

# Research data management (RDM) guidance for i-CONN early stage researchers

15/09/2020

Instructor: Nicholas Syrotiuk



<https://doi.org/d8wz>



# Session plan

F.A.I.R. data  
principles

15:00 [1]

Working  
reproducibly

10:00 [2]

Preserving  
research  
data

10:00 [3]

Summary

10:00 [4]

# What is research data?

One definition: Anything which can be used to validate or replicate a research conclusion, or enrich understanding of the research process.

# Why bother with data management?

An aerial photograph of a large, multi-story research facility. A central building is engulfed in flames, with thick black smoke billowing upwards and spreading across the sky. The surrounding area includes other large buildings, parking lots, and some greenery. The overall scene depicts a major disaster at a research center.

2005/10/30:

Fire destroys Southampton research centre

# CASH REWARD

for returning my lost backpack



*205 Adventure.com*

- Black [AK] Burton Rucksack
- Lost on Friday 15. July at 8 pm in the Panton Arms pub 43, Panton St. Cambridge
- Containing a laptop (white MacBook), a black external hard drive and scientific research documents

The external hard drive is VERY important to me as it contains 5 years of research data which are crucial for my PhD thesis!!!

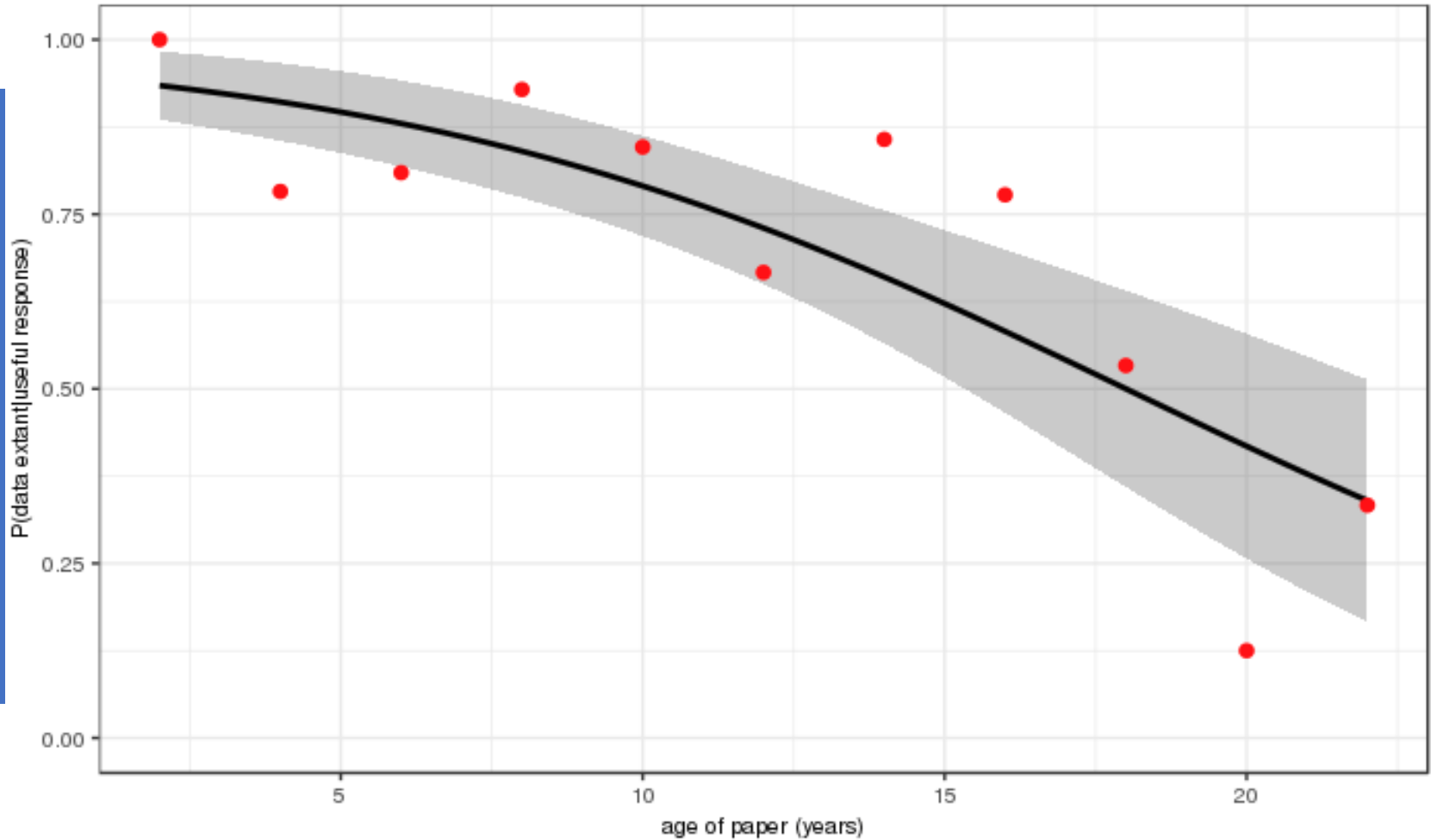
If you found it, I would be extremely grateful if you could return it to the Panton Arms or contact me on: [REDACTED]

Thank you!!

F I V E  
years of  
research  
data lost

"Availability of research data declines rapidly with article age"--Vines et al.

Data extant (percent)



Age of paper (years)

Part one:

# **WHAT ARE THE F.A.I.R. DATA PRINCIPLES?**



## Brief history

**2014: Term coined**

**2016: Fifteen principles  
published in Nature:**

DOI: <http://doi.org/10.1038/sdata.2016.18>

**2020: Guidance on web  
site:** <https://www.go-fair.org/fair-principles/>



F  
indable



A  
ccessible



I  
nteroperable



R  
eusable




Creator(s)

Title

Publisher

Publication year

Subject keywords



BIBLIOGRAPHIC  
METADATA

# Assign a persistent identifier

DOI: Digital object identifier

ARK: Archival resource key

URN: Uniform resource name

PURL: Persistent uniform resource locator

# Minting process:



1. Deposit data

2. Add bibliographic metadata

3. Mint/publish DOI

[4. Cite the data]

<https://search.datacite.org/>



Search for work

Search

F  
indable



A  
ccessible



I  
nteroperable




R  
eusable



As open as  
possible,  
as closed as  
necessary





Funders expect research data to be published openly in order to facilitate reproducible and transparent research



# Exceptions to sharing

Ethical reasons

Public safety reasons

Commercial reasons



# Data access statement: Is your research data published openly?

1. Yes. Here's the DOI.

2. Yes, with restrictions as described in the Non-disclosure Agreement.

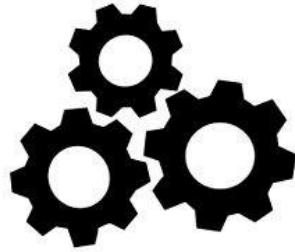
3. No, the data is too sensitive.

4. No new data was generated.

F<sub>indable</sub> A<sub>ccessible</sub>



I<sub>nteroperable</sub>



R<sub>eusable</sub>

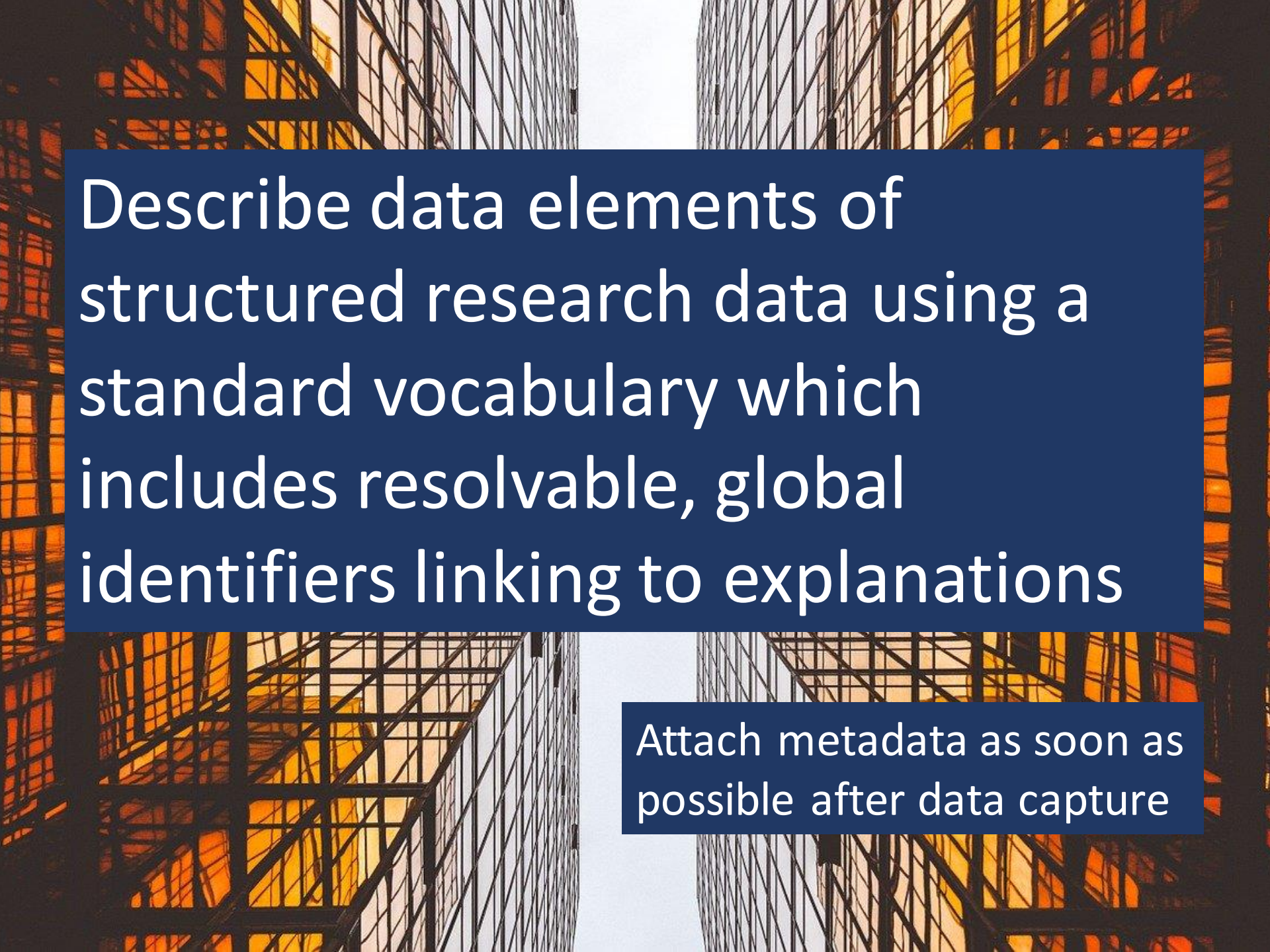


Store research data in an open, structured file format which is machine readable





Avoid  
proprietary  
file formats



Describe data elements of structured research data using a standard vocabulary which includes resolvable, global identifiers linking to explanations

Attach metadata as soon as possible after data capture



# Example of discipline-specific metadata

## Standard

### **Ecological Metadata**

**Language** (EML) for earth, environmental and ecological sciences (in XML format)

## Tool

### **Morpho**

<https://knb.ecoinformatics.org/tools/morpho>

## Use case

**Knowledge Network for Biocomplexity (KNB) data repository**



F  
indable

A  
ccessible

I  
nteroperable

R  
eusable





# Documentation

Good data comes with  
good documentation

The background of the image shows the spines of three antique books bound in dark brown leather with intricate gold-tooled floral and scrollwork patterns. The books are arranged vertically, and the text is overlaid on a white rectangular box in the upper portion of the image.

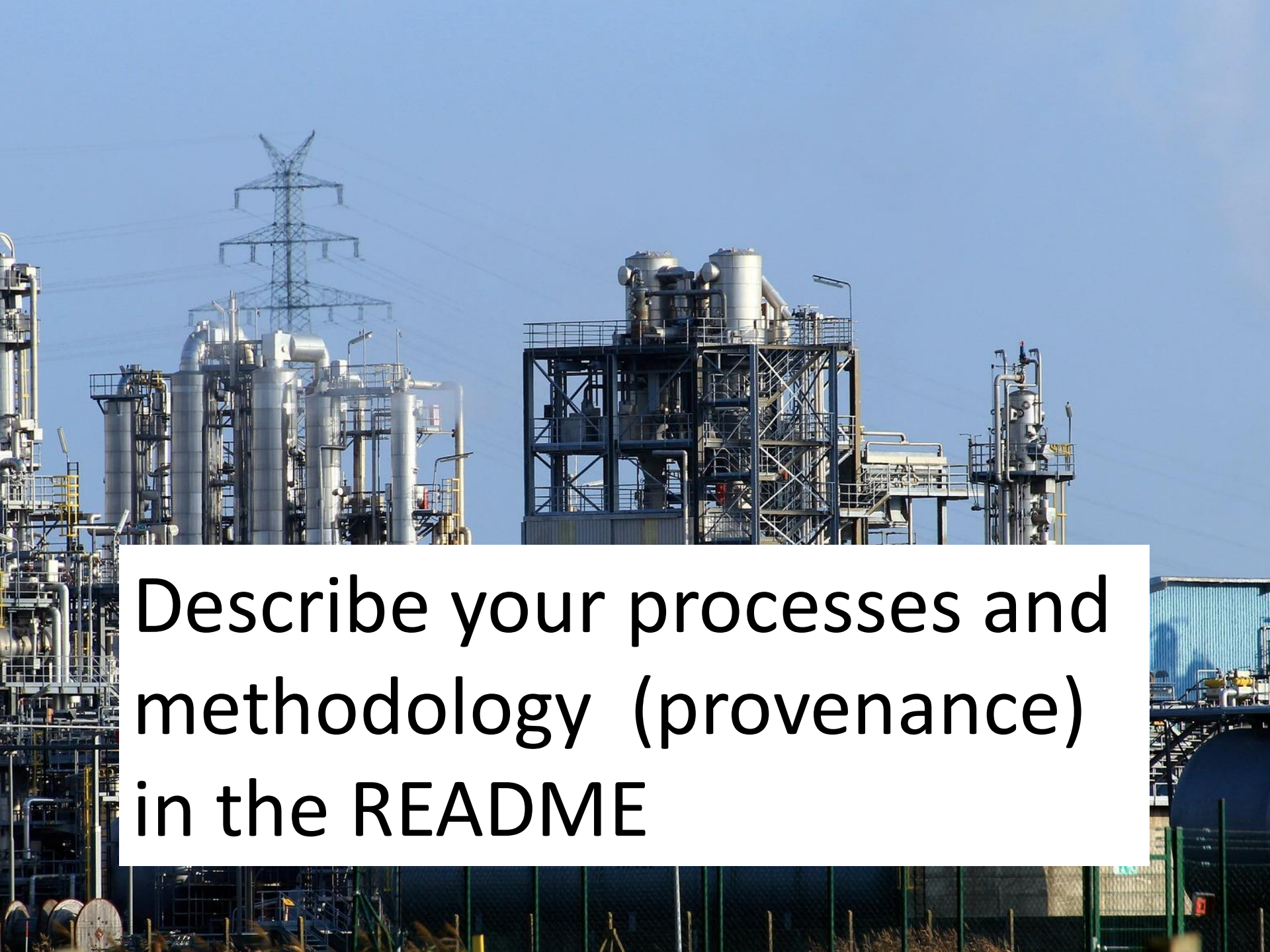
# README Guidelines:

<https://data.research.cornell.edu/content/readme>

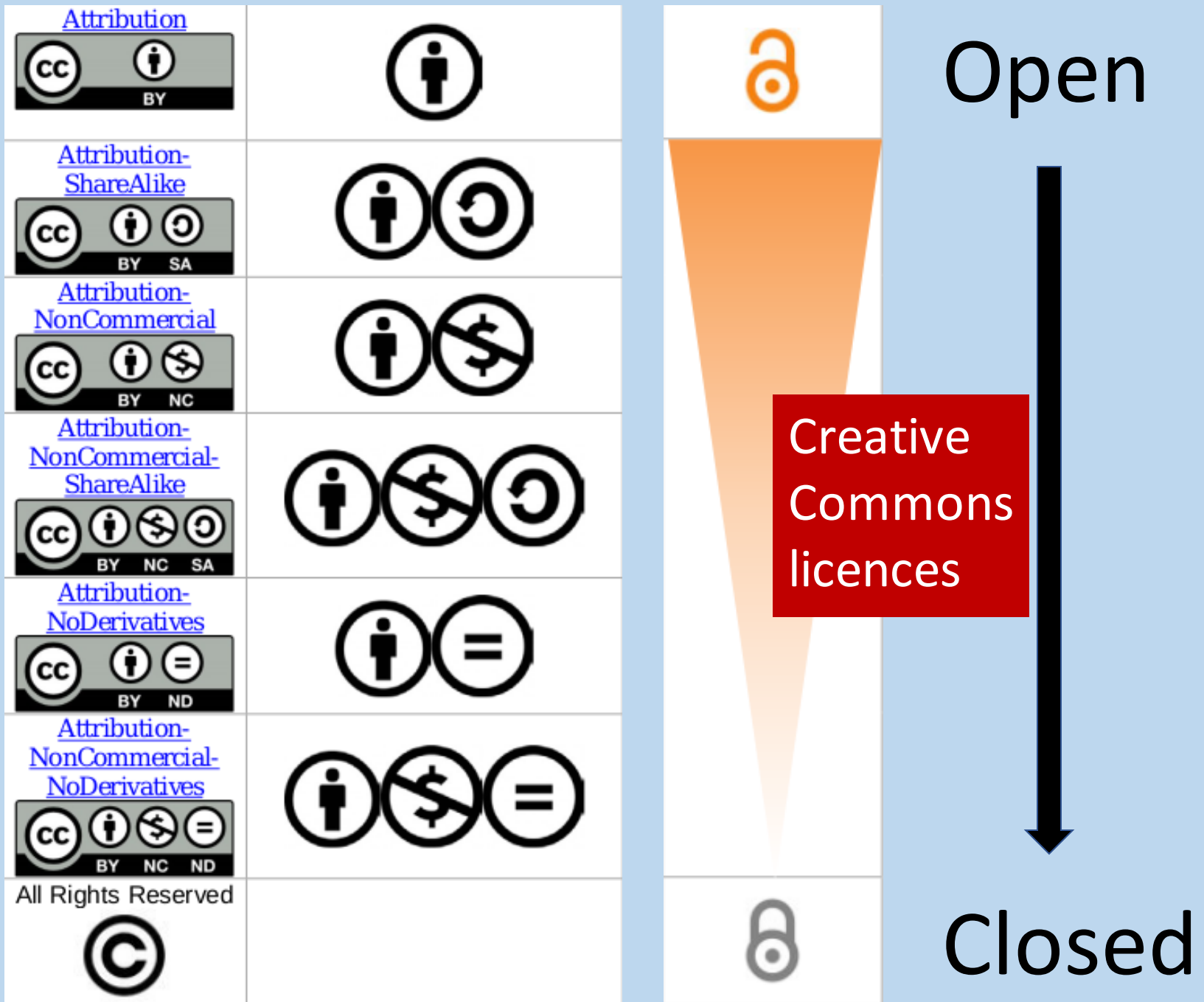
INCLUDE

README

in plain text  
FORMAT

A photograph of an industrial facility, likely a refinery or chemical plant, featuring several tall distillation columns and complex piping. In the background, a large electrical transmission tower is visible against a clear blue sky. The foreground shows a green fence and some vegetation.

**Describe your processes and methodology (provenance) in the README**

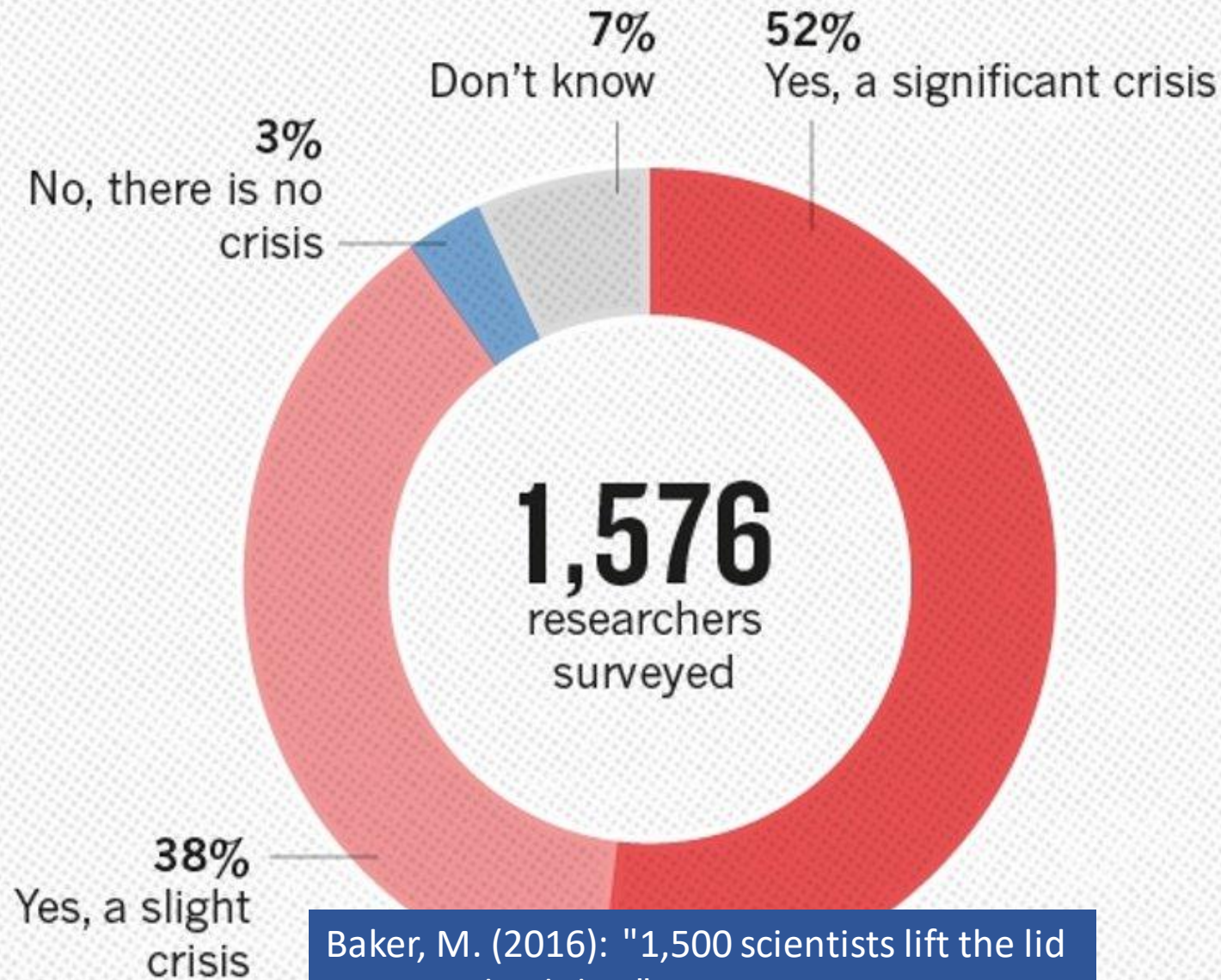


Part two:

**WORKING**

**REPRODUCIBLY**

# IS THERE A REPRODUCIBILITY CRISIS?



Baker, M. (2016): "1,500 scientists lift the lid on reproducibility." *Nature*, 533:7604. DOI: <http://doi.org/10.1038/533452a>



# Levels of reproducibility

According to Florian Markowetz (2015)



# Avoid beginners' mistakes

Keep files organised

Name files in a meaningful way

Avoid scattering files

# START

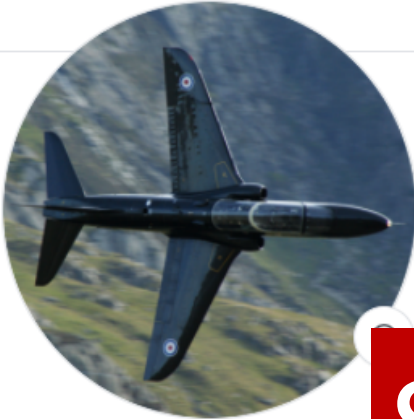
Lowest level of reproducibility

# Computational reproducibility

Scripting tools:  
Python, R, Perl, ...

Notebook tools:  
IPython, Jupyter, ...

Next level of reproducibility



Nicholas Syrotiuk  
sefnyn

Edit profile

4 followers · 6 following · 24 stars

Highlights  
Arctic Code Vault Contributor

Organizations  
OECD

Overview Repositories 22 Projects Packages

Pinned

Repository card for 'scholix' (Python) with 6 stars and 2 forks. Description: Scripts to find links between research papers and corresponding research data.

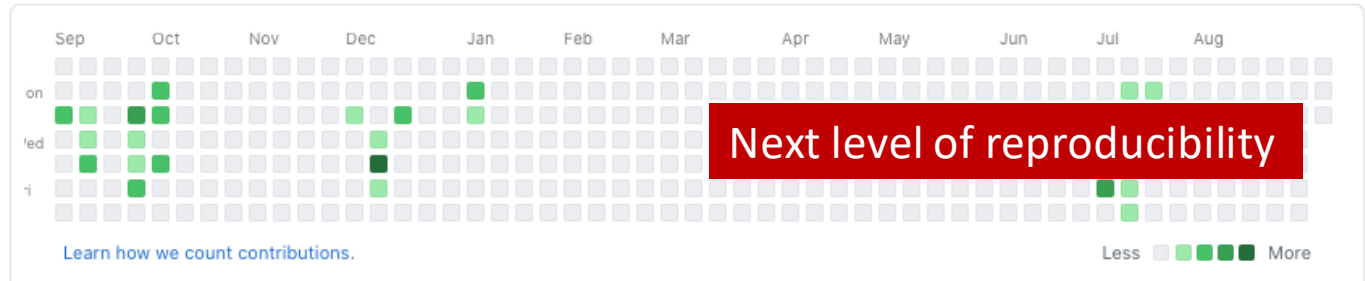
Repository card for 'carousel' (JavaScript) with description: RDM slides published as a simple image carousel.

Software version control systems  
e.g., Git, Mercurial, Subversion

Repository card for JavaScript

Repository card for Ruby

144 contributions in the last year



Next level of reproducibility

An aerial photograph of a busy port. The top half shows a large yard filled with stacks of colorful shipping containers (blue, red, green, white) and several red gantry cranes. The bottom half shows a large container ship docked at a pier, with its deck also covered in stacks of containers. Two red gantry cranes are positioned over the ship's deck. The water is dark blue.

# Containerisation

e.g., Docker

Highest level of reproducibility

Part three:

**PRESERVING**

**RESEARCH**

**DATA** FOR THE LONG TERM

# Preparation and organisation

A top-down view of various professional video production equipment neatly arranged on a light-colored wooden surface. The items include two Canon EOS R5 cameras, a GoPro, a DJI drone with its controller, a gimbal, several lenses, two RED memory cards, a REVOLT battery, a pair of headphones, a tablet, and a smartphone. The equipment is organized in a grid-like pattern, showcasing a variety of high-end gear used in cinematography and videography.

Provide all necessary files

Ensure accessibility of files

Descriptive file names

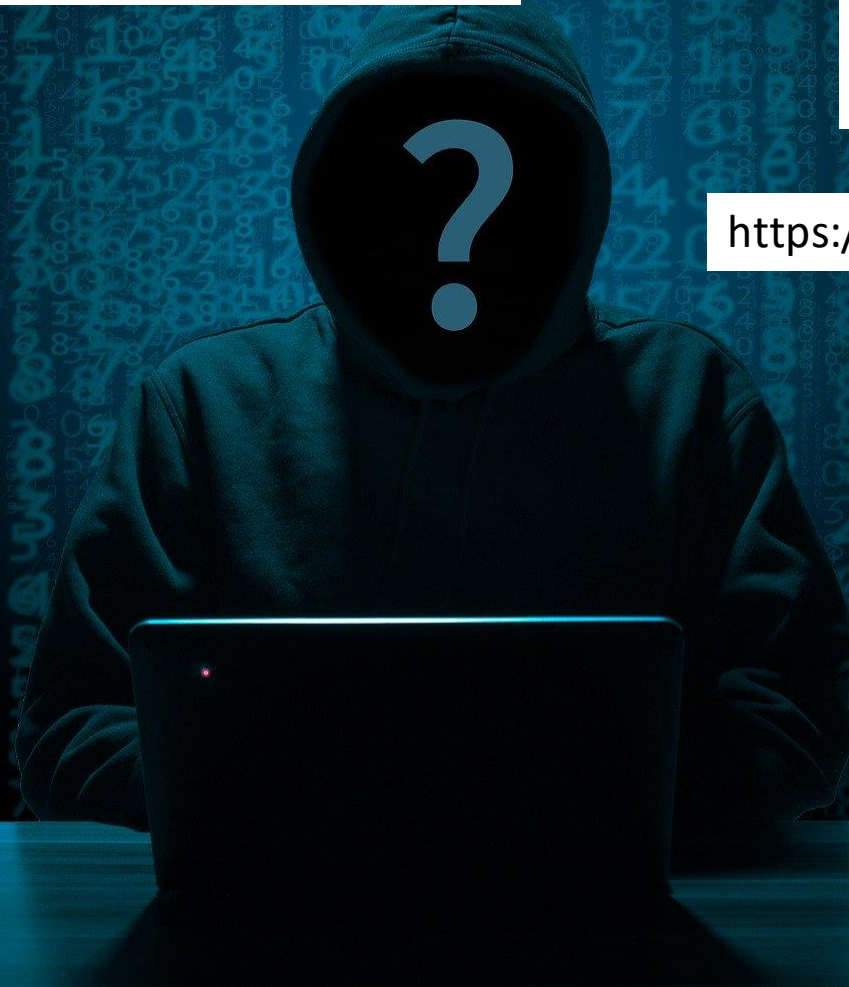
Logical file structure

Include a README

# Anonymisation



<https://amnesia.openaire.eu/>



A hallway with five doors set against a wall with a repeating floral pattern. The floor is made of light-colored wood planks. The doors are white with black frames and handles, except for the second door from the left, which is bright yellow. A red rectangular box is overlaid on the lower half of the image, containing white text.

Choosing a research  
data repository



Three types of repository

1

Subject-specific

2

Multi-disciplinary

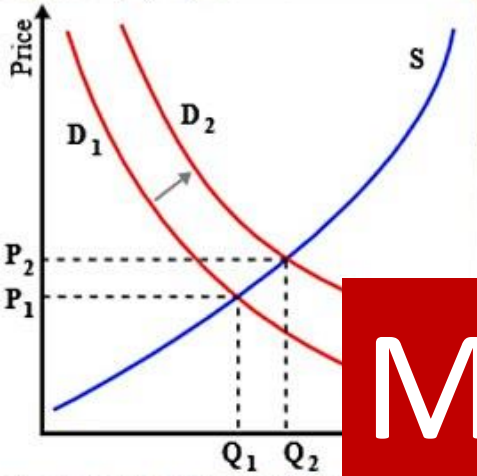
3

Institutional

Find a subject-specific  
repository



re3data.org  
REGISTRY OF RESEARCH DATA REPOSITORIES



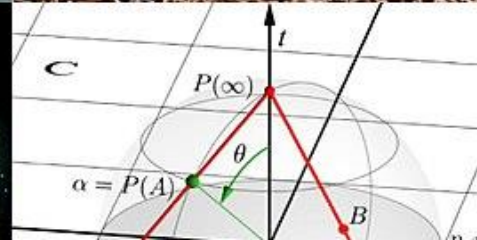
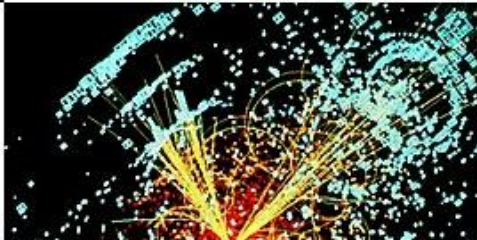
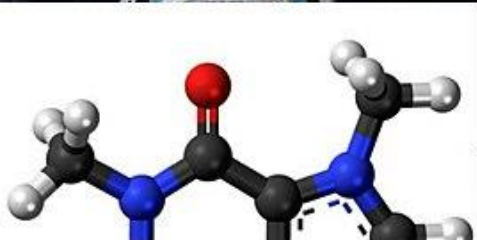
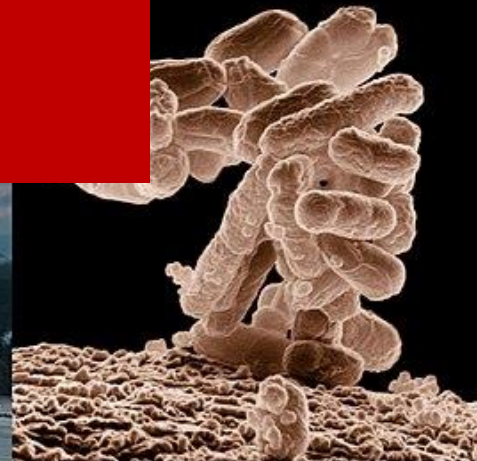
# Multi-disciplinary repository



```

if ast[1].strip():
    print '=' % s; ' % ast[1]
else:
    print ''
else:
    print ''
children = []
for n, child in enumerate(ast[
    children.append(dotwrite(ch
print ' %s -> {' % nodename,
for name in children:
    print ' %s' % name,

```





Mendeley

zenodo

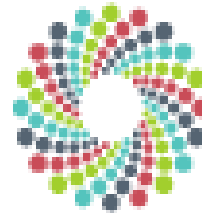


**DRYAD**



OSF

Open Science Framework



figshare

Examples of multi-disciplinary repositories

Enter search terms

All ▾

[About](#) [Help](#) [Contact](#) [Login](#)

# Institutional repo: Durham research data repository



## Recently Uploaded

Depositor	File Details
N.G. Chancellor 	<a href="#"><u>Finding spin-glass ground...</u></a> <a href="#">qwspinglass_data.tar.gz</a> <a href="#">quantum computing</a> , <a href="#">quantum walk</a> , <a href="#">quantum optimisation</a> , <a href="#">quantum algorithms</a> , <a href="#">spin glass</a>

## Tweets by @DurhamRdm

 Durham RDM Retweeted

 **DurhamResearchOnline**  
@DROdurham  
Accepted manuscript now available on DRO: Kilby, Karen E. (2018) 'Julian of Norwich, Hans Urs von Balthasar, and the status of suffering in Christian theology.', *New blackfriars.*, 99 (1081). pp. 298-311. [dro.dur.ac.uk/22207/](http://dro.dur.ac.uk/22207/)

  5h



Metrics

48,176 Downloads

# Another institutional repository

## Welcome to DataverseNL

Store, share and publish research data online. Use the slider below to access the dataverses of the DataverseNL partners. If you want to try out the DataverseNL features, please visit our [demo-site](#).



Utrecht University

Utrecht University



Maastricht University

Maastricht University



Understanding Society

Tilburg University



rijksuniversiteit groningen

University of Groningen



Search this dataverse...

Find

Advanced Search

**Dataverses (370)**

**Datasets (1,442)**

**Files (9,522)**

1 to 10 of 1,812 Results

Sort

### The Attentional Blink is Related to the Microsaccade Rate Signature

Sep 1, 2020 - Cognitive Neuroscience



Roberts, Mark J.; Lange, Gesa; Van Der Veen, Tracey; Lowet, Eric; de Weerd, Peter, 2020, "The Attentional Blink is Related to the Microsaccade Rate Signature", <https://doi.org/10.34894/5YNXRI>, DataverseNL, V1

The reduced detectability of a target T2 following discrimination of a preceding target T1 in the attentional blink (AB) paradigm is classically interpreted as a consequence of reduced attention to T2 due to attentional allocation to T1. Here, we investigated whether AB was relat...

# Data citation

Format:

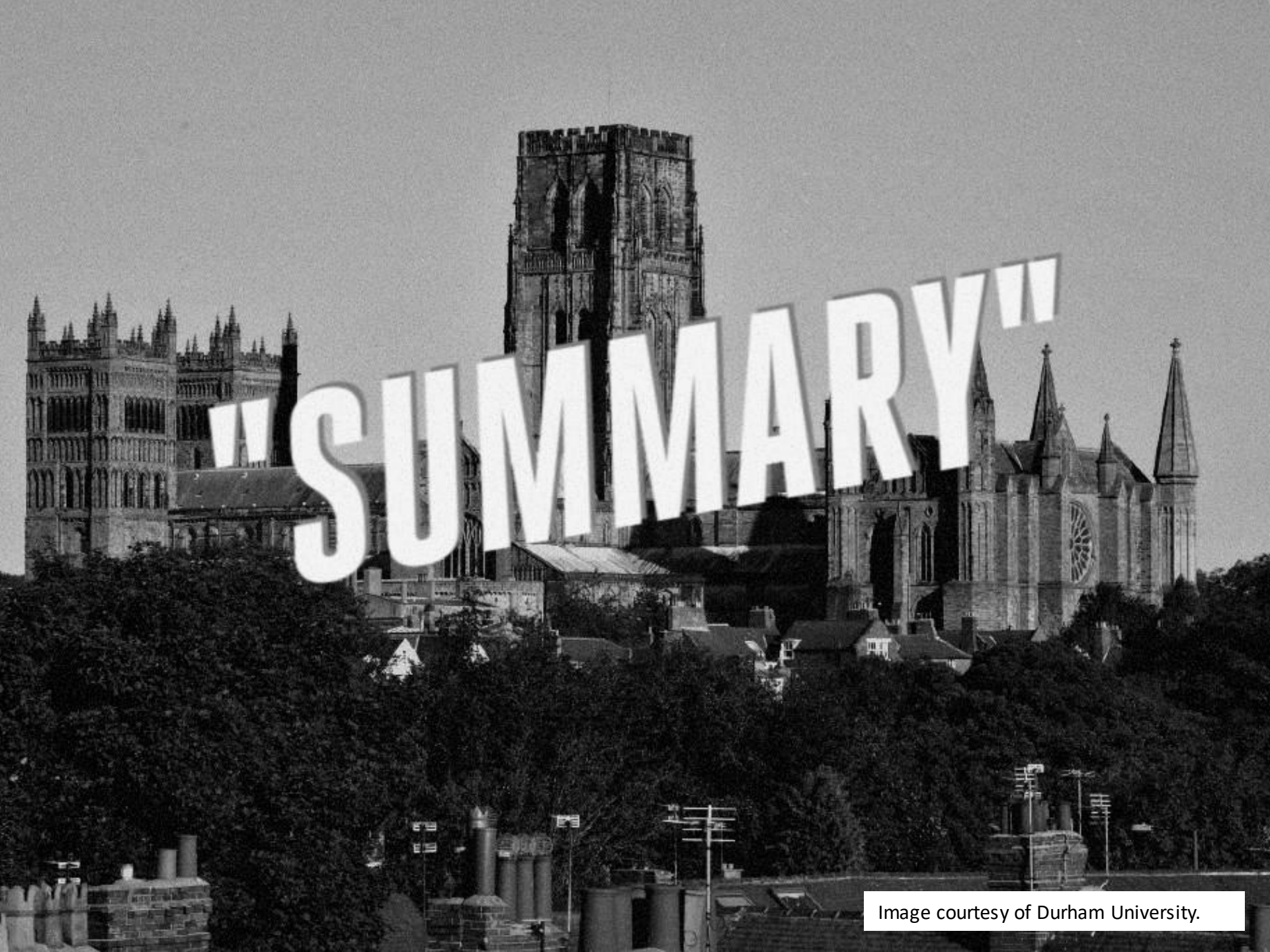
Creator (PublicationYear): Title. Version. Publisher.  
(resourceTypeGeneral). Persistent identifier

Example:

Breckon, T; Tiancheng, G (2018): Pretrained neural network models for Guo 2018 study, TensorFlow format. Durham University. (dataset).

DOI: <http://doi.org/10.15128/r23j333226h>

Source: Metadata Schema Documentation for the Publication and Citation of Research Data, version 4.3. DOI: <https://doi.org/10.14454/7xq3-zf69>

A black and white photograph of Durham Cathedral, a large Gothic-style building with a prominent square tower. The cathedral is set on a hillside, with trees and other buildings visible in the foreground. The word "SUMMARY" is overlaid in large, white, 3D-style capital letters across the center of the image.

# "SUMMARY"

Image courtesy of Durham University.



# Summary (1): Basic do's and don'ts of data management

DO	DON'T
Have a plan for managing research data	Make it up as you go along
Keep backups. Make this easy with automated syncing services like Dropbox, provided your data isn't too sensitive	Carry the only copy around on a memory card, your laptop, your phone, etc
Describe your data as you collect it. This makes it possible for others to interpret it, and for you to do the same a few years down the line	Leave this till the end. The quality of metadata decreases with time, and the best metadata is created at the moment of data capture
Save your work in open file formats, where possible, and use accepted metadata standards to enable like-with-like comparison	Invent new 'standards' where community norms already exist
Deposit your data in a data centre or repository, and link it to your publications	Be afraid to ask for help. This will exist both within institutions, and via national / European support organisations

## Summary (2): Data management rules of thumb

- Without intervention, data + time = no data
  - *See:* Vines, T. H., et al. (2014): "The availability of research data declines rapidly with article age," *Current Biology* 24(1): 94-97. DOI: <http://doi.org/10.1016/j.cub.2013.11.014>
- Following F.A.I.R. data principles and sharing research data can lead to making more progress as a research community collectively
- Working reproducibly and writing good RDM documentation ultimately saves time
- Publish data in a repository in order to preserve it for the long term
  - Not all data should be published or shared
  - Publish in one place only



# Thank you

Nicholas Syrotiuk



@DurhamRdm



<http://doi.org/d8wz>

# References:

Vines, T. H., et al. (2014): "The availability of research data declines rapidly with article age," *Current Biology* 24(1): 94-97. DOI: <https://doi.org/10.1016/j.cub.2013.11.014>

Vines, T. H., et al. (2013): "The availability of research data declines rapidly with article age." Dryad Digital Repository. (dataset). DOI: <https://doi.org/10.5061/dryad.q3g37>

UK Research and Innovation (2018): "Guidance on best practice in the management of research data." UKRI web site.

Markowetz, F. (2015): "Five selfish reasons to work reproducibly." *Genome Biol* 16, 274. DOI: <https://doi.org/10.1186/s13059-015-0850-7>

Piccolo, S.R. and Frampton, M.B. (2016): "Tools and techniques for computational reproducibility." *GigaScience*, Volume 5, Issue 1, s13742–016–0135–4. DOI: <https://doi.org/10.1186/s13742-016-0135-4>

# Credits:

F.A.I.R. data image:

Licensed by Sangya Pundir under CC-BY-SA 4.0.

[https://commons.wikimedia.org/wiki/File:FAIR\\_data\\_principles.jpg](https://commons.wikimedia.org/wiki/File:FAIR_data_principles.jpg)

CC licence image:

Licensed by Creative Commons Australia under CC BY 4.0.

<http://creativecommons.org.au/know-your-rights/>

Image of academic disciplines:

Licensed by the Collective under CC BY-SA 3.0.

<https://commons.wikimedia.org/w/index.php?curid=62858409>