
# Research Data Management: taster

November 14, 2018, Luxemburg

Gwen Franck


# What is “data”? And why does it need managing?

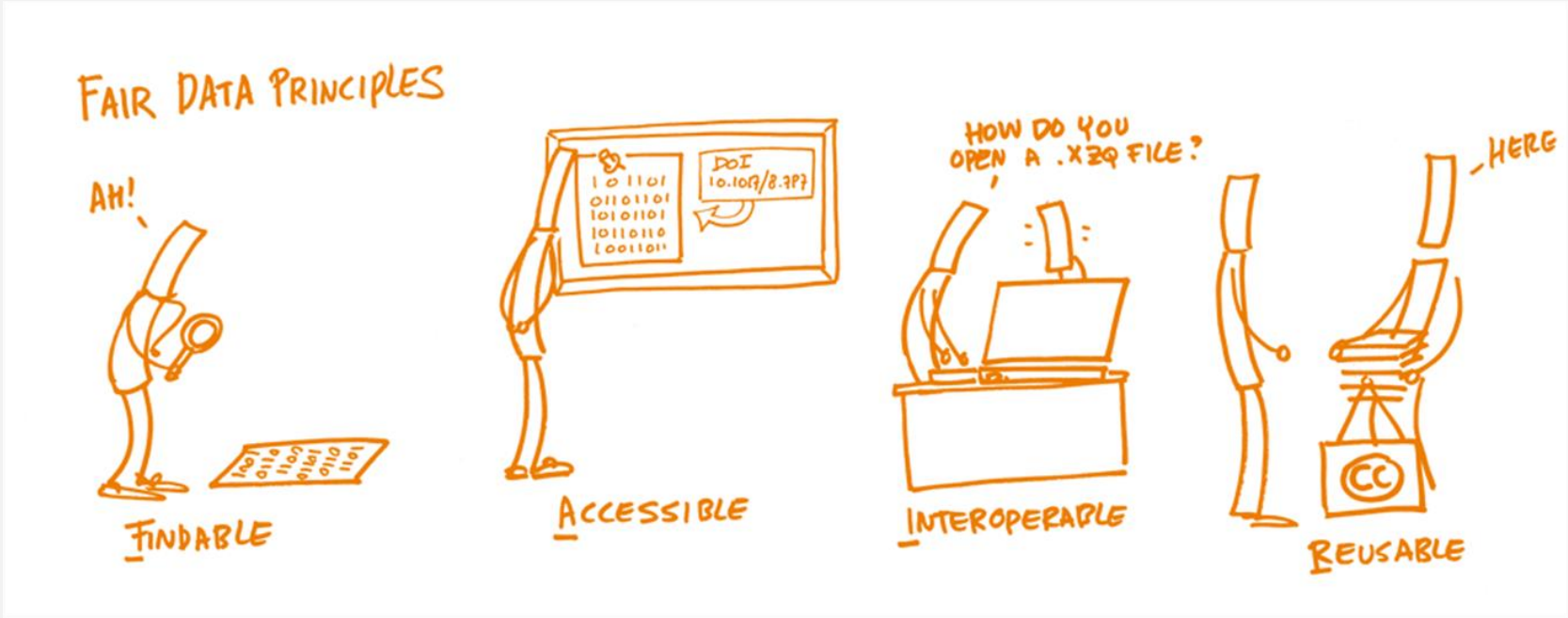

PUBLICATIONS AND DATA

# Tell me: what’s your data?

DATA: 'any information that has been collected, observed, generated or created to validate original research findings'

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.’

So, either I make my data open for everybody or not at all? Is there no middle ground?


FAIR DATA PRINCIPLES
AH!
FINDABLE
DOI
10.1087/8.387
ACCESSIBLE
HOW DO YOU OPEN A .X3Q FILE?
INTEROPERABLE
HERE
CC
REUSABLE

# Making data FAIR

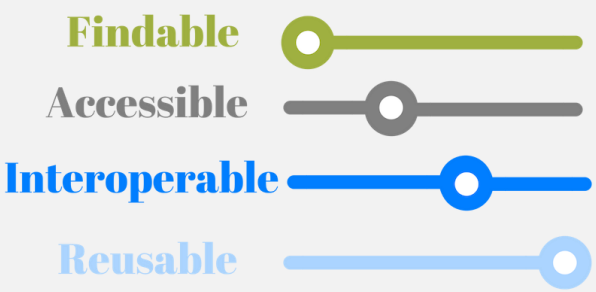
**Findable** – Assign persistent IDs, provide rich metadata, register in a searchable resource,...

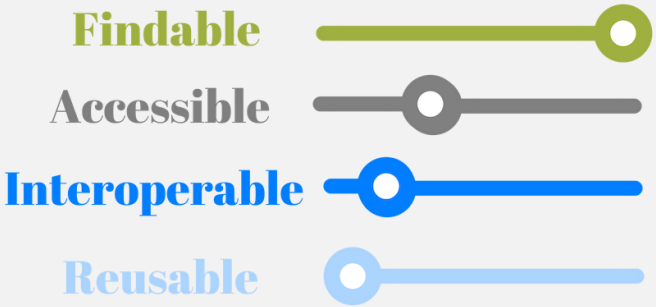
**Accessible** - Retrievable by their ID using a standard protocol, metadata remain accessible even if data aren't...

**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


Findable
Accessible
Interoperable
Reusable


Findable
Accessible
Interoperable
Reusable

# Why manage data?

- 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 my data, ok, but why do I need to make it open?

# 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

nature International weekly journal of science

Login

nature news home | news archive | specials | opinion | features | news blog | nature journal

comments on this story

Stories by subject
- Brain and behaviour
- Lab life

Stories by keywords
- Deiderik Stapel
- Tilburg University
- Academic fraud
- Retractions
- Social psychology

This article elsewhere

Blogs linking to this article

Add to Digg
Add to Facebook
Add to Newsvine
Add to Del.icio.us
Add to Twitter

Published online 1 November 2011 | *Nature* **479**, 15 (2011) |
doi:10.1038/479015a
Updated online: 1 November 2011
Updated online: 8 December 2011

News

## Report finds massive fraud at Dutch universities

**Investigation claims dozens of social-psychology papers contain faked data.**

Ewen Callaway

When colleagues called the work of Dutch psychologist Diederik Stapel too good to be true, they meant it as a compliment. But a preliminary investigative report (go.nature.com/tqmp5c) released on 31 October gives literal meaning to the phrase, detailing years of data manipulation and blatant fabrication by the prominent Tilburg University researcher.

Dutch psychologist Diederik Stapel.
*Persbureau van Eijndhoven*

"We have some 30 papers in peer-reviewed journals where we are actually sure that they are fake, and there are more to come," says Pim Levelt, chair of the committee that investigated Stapel's work at the university.

Stapel's eye-catching studies on aspects of social behaviour such as power and stereotyping garnered wide press coverage. For example, in a recent *Science* paper (which the investigation has not identified as fraudulent), Stapel reported that untidy environments encouraged discrimination (*Science* **332**, 251–253; 2011).

Related stories
- Seven days: 9–15 September 2011
  14 September 2011
- Chaos promotes stereotyping
  07 April 2011

Naturejobs

**Tenure-Track Faculty Positions (Assistant / Associate / Full Professor)Yale University, Department of Genetics**
Yale University School of Medicine

**Assistant Professor**
Harvard Medical School

More science jobs
Post a job for free

Resources
- PDF Format
- Send to a Friend
- Reprints & Permissions
- RSS Feeds

external links
- Tilburg University
- Interim investigation report

FOSTER

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

# 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 **Phillip Inman**
The Guardian, Thursday 18 April 2013 21.10 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

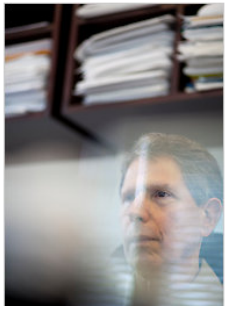
# More scientific breakthroughs

## Sharing of Data Leads to Progress on Alzheimer's

By GINA KOLATA
Published: August 12, 2010

In 2003, 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.

Enlarge This Image



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 600 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

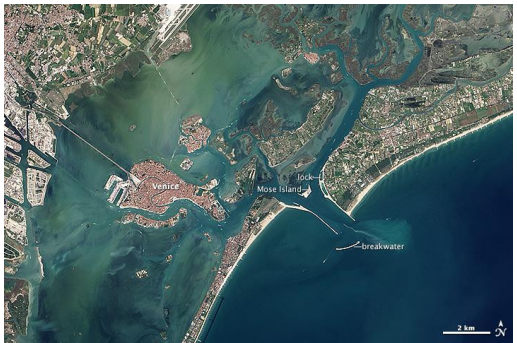
FOSTER

# Increased use and economic benefit

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

## Up to 2008

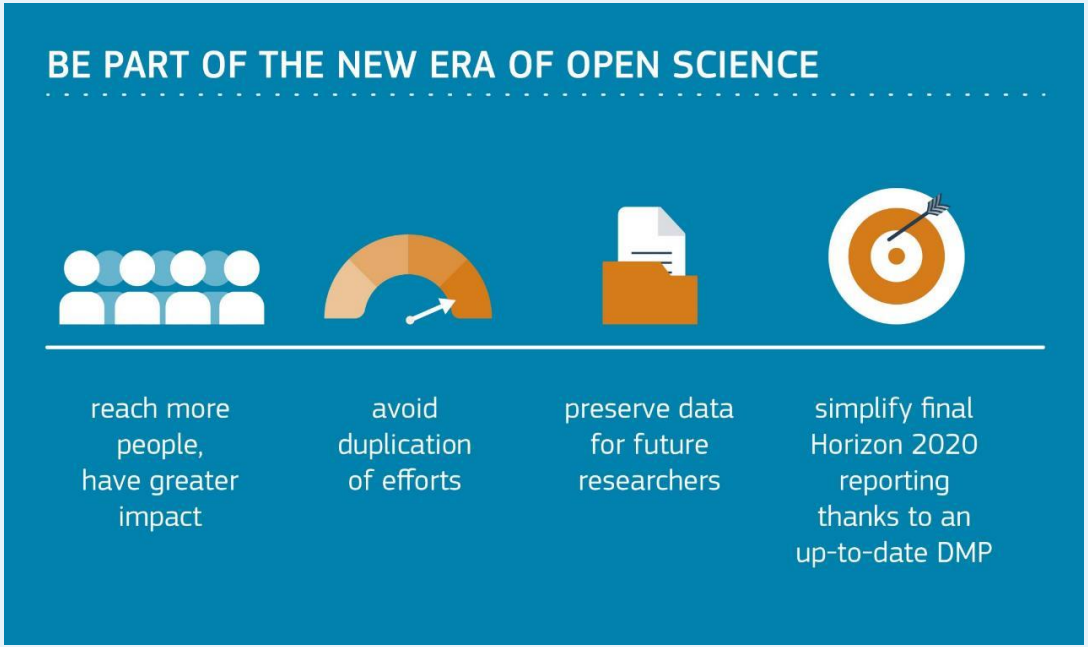
- Sold through the US Geological Survey for US$600 per scene
- Sales of 19,000 scenes per year
- Annual revenue of $11.4 million




## 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

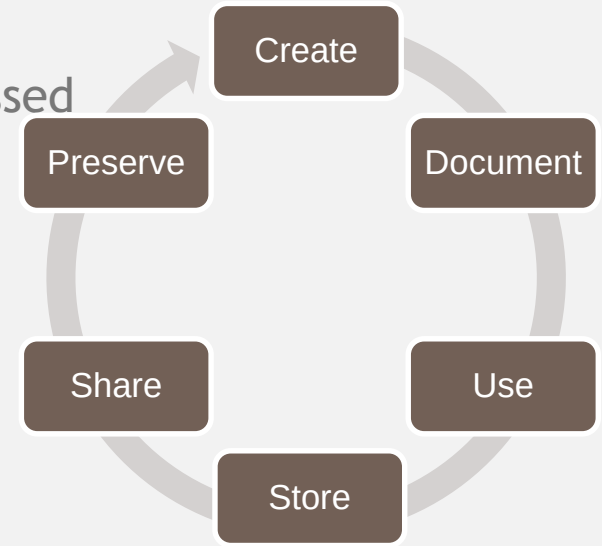

BE PART OF THE NEW ERA OF OPEN SCIENCE
reach more people, have greater impact
avoid duplication of efforts
preserve data for future researchers
simplify final Horizon 2020 reporting thanks to an up-to-date DMP

FOSTER

# Data Management Plans

# 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



# Data Management Planning

- Data Types, Formats, Standards and Capture Methods
- Ethics and Intellectual Property
- Access, Data Sharing and Reuse
- Short-Term Storage and Data Management
- Deposit and Long-Term Preservation

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


https://dmponline.dcc.ac.uk

Managing and Sharing Research Da

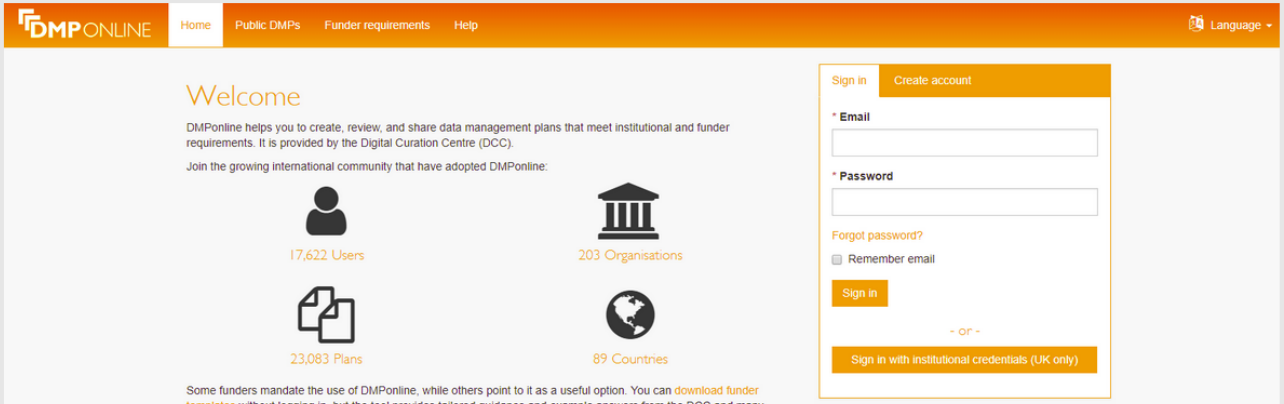
https://www.fosteropenscience.eu/learning/managing-and 120% Search

FOSTER

# 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

# 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

ICPSR framework for a DMP www.icpsr.umich.edu/icpsrweb/content/datamanagement/dmp/framework.html

So, I am managing my data. Does that mean I have to share it as well? And how do I do that?

FOSTER

# Sharing your data

## Explain which data will be shared and how (e.g. via repository, under what licence)

# 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

# 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.

# Research data lifecycle

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

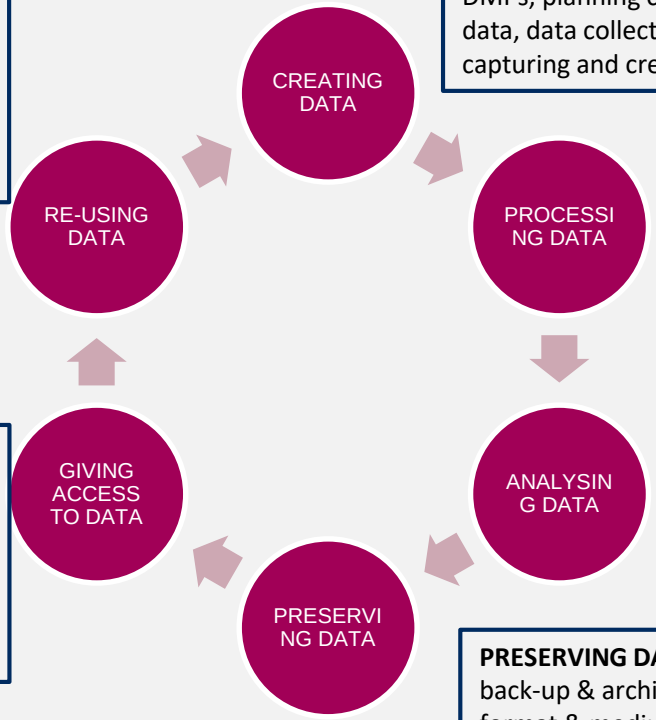
**CREATING DATA**: designing research, DMPs, planning consent, locate existing data, data collection and management, capturing and creating metadata

**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

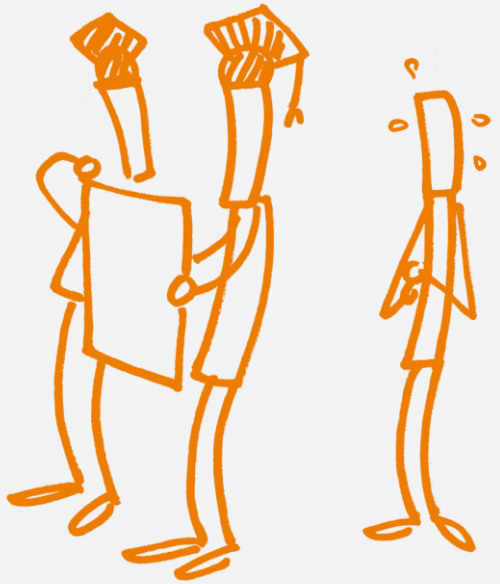
**ACCESS TO DATA**: distributing data, sharing data, controlling access, establishing copyright, promoting data




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

# 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

# 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)

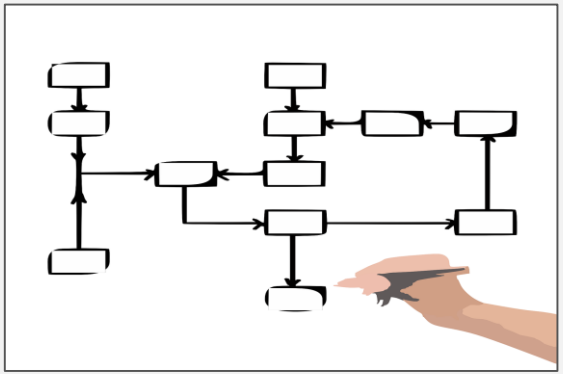
## Open Data

To support validation and facilitate reuse

## Open Code

Software created to analyse and/or visualise the data

## Open Workflows

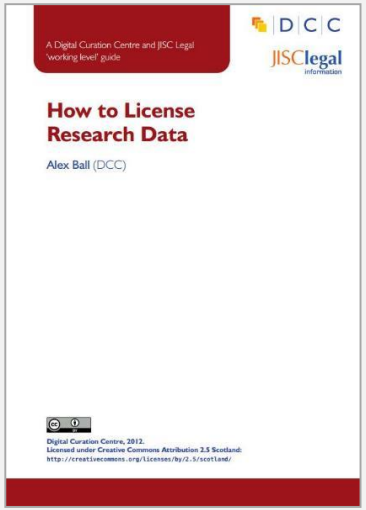


What steps were taken and in what order?

# Consider who else has a say about sharing data

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

# 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?

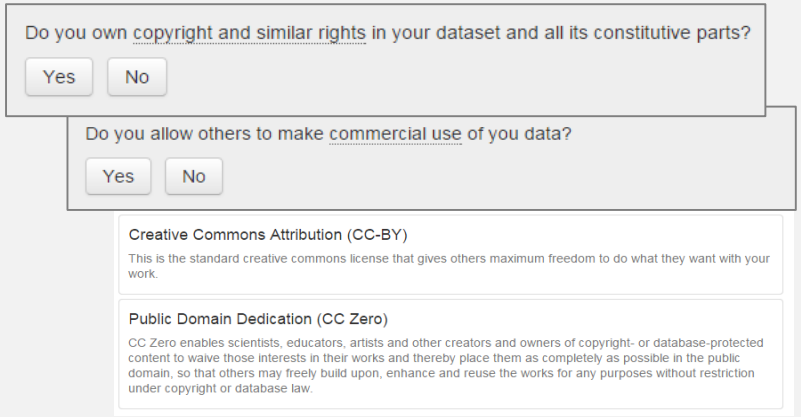
ND
No Derivatives
Severely restricts use

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

**These clauses are not open licenses**

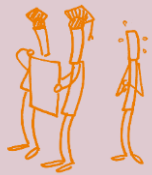
# EUDAT licensing tool

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




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

# www.fosteropenscience.eu/toolkit

| 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** |
|  |  |  |  |  |

FOSTER

Data Protection and Ethics | Data P

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

Search

FOSTER

# Data Protection and Ethics

This course helps you to get to grips with data protection and the ethics around responsible data sharing.

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

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

Expert Tour Guide on Data Mar

https://www.cessda.eu/Research-Infrastructure/Training/Expert-Tour-Guide-

Search

cessda eric

About Consortium Projects Research Infrastructure Contact

Home / Research Infrastructure / Training / Expert Tour Guide on Data Management

# Expert Tour Guide on Data Management

## About this expert tour guide

This tour guide by CESSDA ERIC (the Consortium of European Social Science Data Archives European Infrastructure Consortium) aims to put social scientists like yourself at the heart of making their research data findable, understandable, sustainably accessible and reusable.

You will be guided by European experts who are - on a daily basis - busy ensuring long-term access to valuable social science datasets, available for discovery and reuse at one of the 17 CESSDA social science data archives. With this guide and the training events being held across Europe, we want to accompany and inspire you in your journey through the research data life cycle.

https://www.cessda.eu/Research-Infrastructure/Training/Expert-Tour-Guide-on-Data-Management

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 741839


Thank you! Questions?

gwen.franck@eifl.net (g_fra)

Facebook: @fosteropenscience

Twitter: @fosterscience