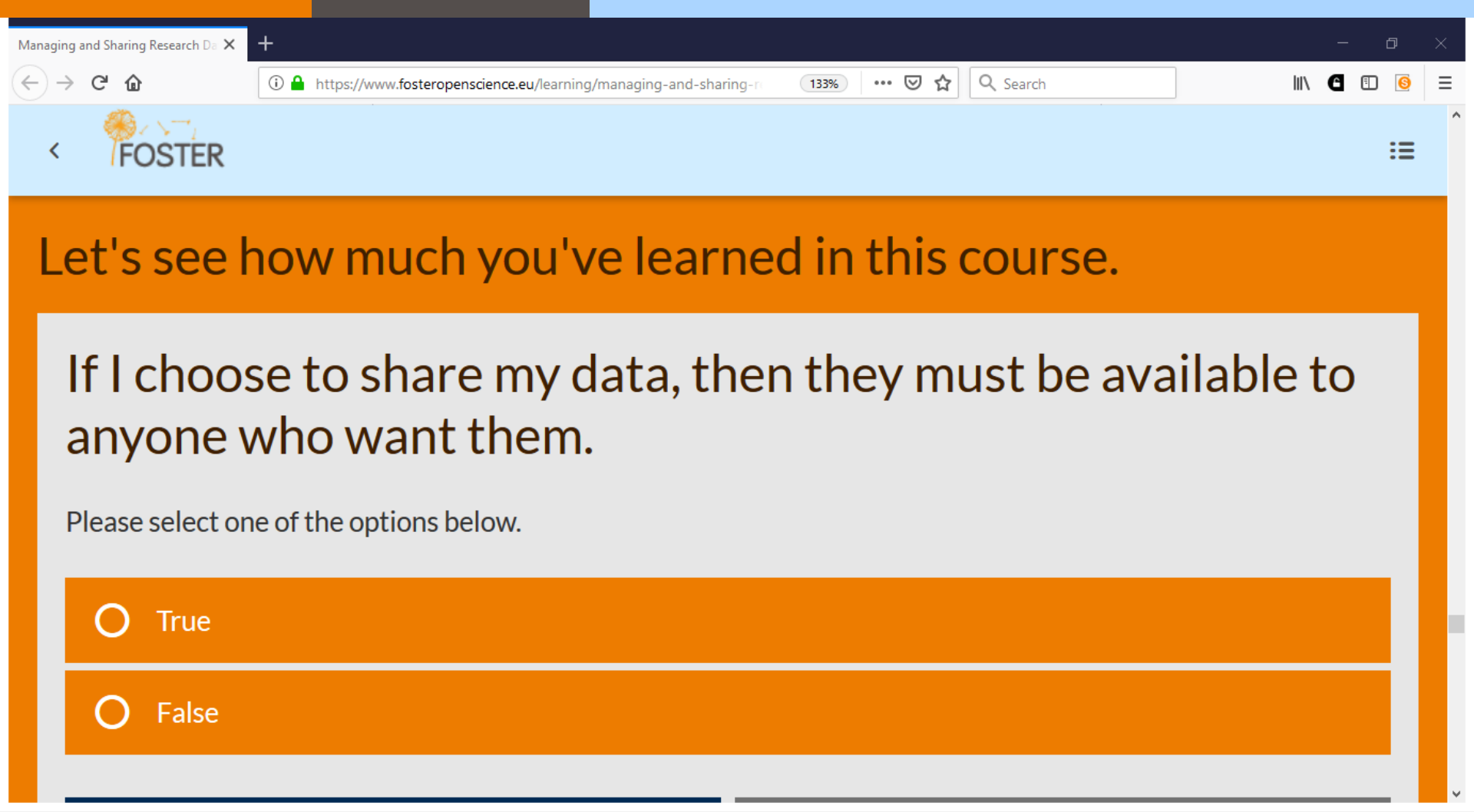
Why manage data?

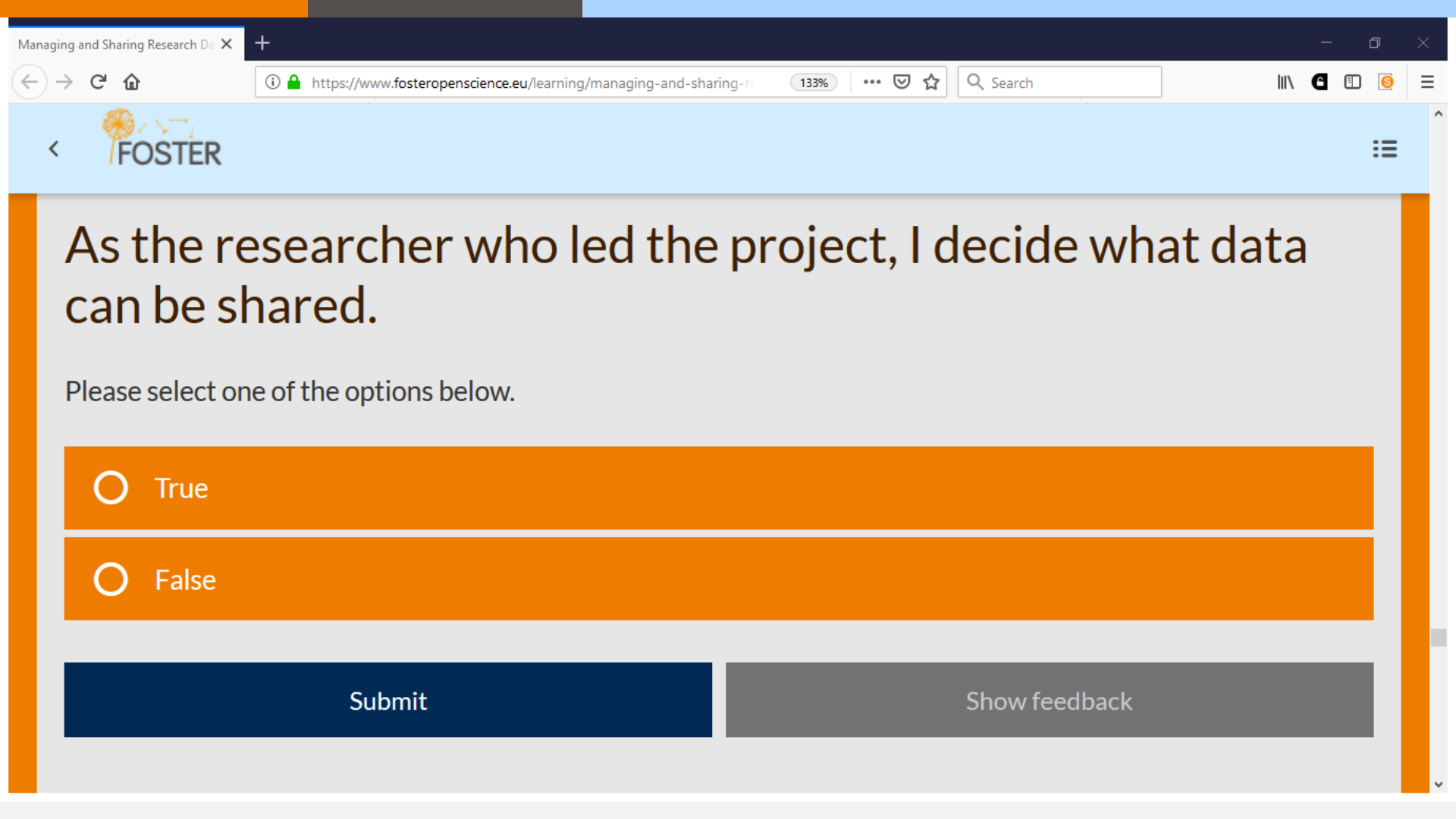
NON PECUNIAE INVESTIGATIONIS CURATORE

SED VITAE FACIMUS PROGRAMMAS DATORUM PROCURATIONIS

(Not for the research funder, but for life we make data management plans)

- Make your research easier
- Stop yourself drowning in irrelevant stuff
- Save data for later
- Avoid accusations of fraud or bad science
- Write a data paper
- Share your data for re-use
- Get credit for it


Managing and Sharing Research Da
https://www.fosteropenscience.eu/learning/managing-and-sharing-r
133%
Search
FOSTER
Let's see how much you've learned in this course.
If I choose to share my data, then they must be available to anyone who want them.
Please select one of the options below.
True
False

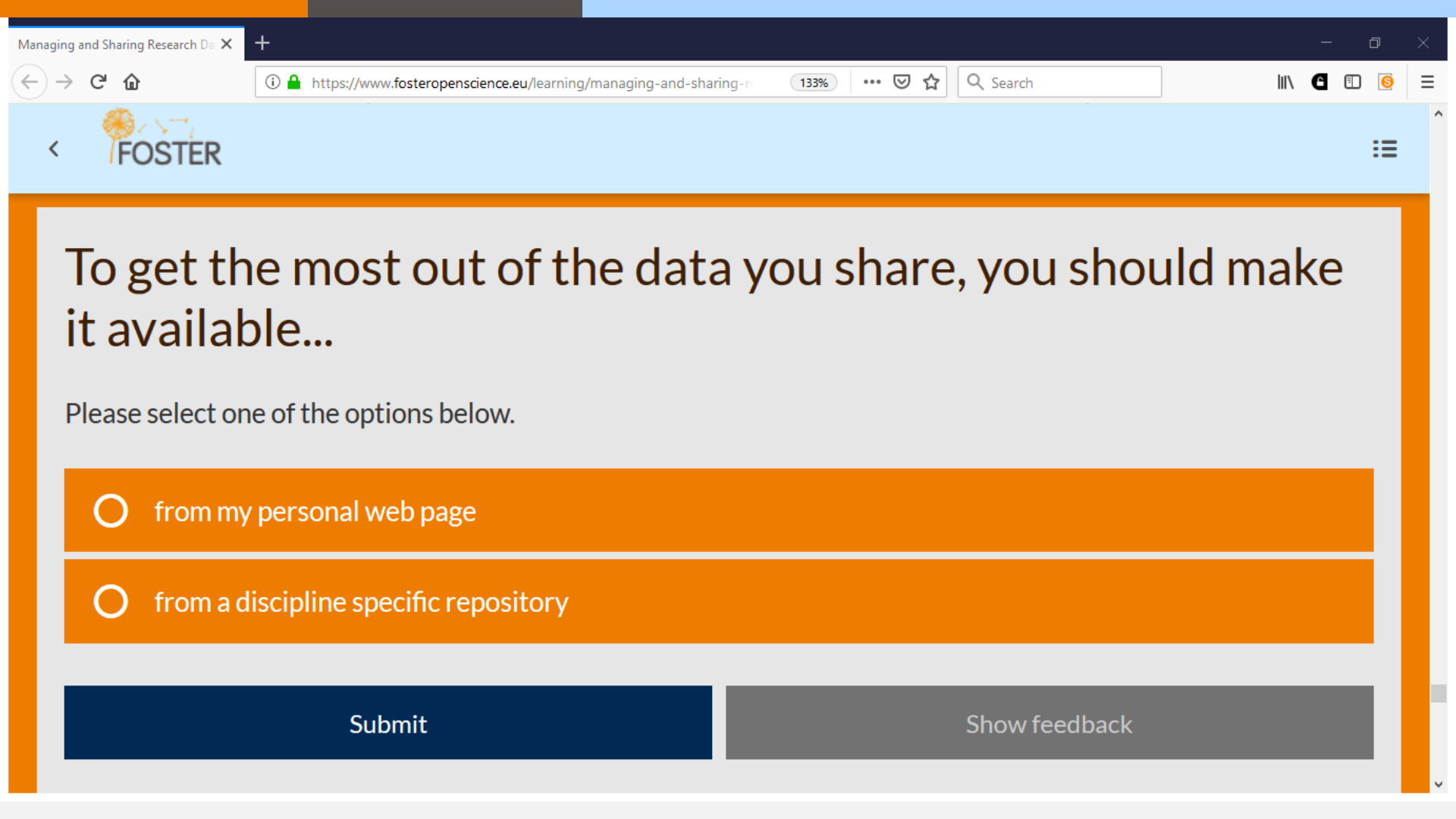

Managing and Sharing Research Da
https://www.fosteropenscience.eu/learning/managing-and-sharing-r
133%
Search
FOSTER
As the researcher who led the project, I decide what data can be shared.
Please select one of the options below.
True
False
Submit
Show feedback

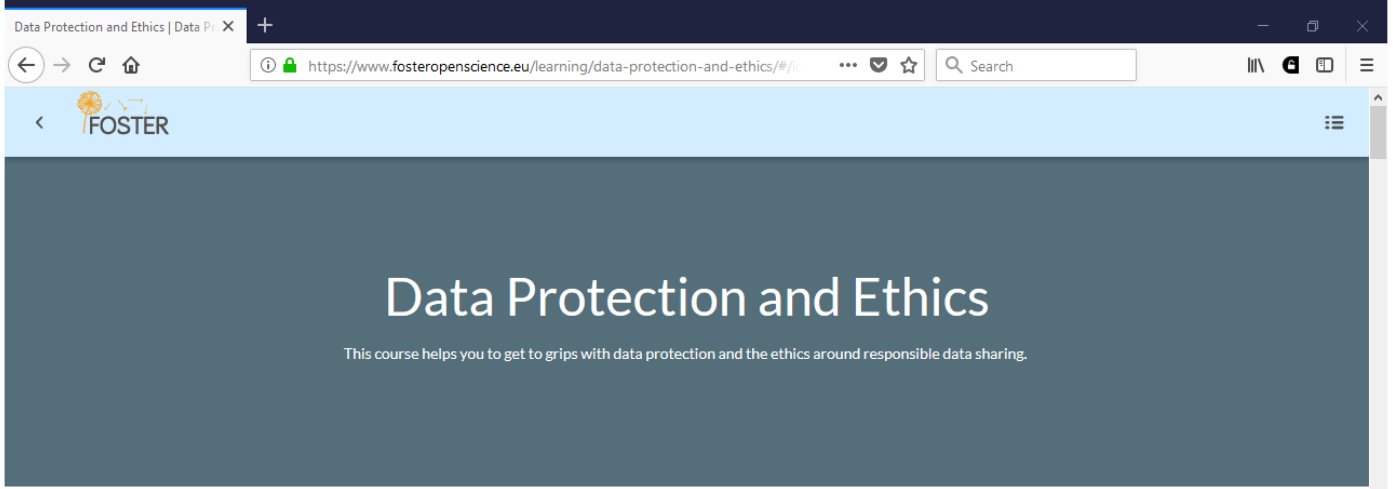
FOSTER

FAIR data are:

Please select one of the options below.

- Findable, accurate, interoperable, reusable
- Findable, accessible, interoperable, replicable
- Findable, accessible, interoperable, reusable
- Formatted, accessible, interoperable, reusable


Managing and Sharing Research Da
https://www.fosteropenscience.eu/learning/managing-and-sharing-r
133%
Search
FOSTER
To get the most out of the data you share, you should make it available...
Please select one of the options below.
from my personal web page
from a discipline specific repository
Submit
Show feedback



https://www.fosteropenscience.eu/learning/data-protection-and-ethics

Introduction

This course covers data protection in particular and ethics more generally. It will help you understand the basic principles of data protection and introduces techniques for implementing data protection in your research processes. Upon completing this course, you will know:

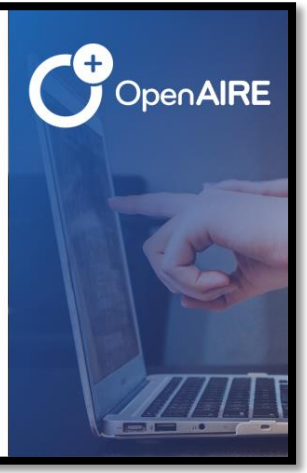
- what personal data are and how you can protect them
- what to consider when developing consent forms
- how to store your data securely
- how to anonymise your data

Data protection and ethics

How do I know if my research data is protected?

- What is Research Data?
- Protection of Research Data
- When is research data protected?
- How do rules on research data impact on use?
- Sui Generis Database Right (SGDR)
- Copyright
- What is Copyright law?
- How can copyright protected works be used?

www.openaire.eu/how-do-i-know-if-my-research-data-is-protected



How do I license my research data?

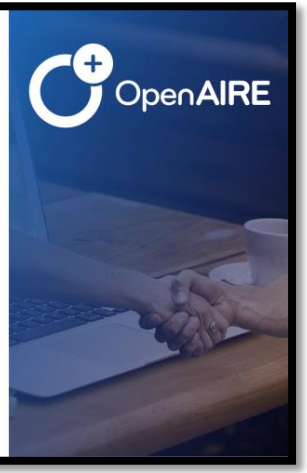
- Licenses for Research Data
- What licence should be applied to the research data?
- What is a Creative Commons licence?
- How to apply licenses for Research Data
- How are licences applied to research data?
- How can I make sure others cite me as the source for my research?
- Specifications of licensing Research Data
- Is there any part of the research data that cannot be made available?
- How should I licence my data for the purposes of Open Science?

www.openaire.eu/how-do-i-license-my-research-data

Can I reuse someone else's research data?

- How can a protected dataset be used?
- Where are licenses found?
- Interoperability and stacking
- What happens if I use 'Share Alike' (SA) licensed material in my work? Does that mean I have to make my work available under the same SA license?
- Can a dataset be used if there is no licence?
- What are the risks of using a dataset without a license?

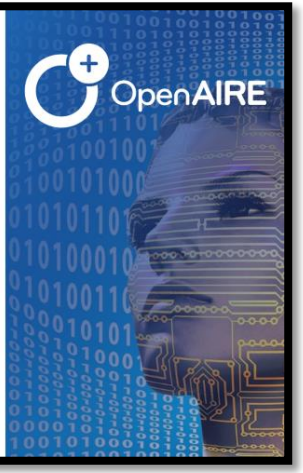
www.openaire.eu/can-i-reuse-someone-else-research-data



How to deal with sensitive data

- What is Sensitive data
- How to prepare sensitive data for storage and sharing?
- Storing sensitive data

www.openaire.eu/sensitive-data-guide



https://www.openaire.eu/guides

How do I know if my research data is protected?

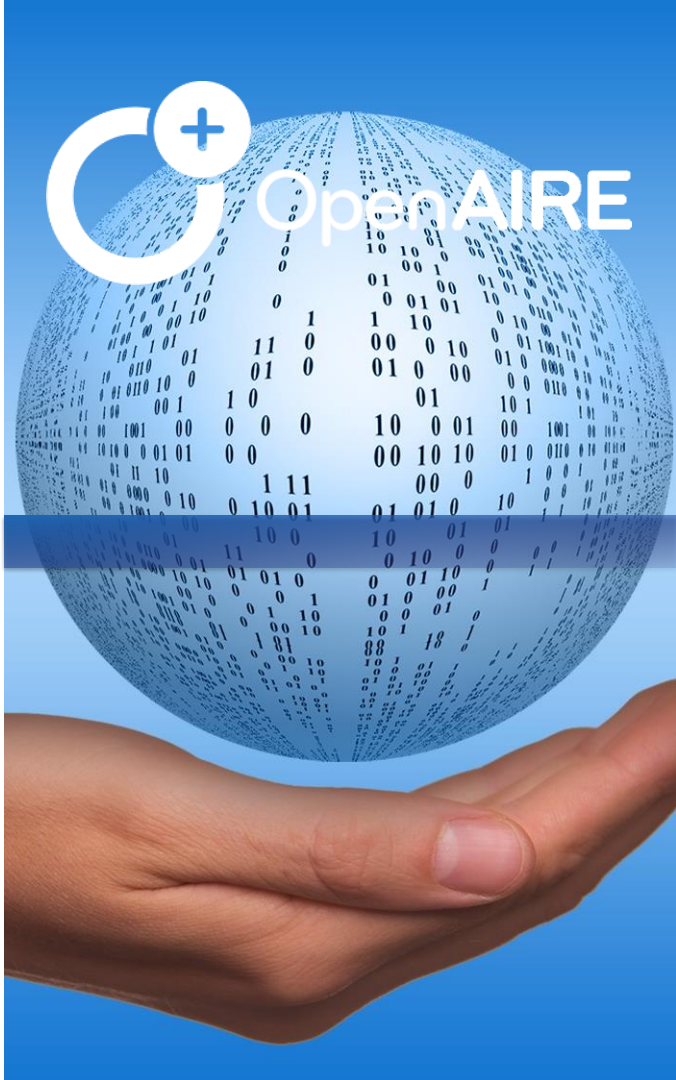
- What is Research Data?
- Protection of Research Data
- When is research data protected?
- How do rules on research data impact on use?
- Sui Generis Database Right (SGDR)
- Copyright
- What is Copyright law?
- How can copyright protected works be used?

www.openaire.eu/how-do-i-know-if-my-research-data-is-protected

How do I license my research data?

- Licenses for Research Data
- What licence should be applied to the research data?
- What is a Creative Commons licence?
- How to apply licenses for Research Data
- How are licences applied to research data?
- How can I make sure others cite me as the source for my research?
- Specifications of licensing Research Data
- Is there any part of the research data that cannot be made available?
- How should I licence my data for the purposes of Open Science?

www.openaire.eu/how-do-i-license-my-research-data



Can I reuse someone else's research data?

- How can a protected dataset be used?
- Where are licenses found?
- Interoperability and stacking
- What happens if I use 'Share Alike' (SA) licensed material in my work? Does that mean I have to make my work available under the same SA license?
- Can a dataset be used if there is no licence?
- What are the risks of using a dataset without a license?

www.openaire.eu/can-i-reuse-someone-else-research-data

How to deal with sensitive data

- What is Sensitive data
- How to prepare sensitive data for storage and sharing?
- Storing sensitive data

www.openaire.eu/sensitive-data-guide

Open Licensing | Open Licensing

https://www.fosteropenscience.eu/learning/open-licensing/#/id/5b618775

Search

FOSTER

Open Licensing

Open Licensing course

https://www.fosteropenscience.eu/learning/open-licensing

Introduction

Licensing your research outputs is an important part of practicing Open Science. In this course, you will:

- know what licenses are, how they work, and how to apply them
- understand how different types of licenses can affect research output reuse
- know how to select the appropriate license for your research

Why do you need apply a license?

Licensing is an important aspect of practising Open Science. By applying licenses to your outputs, you remove any ambiguity over what others can - and can't - do with your work.

An open license, Creative Commons or any other open license, consists of different elements that can be combined. Each element consists of a condition that needs to be followed by the re-user. The different combinations allow for great variation in the type of open license you apply: some being very open, others being very restrictive.

Open licenses

Guidelines on DMPs

How to develop a DMP www.dcc.ac.uk/resources/how-guides/develop-data-plan

RDM brochure and template https://dans.knaw.nl/en/about/organisation-and-policy/information-material?set_language=en

OpenAIRE RDM Handbook https://www.openaire.eu/rdm-handbook

With thanks to

Joy Davidson, University of Glasgow
Marjan Grootveld, DANS
Sarah Jones, DCC
Acknowledgements:
Jonathan Rans, DCC
Thanks to DANS and DCC for reuse of slide

Thank you!

Iryna Kuchma

iryna.kuchma@eifl.net