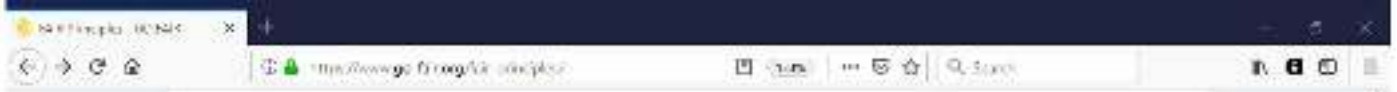
FAIR Data for Open Science

Iryna Kuchma, EIFL Open Access Programme Manager, @irynakuchma

The Role of Library in Open Science and Open Access, EKO-KONNECT USERS CONFERENCE & AGM 2019, 21 January 2019, University of Lago

PUBLICATIONS AND DATA


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



Findable

The first step in (re)using data is to find them. Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services, so this is an essential component of the FAIRification process.

F1. (Meta)data are assigned a globally unique and persistent identifier

F2. Data are described with rich metadata

F3. Metadata clearly and explicitly include the identifier of the data they describe

F4. (Meta)data are registered or indexed in a searchable resource

Accessible

Once the user finds the required data, she/he needs to know how can they be accessed, possibly including authentication and authorisation.

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

https://www.go-fair.org/fair-principles

Interoperable

The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.

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

The ultimate goal of FAIR is to optimise the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.

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

The principles refer to three types of entities: data (or any digital object), metadata (information about that digital object), and infrastructure. For instance, principle F4 defines that both metadata and data are registered or indexed in a searchable resource (the infrastructure component).

What does this mean?

Principle F1 is arguably the most important because it will be hard to achieve other aspects of FAIR without globally unique and persistent identifiers. Hence, compliance with F1 will already take you a long way towards publishing FAIR data (see **10 ways identifiers can help with data integration**).

Globally unique and persistent identifiers remove ambiguity in the meaning of your published data by assigning a unique identifier to every element of metadata and every concept/measurement in your dataset. In this context, identifiers consist of an internet link (e.g., a URL that resolves to a web page that defines the concept such as a particular **human protein**). Many data repositories will automatically generate globally unique and persistent identifiers to deposited datasets. Identifiers can help other people understand exactly what you mean, and they allow computers to interpret your data in a meaningful way (i.e., computers that are searching for your data or trying to automatically integrate them). Identifiers are essential to the human-machine interoperation that is key to the vision of **Open Science**. In addition, identifiers will help others to properly cite your work when reusing your data.

Of course, identifiers are one thing, but their meaning is another (see principles I1-I3). I1 stipulates two conditions for your identifier:

1. It must be globally unique (i.e., someone else could not reuse/reassign the same identifier without referring to your data). You can obtain globally unique identifiers from a registry service that uses algorithms guaranteeing the uniqueness of newly minted identifiers.
2. It must be persistent. It takes time and money to keep web links active, so links tend to become invalid over time. Registry services guarantee resolvability of that link into the future, at least to some degree.

Examples of globally unique and persistent identifiers

- One particular person on planet earth has this globally unique and persistent identifier: **https://orcid.org/0000-0001-8888-635X**
- Here is an identifier that uniquely links to the results of a study estimating the FAIRness of different data repositories: **doi:10.4121/uuid:5146dd06-98e4-426c-9ae5-dc8fa65c549f**
- The human polycystin-1 protein has a globally unique and persistent identifier given by the UniProt database: **http://www.uniprot.org/uniprot/P98161**
- Polycystic kidney disease type 1 has a globally unique and persistent identifier given by the OMIM database: **http://omim.org/entry/173900**
- The number 163485 refers to the undergraduate student ID of **Mark Wilkinson**, the NCBI gi number for a bovine protease, and a part number for a Singer sewing machine. Hence, this is a poor example of GUID

Example services that supply globally unique and persistent identifiers

- Identifiers.org provides resolvable identifiers in the form of URIs and CURIEs: **http://identifiers.org**
- Universally unique identifier: **https://en.wikipedia.org/wiki/Universally_unique_identifier**
- Persistent URLs: **http://www.purlz.org**
- Digital Object Identifier: **http://www.doi.org**
- Archival Resource Key: **https://escholarship.org/uc/item/9p9863nc**
- Research Resource Identifiers: **https://scicrunch.org/resources**
- Identifiers for funding organisations (see F3 & R1): **https://www.crossref.org/services/funder-registry/**
- Identifiers for the world's research organisations (see F3 & R1): **https://www.grid.ac**

Definition of Open Data

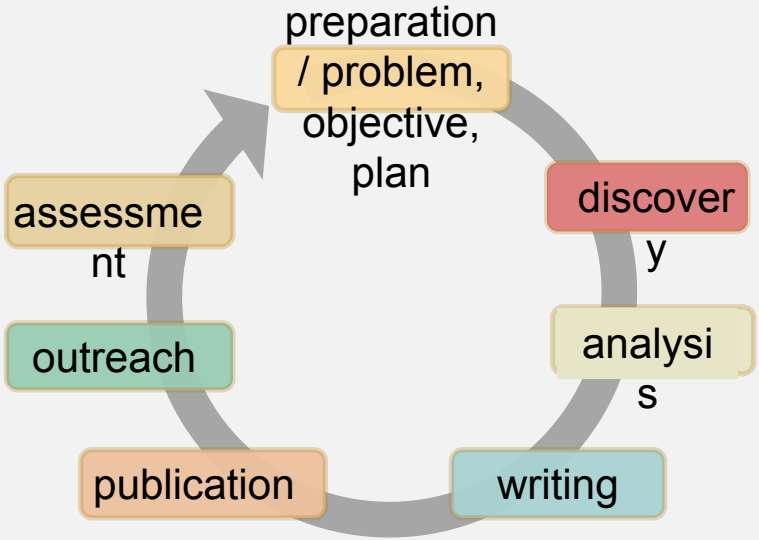
Open Data are online, free of cost, accessible data that can be used, reused and distributed provided that the data source is attributed.

Tip – use 5 Star Open Data Model to explain FAIR

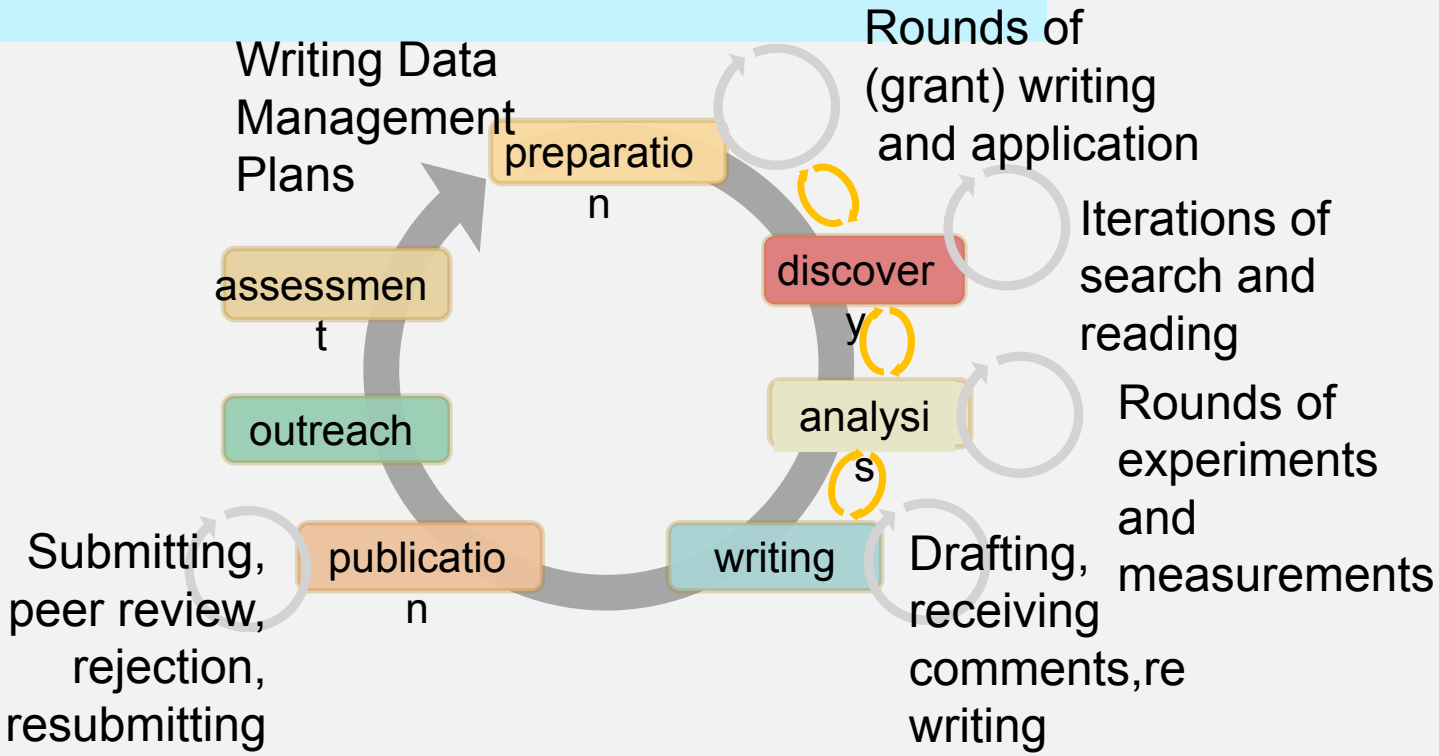
- ★ make your stuff available on the Web (whatever format) under an open license
- ★★ make it available as structured data (e.g., Excel instead of image scan of a table)
- ★★★ make it available in a non-proprietary open format (e.g., CSV instead of Excel)
- ★★★★ use URIs to denote things, so that people can point at your stuff
- ★★★★★ link your data to other data to provide context

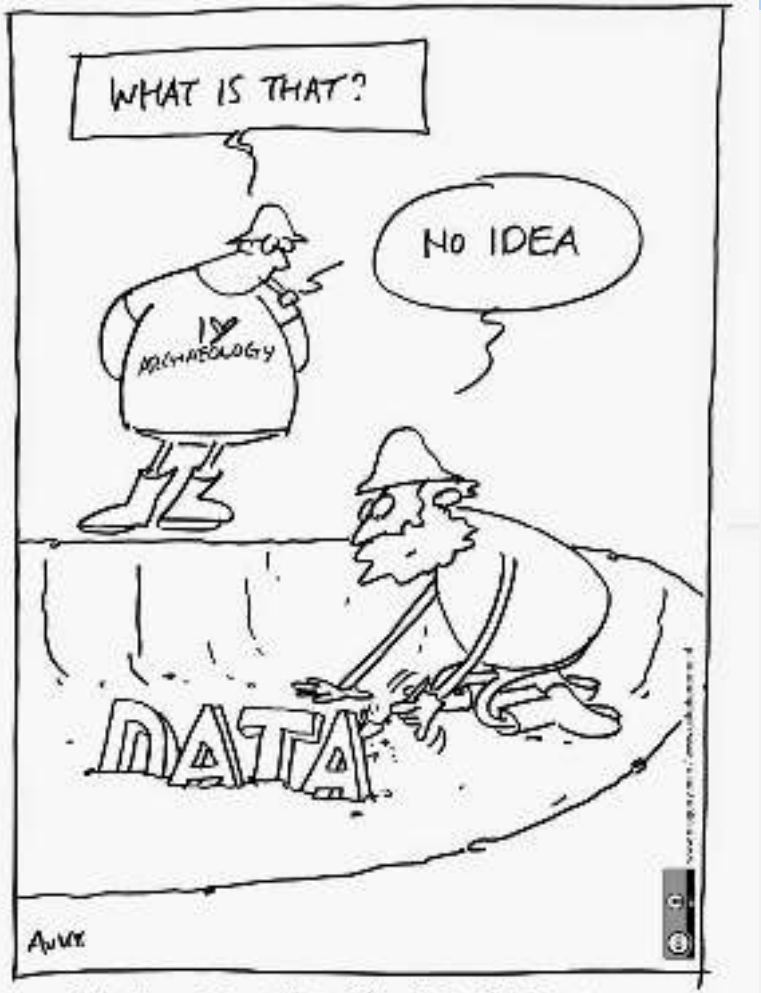
Tim Berners-Lee's proposal for five star open data - http://5stardata.info

A model of the research workflow



A model of the research workflow





DATA FOR FUTURE GENERATIONS

WHY SHOULD YOU BE OPEN?

Image by wonderwebby CC-BY-NC-SA www.flickr.com/photos/wonderwebby/2723279491

It’s part of good research practice

"It was *never* acceptable to publish papers without making data available."

- Ewan Birney

#OpenData
#OpenScience

Original image via doi:10.1038/461145a. “Research cannot flourish if data are not preserved and made accessible. Data management should be woven into every course in science.” - *Nature* 461, 145

FOSTER

Cut down on academic fraud

www.nature.com/news/2011/111101/full/479015a.html

FOSTER

Validation of results

"It was a mistake in a spreadsheet that could have been easily overlooked: a few rows left out of an equation to average the values in a column.

The spreadsheet was used to draw the conclusion of an influential 2010 economics paper: that public debt of more than 90% of GDP slows down growth. This conclusion was later cited by the International Monetary Fund and the UK Treasury to justify programmes of austerity that have arguably led to riots, poverty and lost jobs."

The error that could subvert George Osborne's austerity programme

The theories on which the chancellor based his cuts policies have been shown to be based on an embarrassing mistake

Charles Arthur and Philip Inman
The Guardian, Thursday 18 April 2013 21.18 BST

George Osborne says that Ken Rogoff, the man whose economic error has been uncovered, has strongly influenced his thinking. Photograph: Stefan Wermuth/PA

www.guardian.co.uk/politics/2013/apr/18/uncovered-error-george-osborne-austerity

FOSTER

More scientific breakthroughs

Sharing of Data Leads to Progress on Alzheimer's

In 2005, a group of scientists and executives from the National Institutes of Health, the Food and Drug Administration, the drug and medical imaging industries, universities and nonprofit groups joined in a project that experts say had no precedent: a collaborative effort to find the biological markers that show the progression of Alzheimer's disease in the human brain.

Now, the effort is bearing fruit with a wealth of recent scientific papers on the early diagnosis of Alzheimer's using methods like PET scans and tests of spinal fluid. More than 100 studies are under way to test drugs that might slow or stop the disease.

And the collaboration is already serving as a model for similar efforts against Parkinson's disease. A $40 million project to look for biomarkers for Parkinson's, sponsored by the Michael J. Fox Foundation, plans to enroll 800 study subjects in the United States and Europe.

"It was unbelievable. Its not science the way most of us have practiced in our careers. But we all realised that we would never get biomarkers unless all of us parked our egos and intellectual property noses outside the door and agreed that all of our data would be public immediately."

Dr John Trojanowski, University of Pennsylvania

www.nytimes.com/2010/08/13/health/research/13alzheimer.html?pagewanted=all&_r=0

FOSTER

A citation advantage

A study that analysed the citation counts of 10,555 papers on gene expression studies that created microarray data, showed:

"studies that made data available in a public repository received 9% more citations than similar studies for which the data was not made available"

Data reuse and the open data citation advantage, Piwowar, H. & Vision, T. https://peerj.com/articles/175

FOSTER

Increased use and economic benefit

The case of NASA Landsat satellite imagery of the Earth's surface:

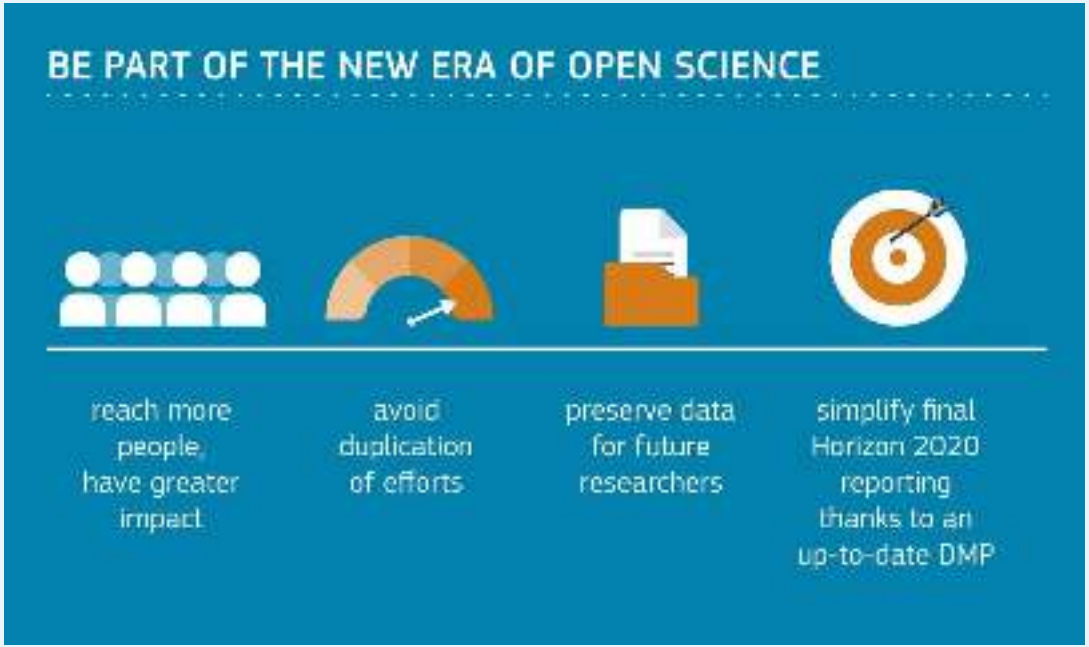
Up to 2008

- Sold through the US Geological Survey for US$600 per scene
- Sales of 19,000 scenes per year
- Annual revenue of $11.4 million

Since 2009

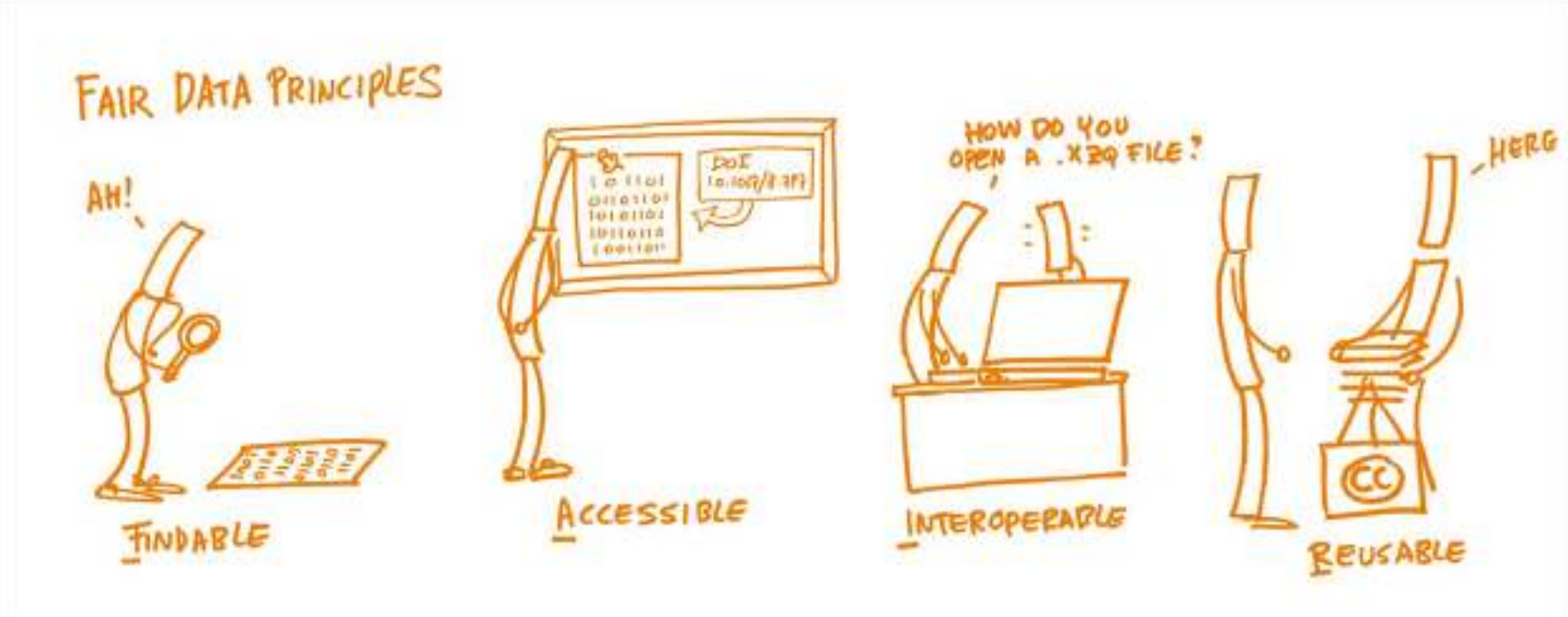
- Freely available over the internet
- Google Earth now uses the images
- Transmission of 2,100,000 scenes per year.
- Estimated to have created value for the environmental management industry of $935 million, with direct benefit of more than $100 million per year to the US economy
- Has stimulated the development of applications from a large number of companies worldwide

http://earthobservatory.nasa.gov/IOTD/view.php?id=83394&src=ve


BE PART OF THE NEW ERA OF OPEN SCIENCE
reach more people, have greater impact
avoid duplication of efforts
preserve data for future researchers
simplify final Horizon 2020 reporting thanks to an up-to-date DMP

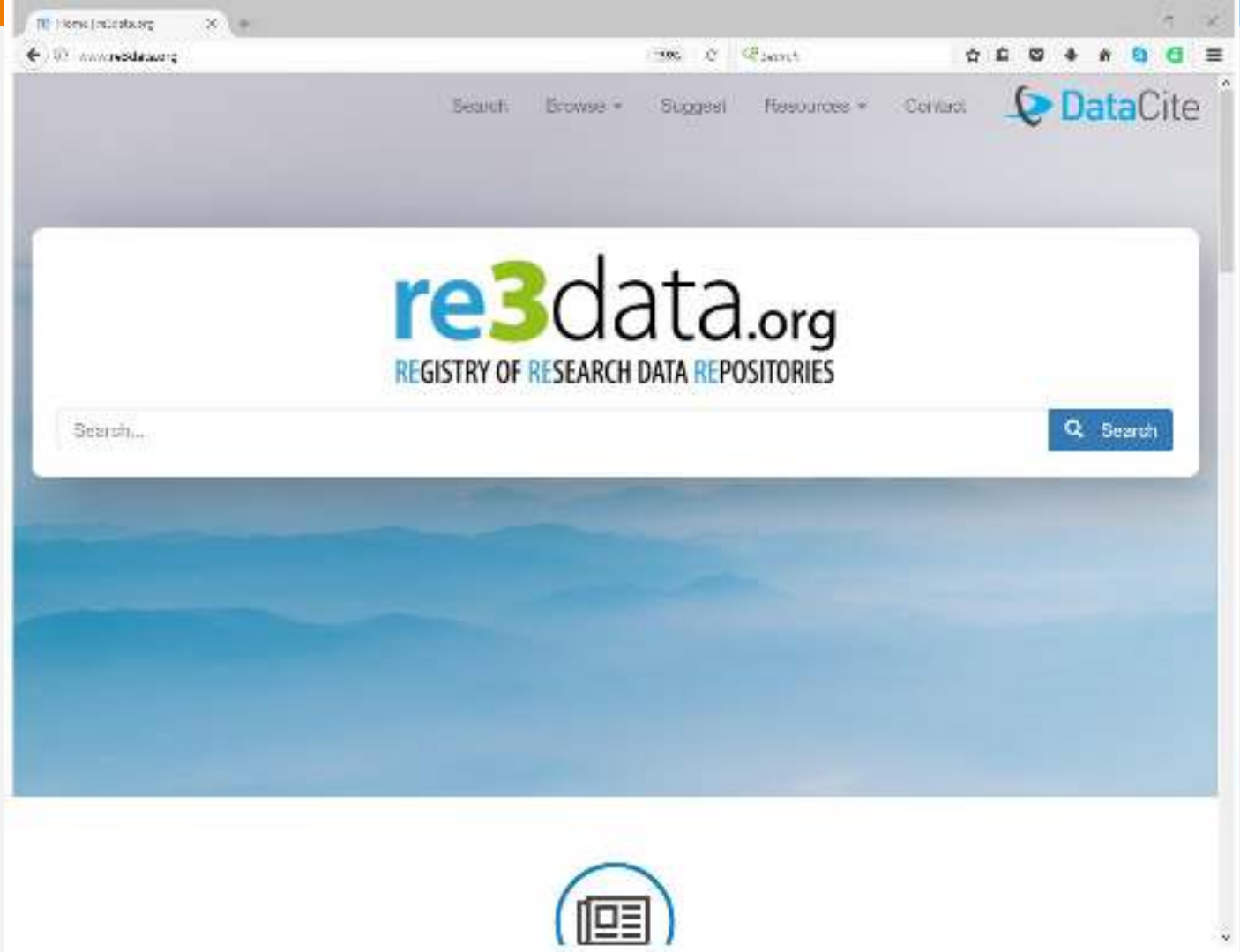
Misconception #1:

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

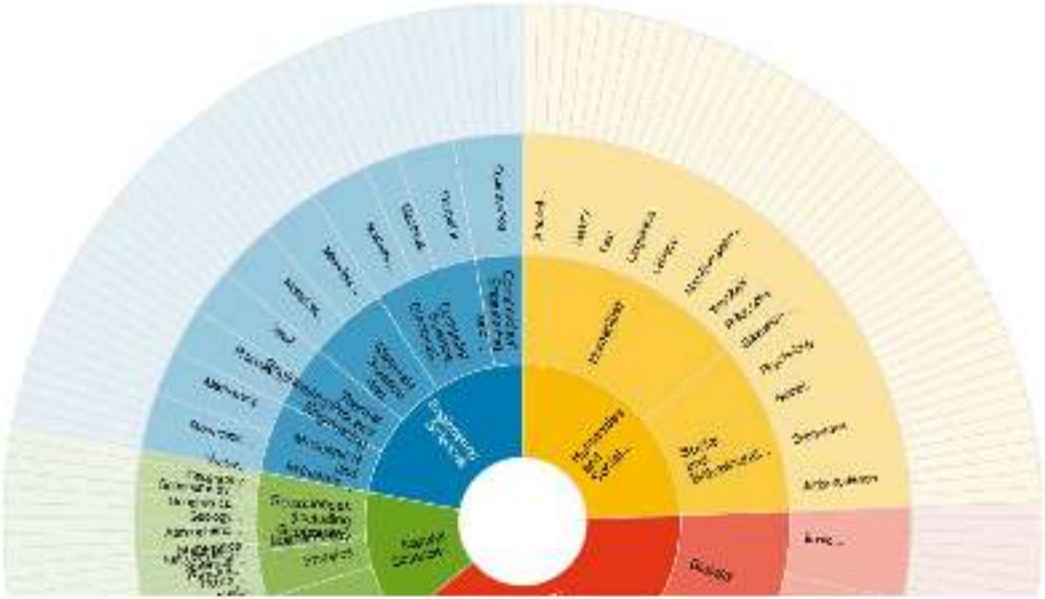

Search
Browse
Suggest
Resources
Contact
DataCite
re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES
Search...
Search

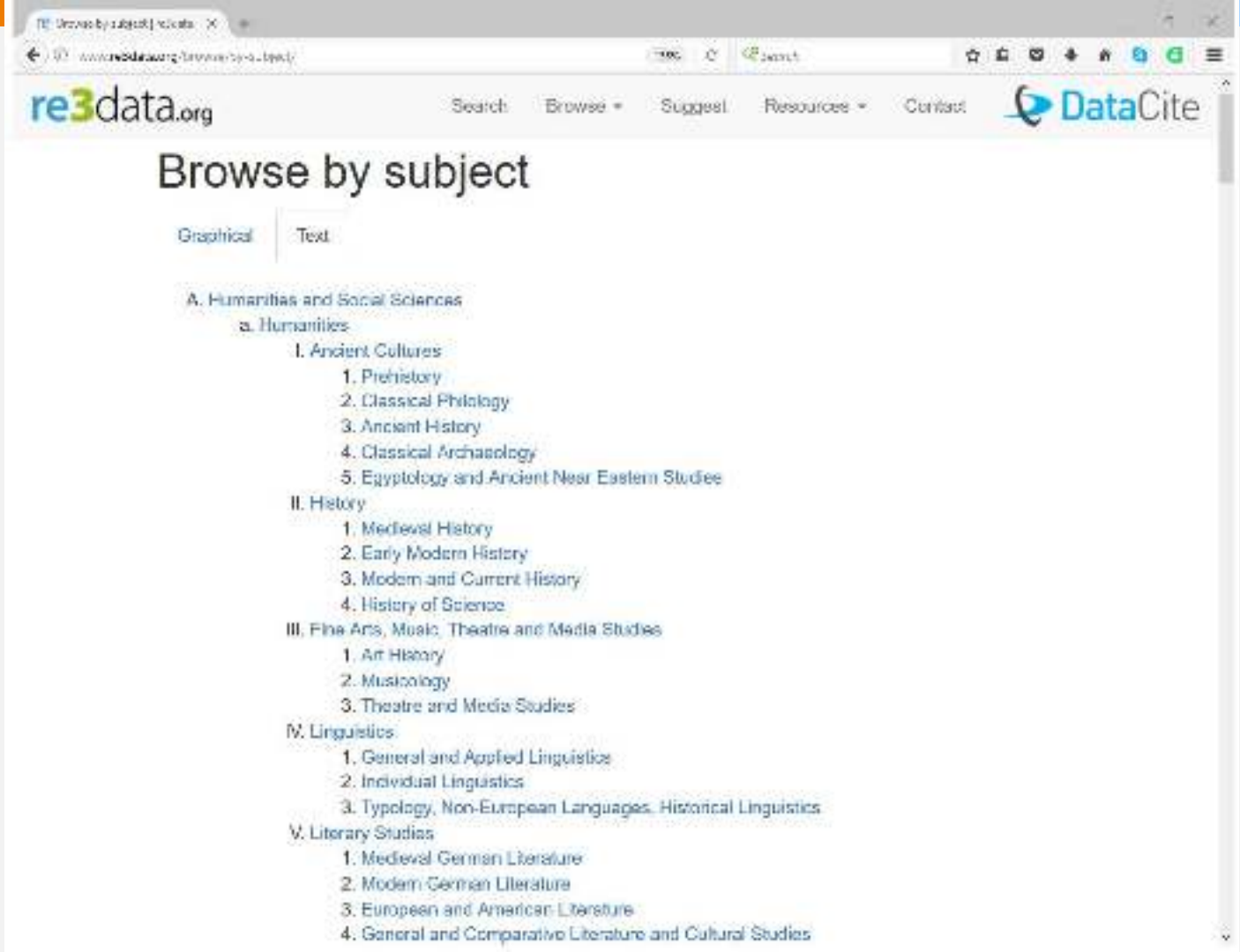
re3data.org Search Browse Suggest Resources Contact DataCite

Browse by subject

Graphical Text

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




re3data.org
Search
Browse
Suggest
Resources
Contact
DataCite
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

re3data.org Search Browse Suggest Resources Contact DataCite

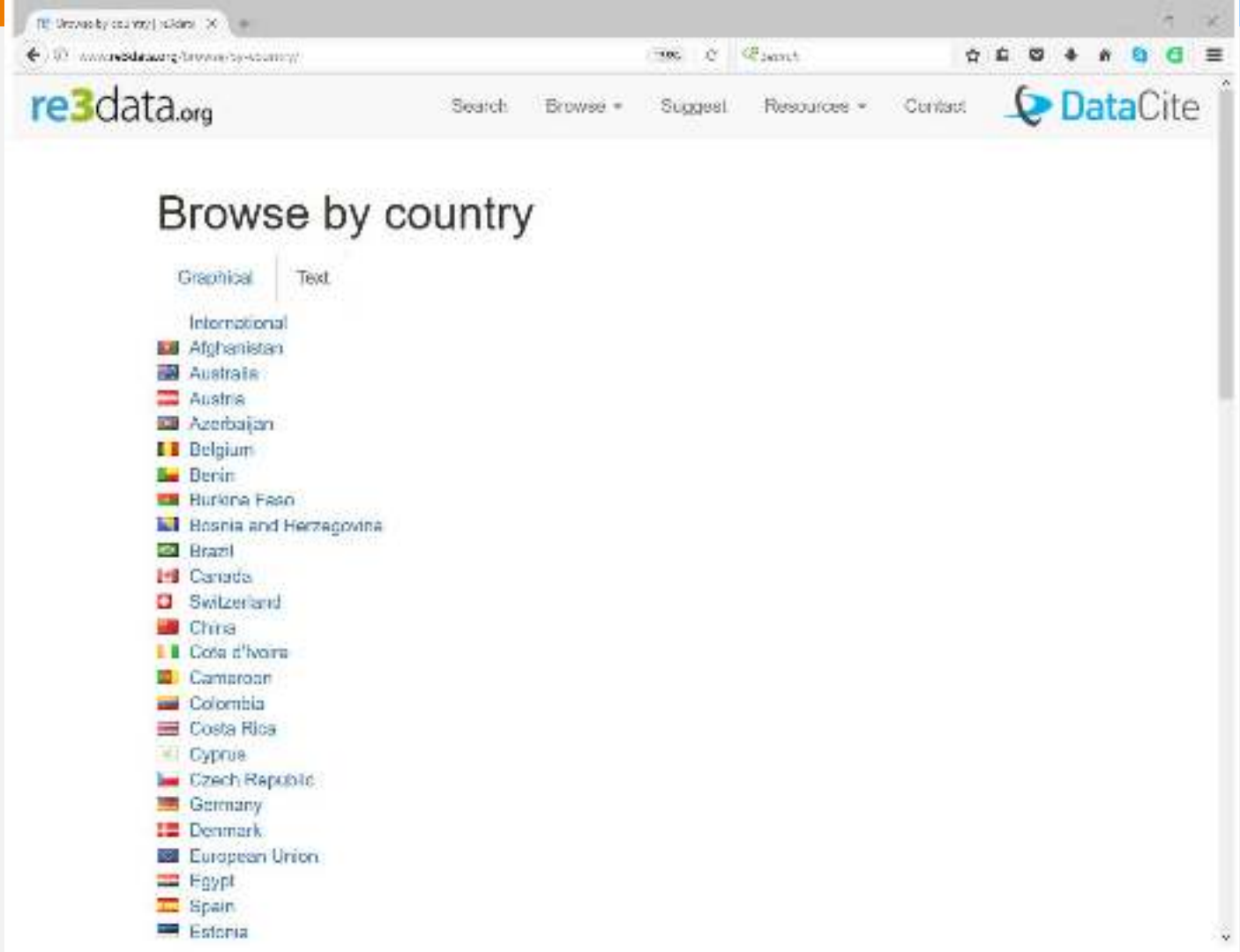
Browse by content type

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

re3data.org Search Browse Suggest Resources Contact DataCite

Browse by country

Graphical Text

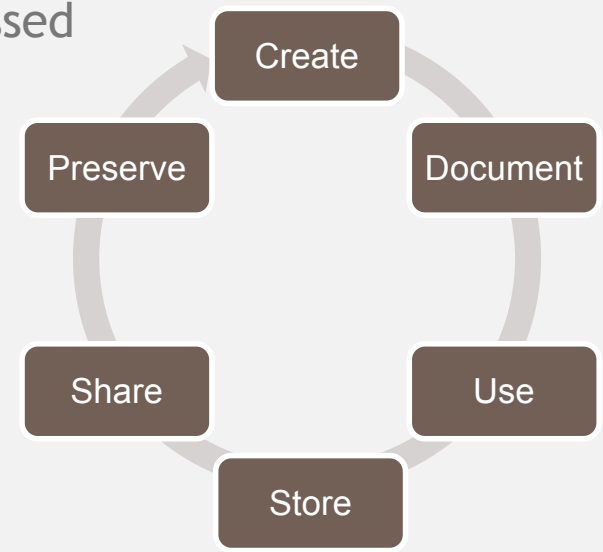

re3data.org
Search
Browse
Suggest
Resources
Contact
DataCite
Browse by country
Graphical
Text
International
Afghanistan
Australia
Austria
Azerbaijan
Belgium
Benin
Burkina Faso
Bosnia and Herzegovina
Brazil
Canada
Switzerland
China
Cote d'Ivoire
Cameroon
Colombia
Costa Rica
Cyprus
Czech Republic
Germany
Denmark
European Union
Egypt
Spain
Estonia

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

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.)

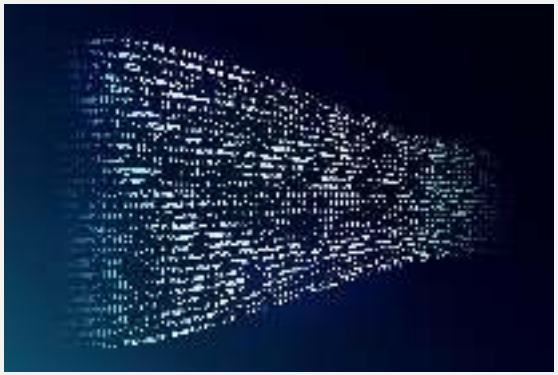
Basically, everything that is needed to replicate a study should be available. Plus everything that is potentially useful for others.

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

Open Data

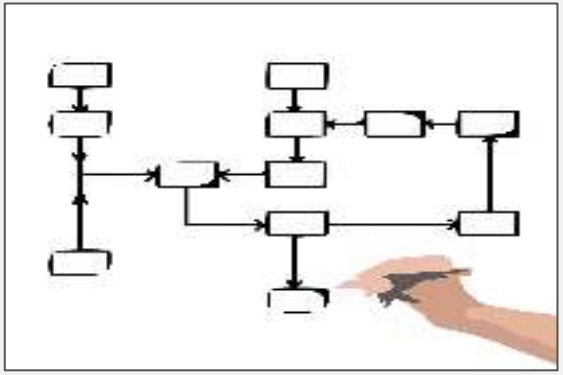
To support validation and facilitate reuse

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?

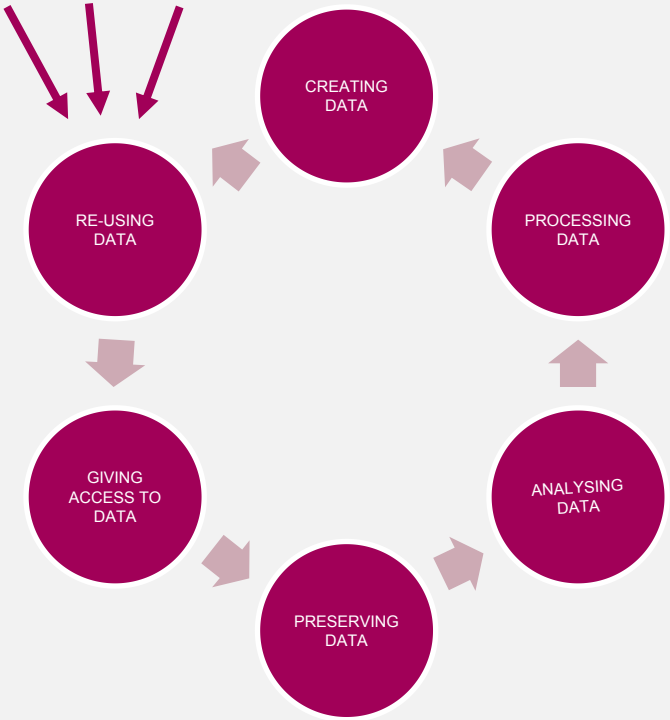
https://okfn.org

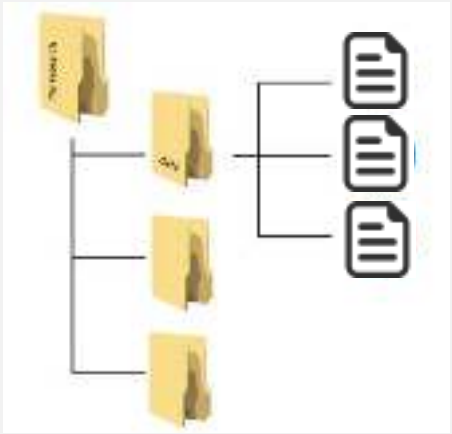
1. Choose your dataset(s)
   - What can you open? You may need to revisit this step if you encounter problems later.
2. Apply an open license
   - Determine what IP exists. Apply a suitable licence e.g. CC-BY
3. Make the data available
   - Provide the data in a suitable format. Use repositories.
4.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

FOSTER

Planning trick: think backwards

What data organisation would a re-user like?





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

(All) Research. Shared.

— your one stop research shop!

All research outputs from across all fields of science are welcome! Zenodo accept any file format as well as both positive and negative results. However, we do promote peer-reviewed openly accessible research, and we do curate your upload before putting it on the front-page.

Citeable. Discoverable.

— be found!

Zenodo assigns all publicly available uploads a Digital Object Identifier (DOI) to make the upload easily and uniquely citeable. Zenodo further supports harvesting of all content via the OAI-PMH protocol.

Community Collections

— create your own repository

Zenodo allows you to create your own collection and accept or reject all uploads to it. Creating a space for your next workshop or project have never been easier. Plus, everything is citeable and discoverable.

www.zenodo.org

Safe

— more than just a drop box!

Your research output is stored safely for the future in same cloud infrastructure as research data from CERN's Large Hadron Collider using a CERN's battle-tested repository software INVENIO used by some of the world's largest repositories such as INSPIRE HEP and CERN Document Server.

Reporting

— tell your funding agency!

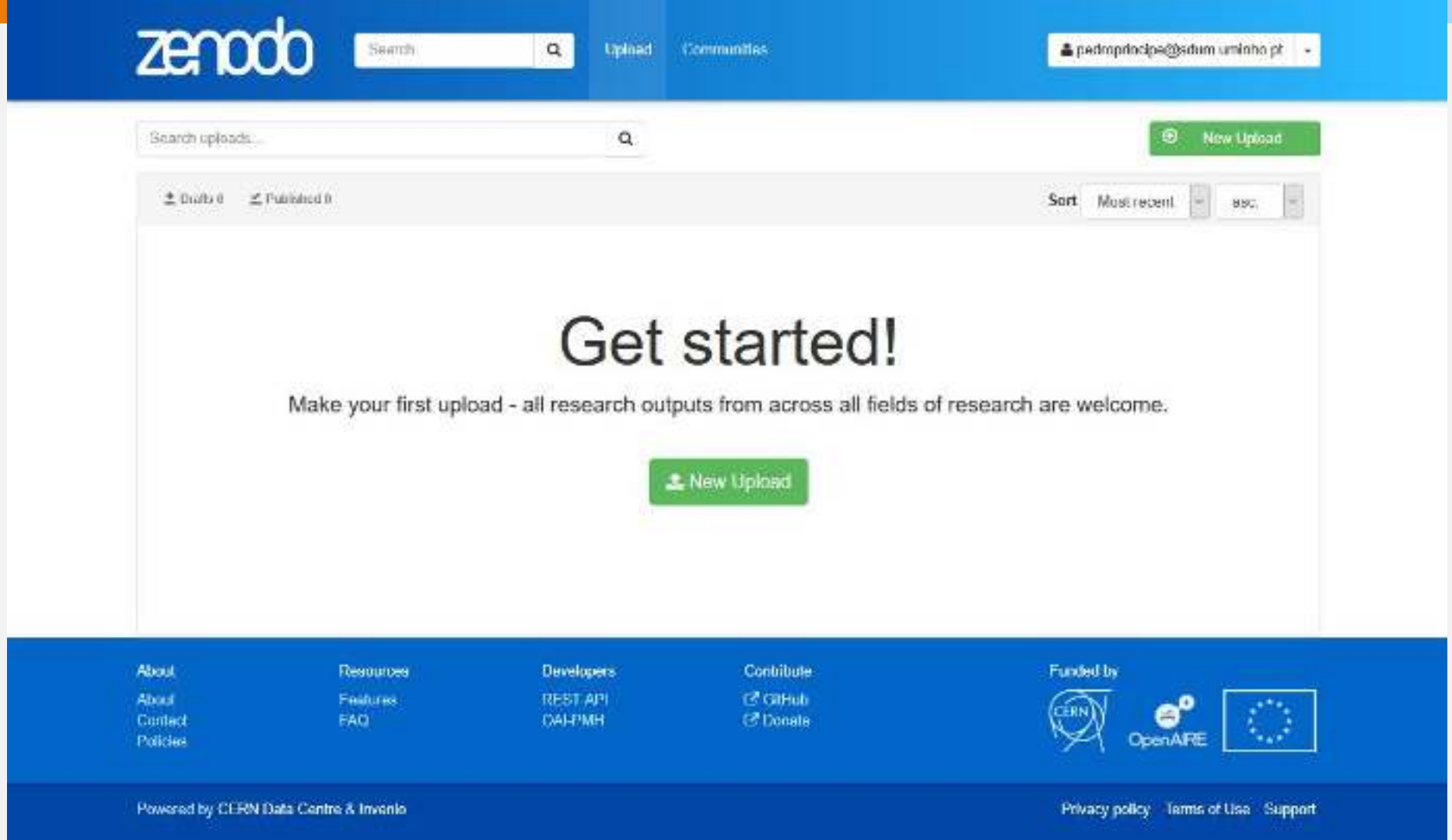
Zenodo is integrated into reporting lines for research funded by the European Commission via OpenAIRE. Just upload your research on Zenodo and we will take care of the reporting for you. We plan to extend with futher funding agencies in the future so stay tuned!

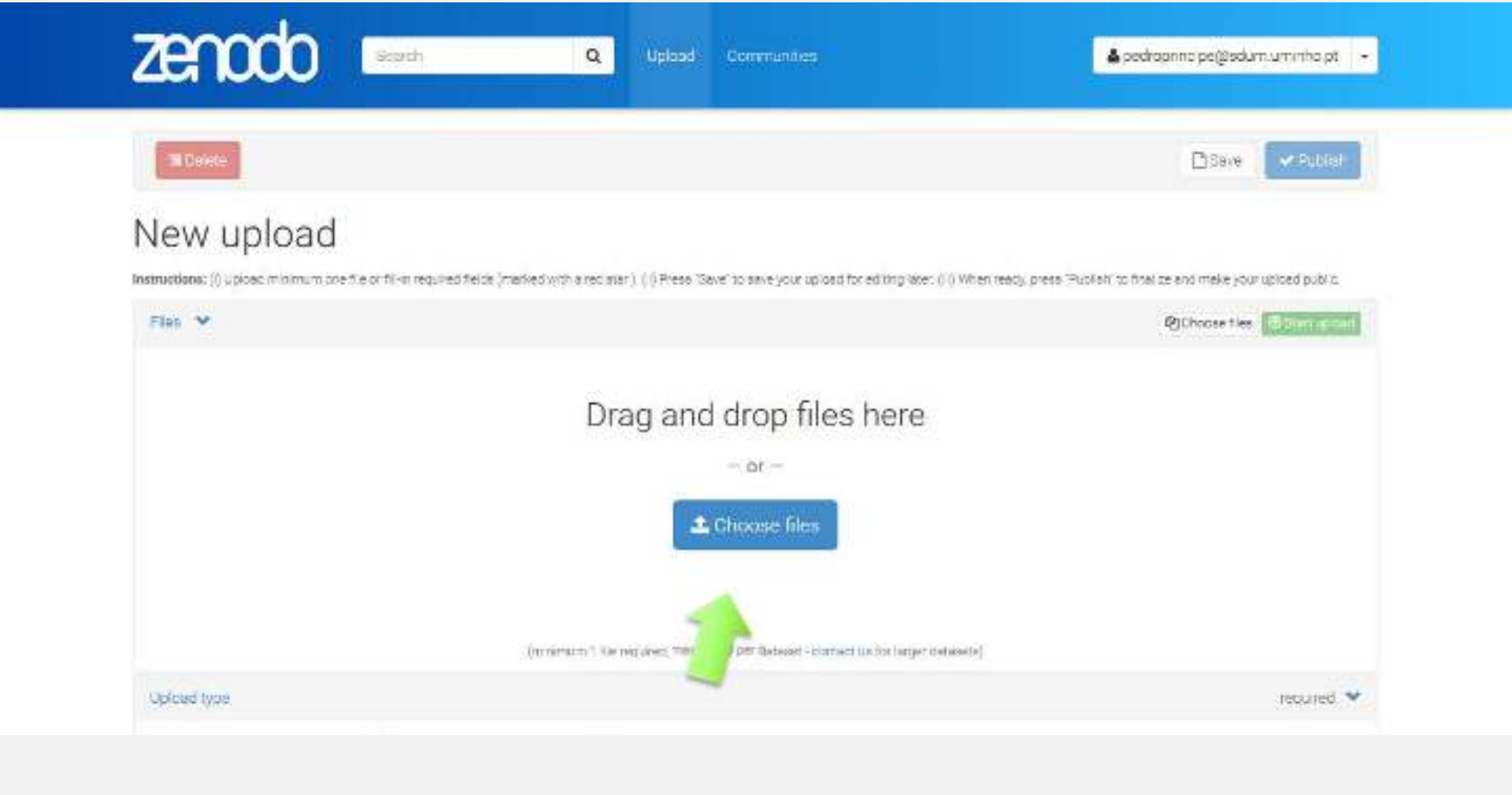
Flexible Licensing

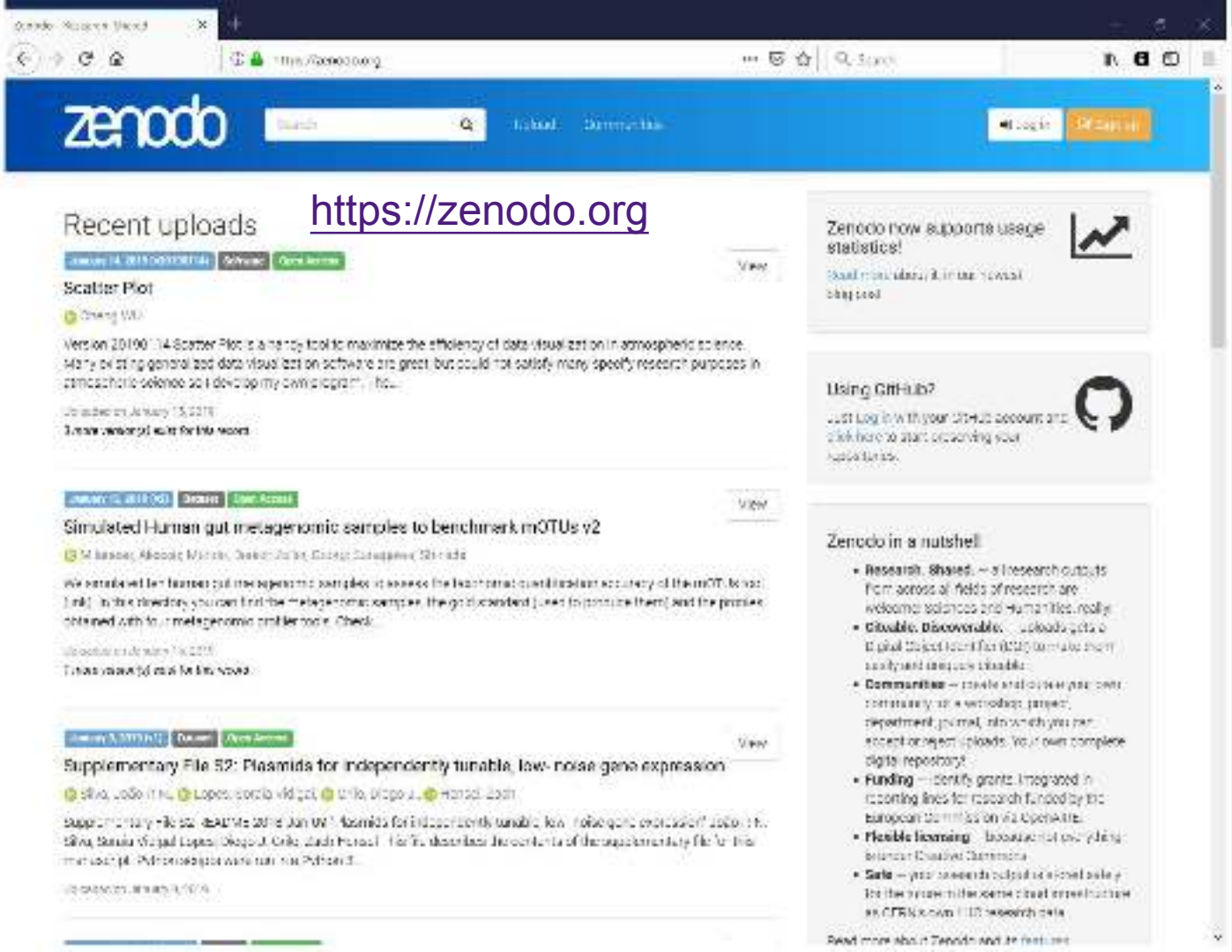
— not everything is under Creative Commons

Zenodo encourage you to share your research as openly as possible to maximize use and re-use of your research results. However, we also acknowledge that one size does not fit all, and therefore allow for uploading under a multitude of different licenses and access levels*.

* You are responsible for respecting applicable copyright and license conditions for the files you upload.


zenodo
Search
Upload
Communities
pedroprioclpe@dium.uminho.pt
Search uploads...
New Upload
Drafts 0
Published 0
Sort
Most recent
asc.
Get started!
Make your first upload - all research outputs from across all fields of research are welcome.
New Upload
About
About
Contact
Policies
Resources
Features
FAQ
Developers
REST API
OAI-PMH
Contribute
GitHub
Donate
Funded by
CERN
OpenAIRE
Powered by CERN Data Centre & Invenio
Privacy policy
Terms of Use
Support


zenodo
Search
Upload
Communities
Delete
Save
Publish
New upload
Files
Choose files
Drag and drop files here
— or —
Choose files
Upload type
required


zenodo
https://zenodo.org
Recent uploads
Scatter Plot
Simulated Human gut metagenomic samples to benchmark mOTUs v2
Supplementary File S2: Plasmids for independently tunable, low-noise gene expression
Zenodo now supports usage statistics!
Using GitHub?
Zenodo in a nutshell
Research. Shared.
Citable. Discoverable.
Communities
Funding
Flexible licensing
Safe

Licensing research data

Horizon 2020 Open Access guidelines point to:

 or 

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

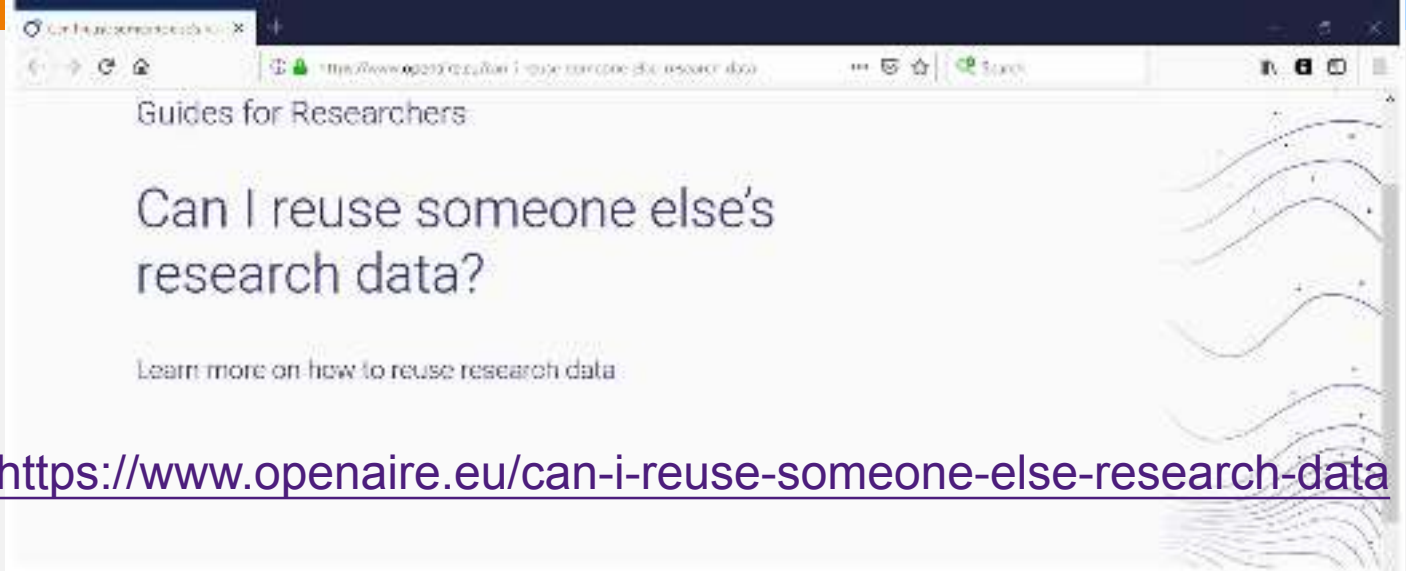
www.dcc.ac.uk/resources/how-guides/license-research-data

Guides for Researchers

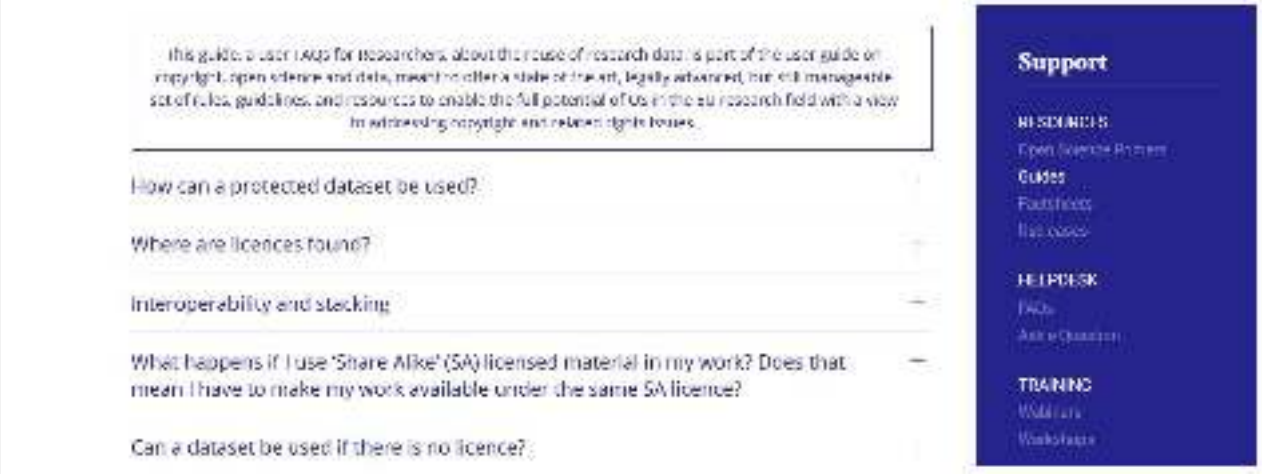
How do I license my research data?

Learn more about licenses for research data and how to apply it

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



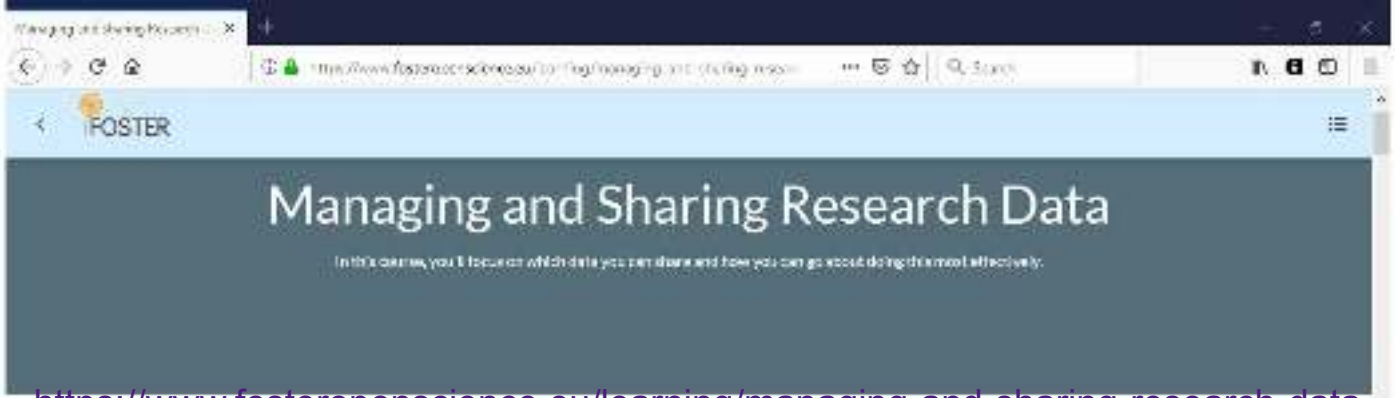
https://www.openaire.eu/can-i-reuse-someone-else-research-data



EUDAT licensing tool

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

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



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

Introduction

Data-driven research is becoming increasingly common in a wide range of academic disciplines, from Archaeology to Zoology, and spanning Arts and Science subject areas alike. To support good research, we need to ensure that researchers have access to good data. Upon completing this course, you will:

- understand the differences between open, closed, and shared data
- be able to make decisions about which data you can share
- know what a data management plan is
- be aware of the FAIR principles
- know how to get maximum impact from sharing your research data

What are research data?

The University of Leeds describes research data as 'any information that has been collected, observed, generated or created to validate original research findings.' Research data can include things like...

Click the arrows to navigate through the content

...raw data

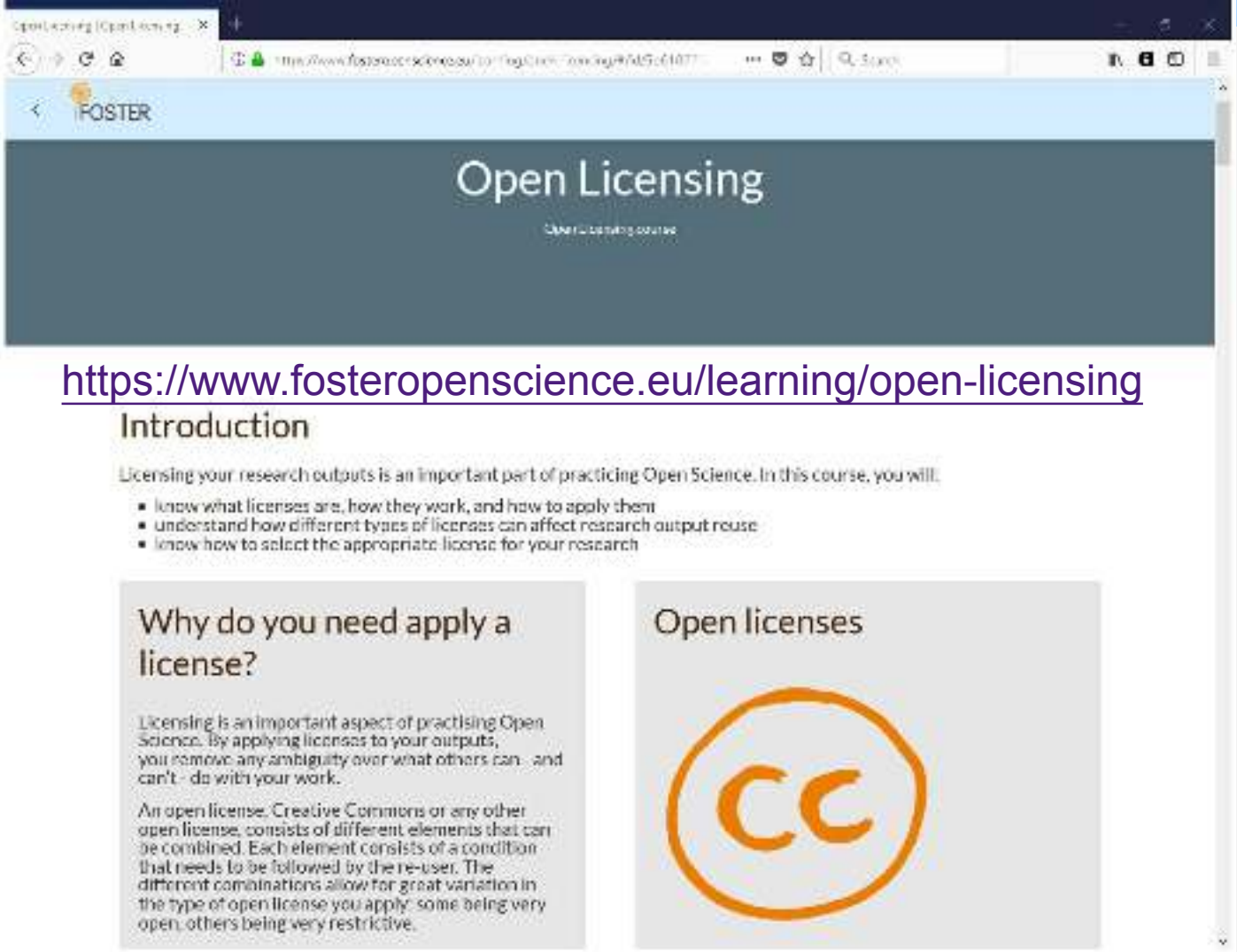
Raw data are those which are captured from instruments and

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

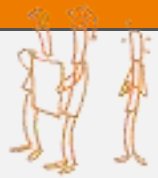
Introduction

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


FOSTER
Open Licensing
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

Self-learning courses

What is Open Science?	Best Practice in Open Research	Open Access Publishing	Open Peer Review	Sharing Preprints
				
Data Protection & Ethics	**Open Source Software & Workflows**	**Managing & Sharing Research Data**	**Open Science & Innovation**	**Open Licensing**
				

www.fosteropenscience.eu/toolkit

FOSTER

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
Bianca Kramer & Jeroen Bosman, Utrecht University Library
Marjan Grootveld, DANS
Sarah Jones, DCC
Acknowledgements:
Jonathan Rans, DCC
Thanks to DANS and DCC for reuse of slide

Thank you! Questions?

iryna.kuchma@eifl.net

Facebook: @fosteropenscience

Twitter: @fosterscience

Youtube: FOSTER Open Science