NBIS
NATIONAL BIOINFORMATICS
INFRASTRUCTURE SWEDEN

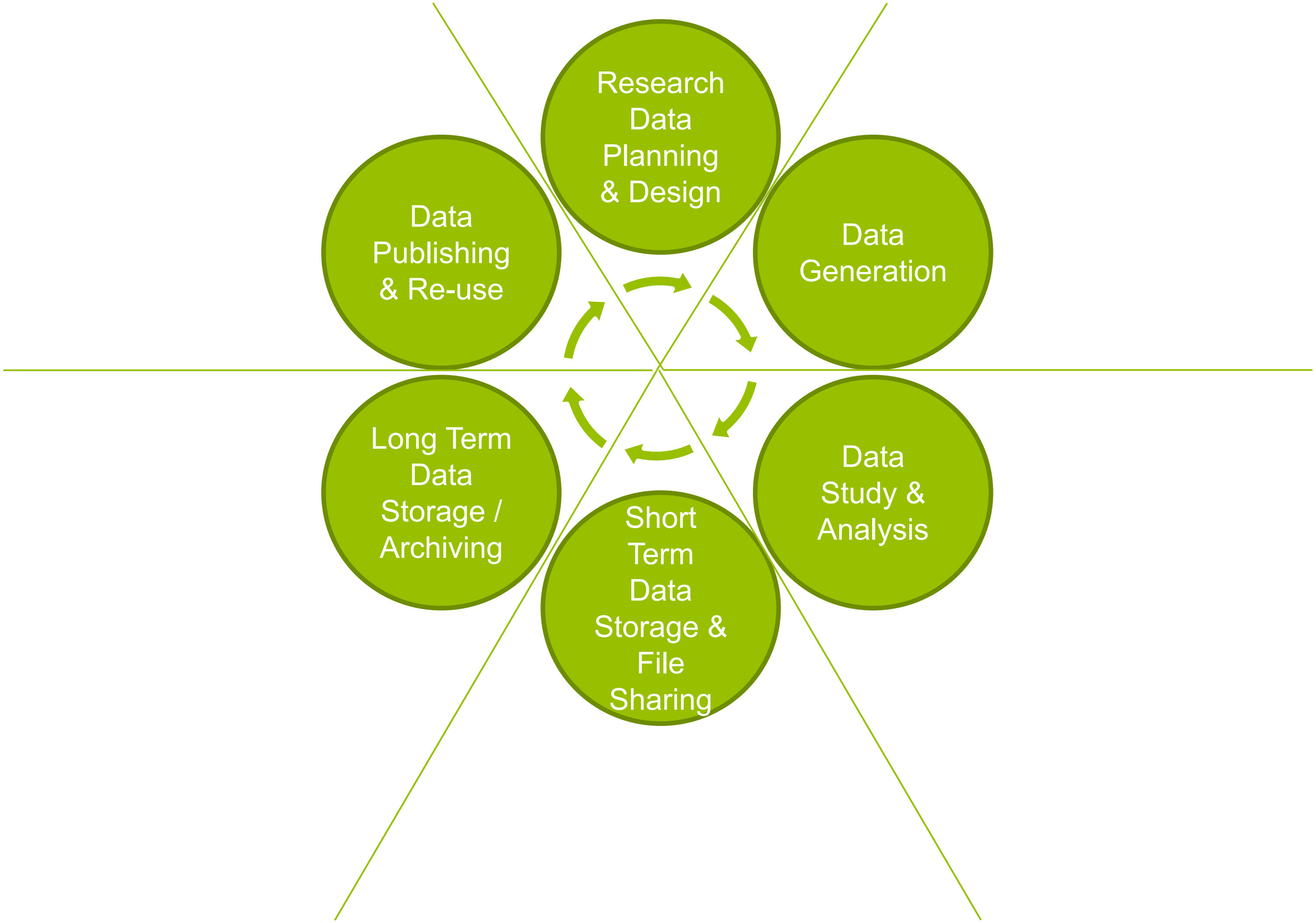
SciLifeLab

Managing your data

Niclas Jareborg, NBIS
niclas.jareborg@nbis.se

How do you know how an old result was generated?

The Research Data Life Cycle



Research Data Planning & Design

Data Generation

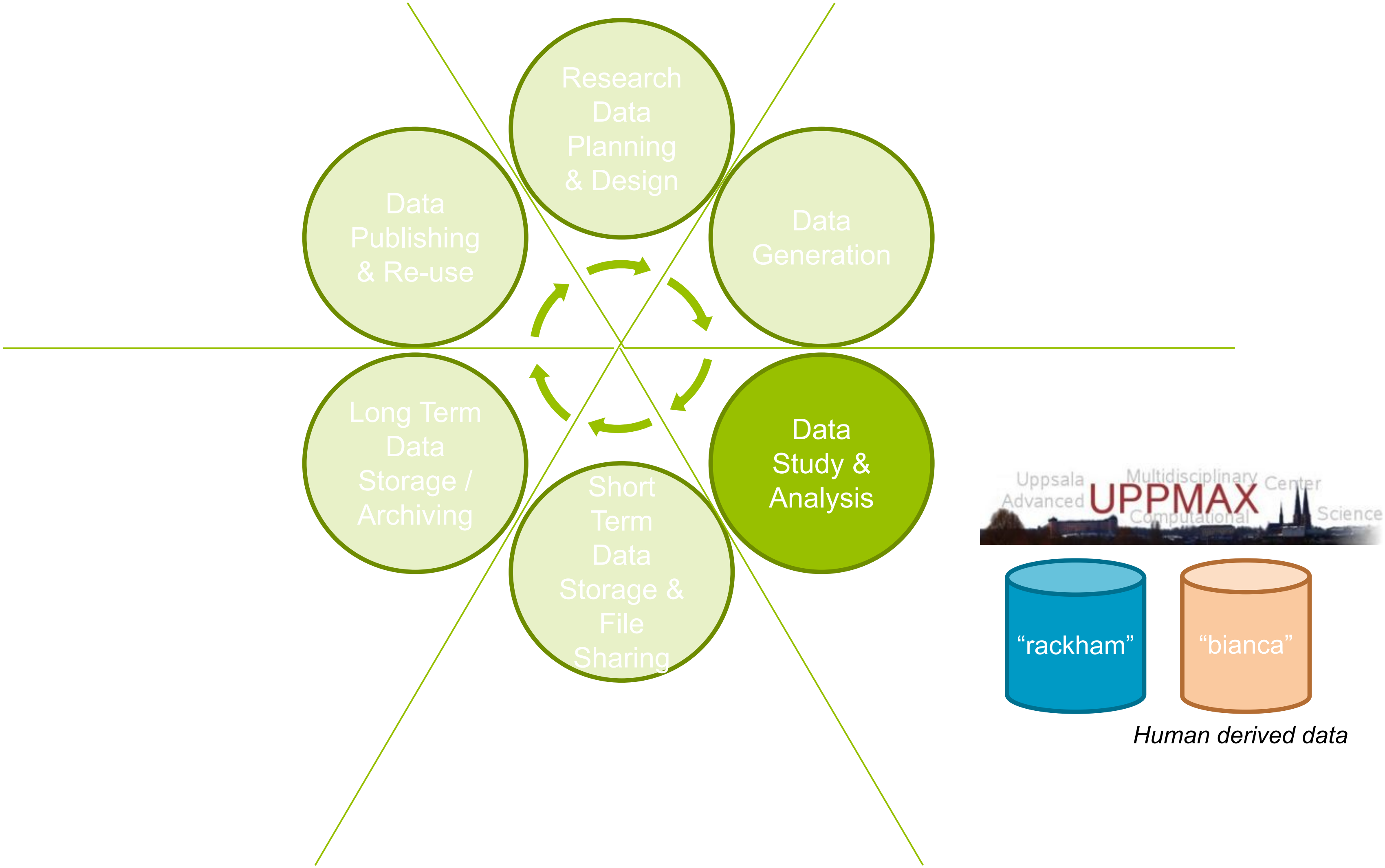
Data Study & Analysis

Short Term Data Storage & File Sharing

Long Term Data Storage / Archiving

Data Publishing & Re-use

Study & Analysis



Research Data Planning & Design

Data Generation

Data Study & Analysis

Short Term Data Storage & File Sharing

Long Term Data Storage / Archiving

Data Publishing & Re-use

Uppsala Multidisciplinary Center Advanced UPPMAX Computational Science

“rackham”

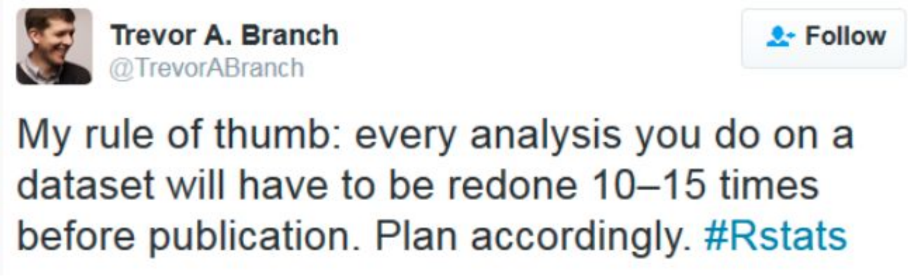
“bianca”

Human derived data

Structuring data for analysis

- Guiding principle
  - *“Someone unfamiliar with your project should be able to look at your computer files and understand in detail what you did and why.”*

- Research reality
  - *"Everything you do, you will have to do over and over again”* – Murphy’s law



Trevor A. Branch
@TrevorABranch

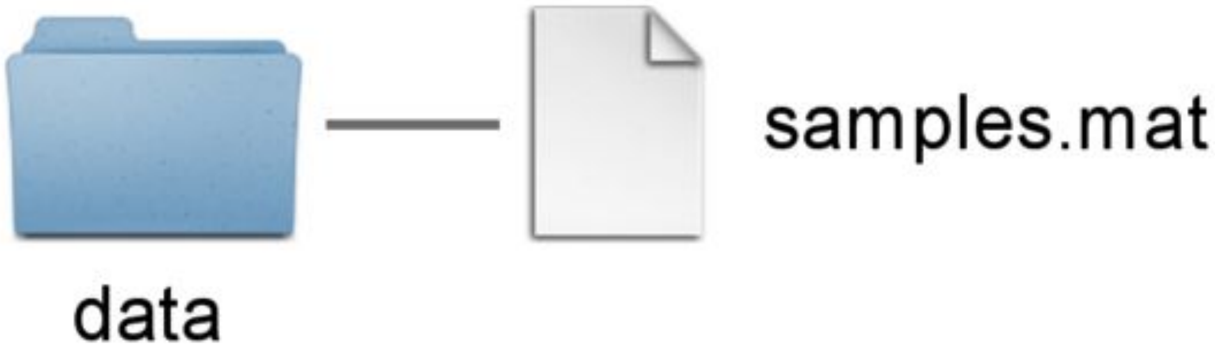
My rule of thumb: every analysis you do on a dataset will have to be redone 10–15 times before publication. Plan accordingly. #Rstats

Structuring data for analysis

Poor organizational choices lead to significantly slower research progress

“Your primary collaborator is yourself six months from now, and your past self doesn’t answer e-mails.”

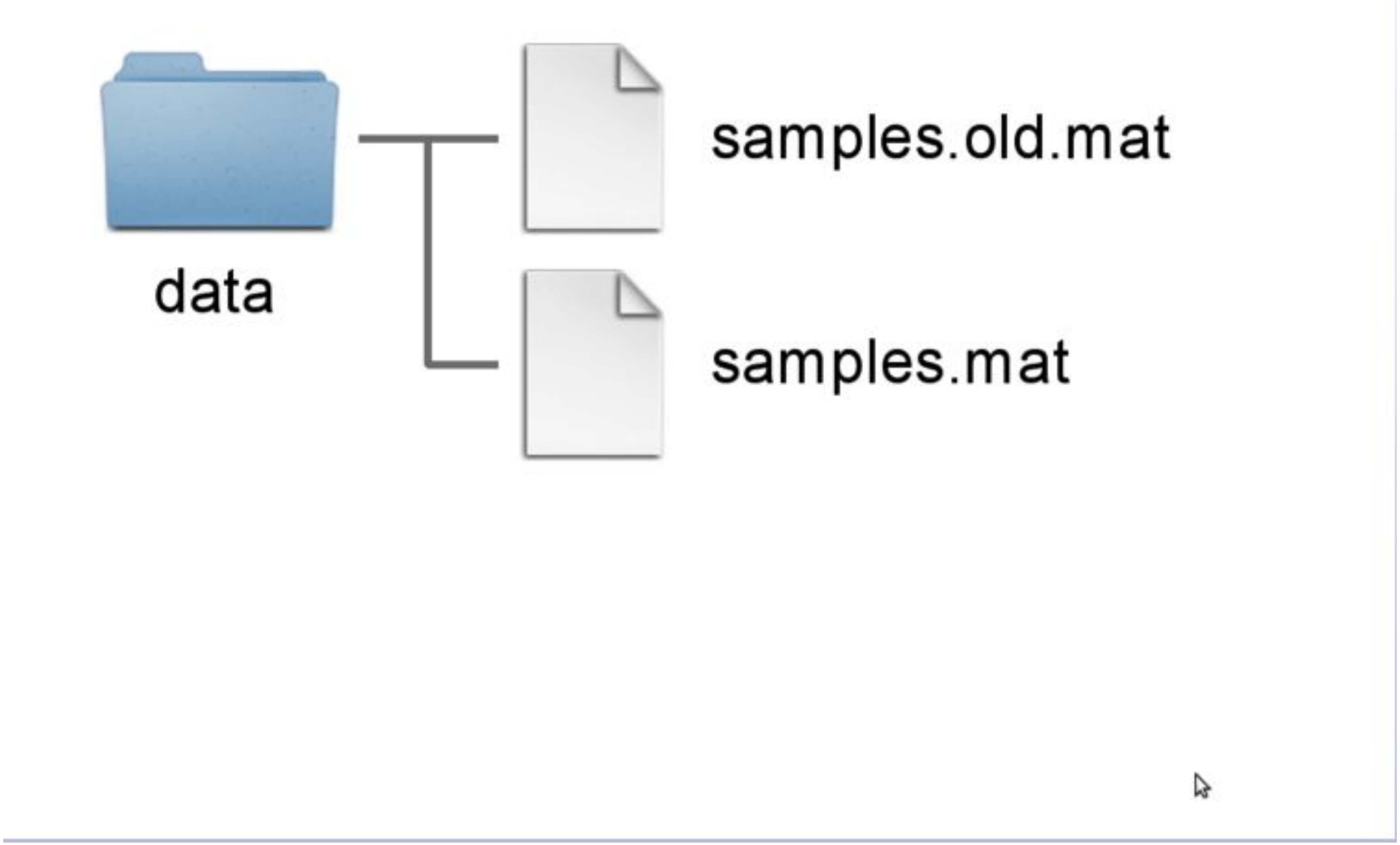
First step - Organization



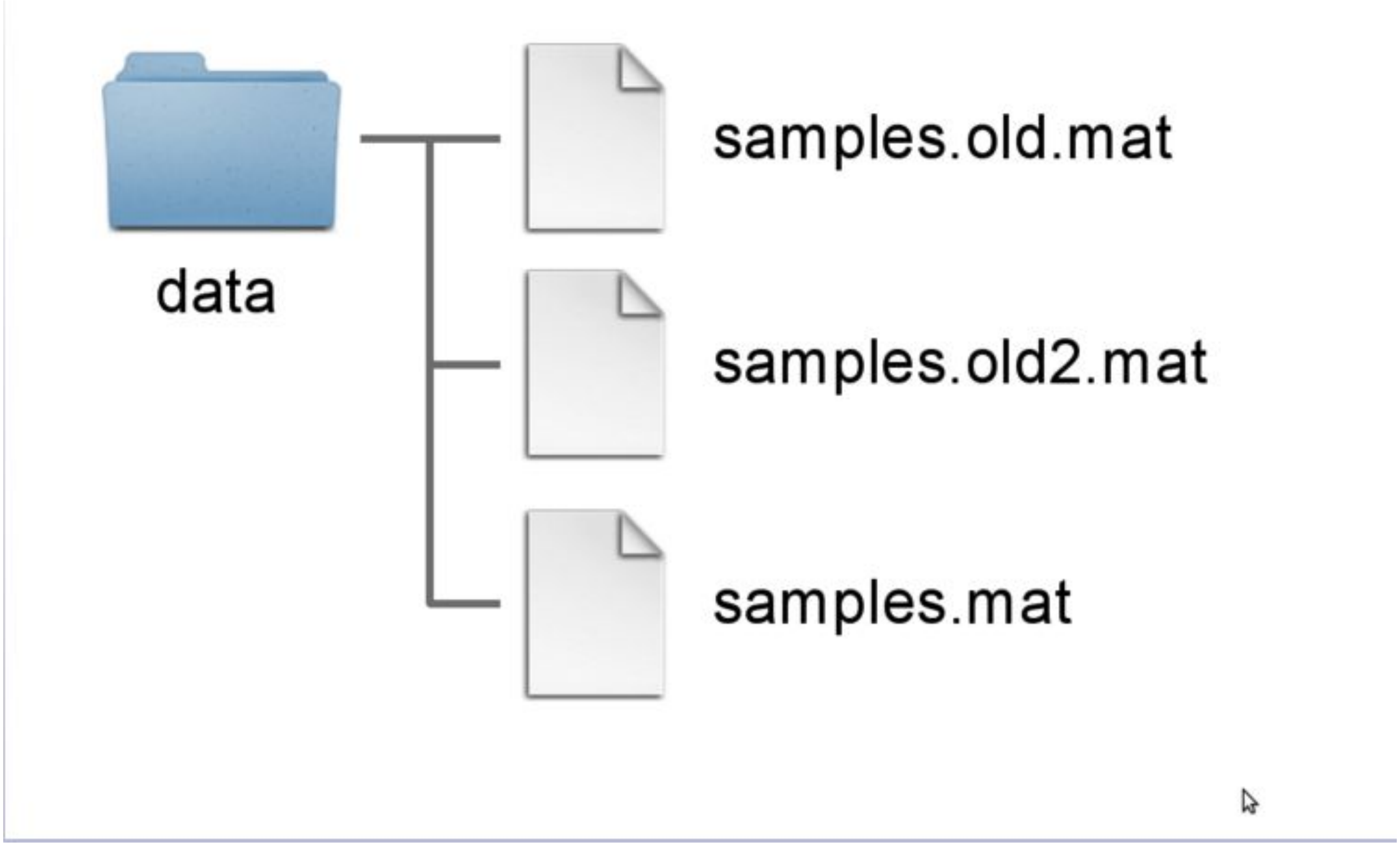
samples.mat

data

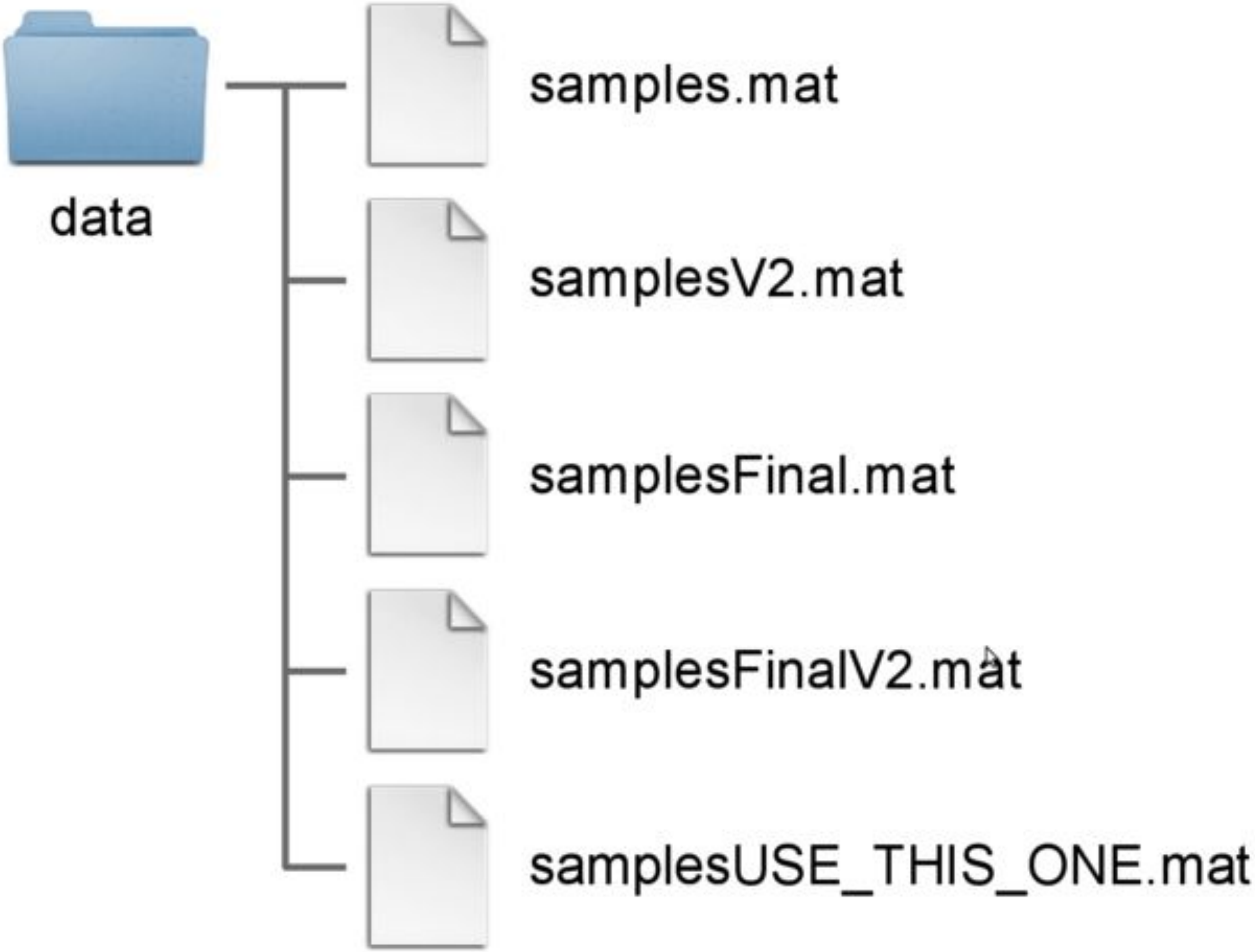
Now what?



I guess this is alright



Which one is the most recent?



data

samples.mat

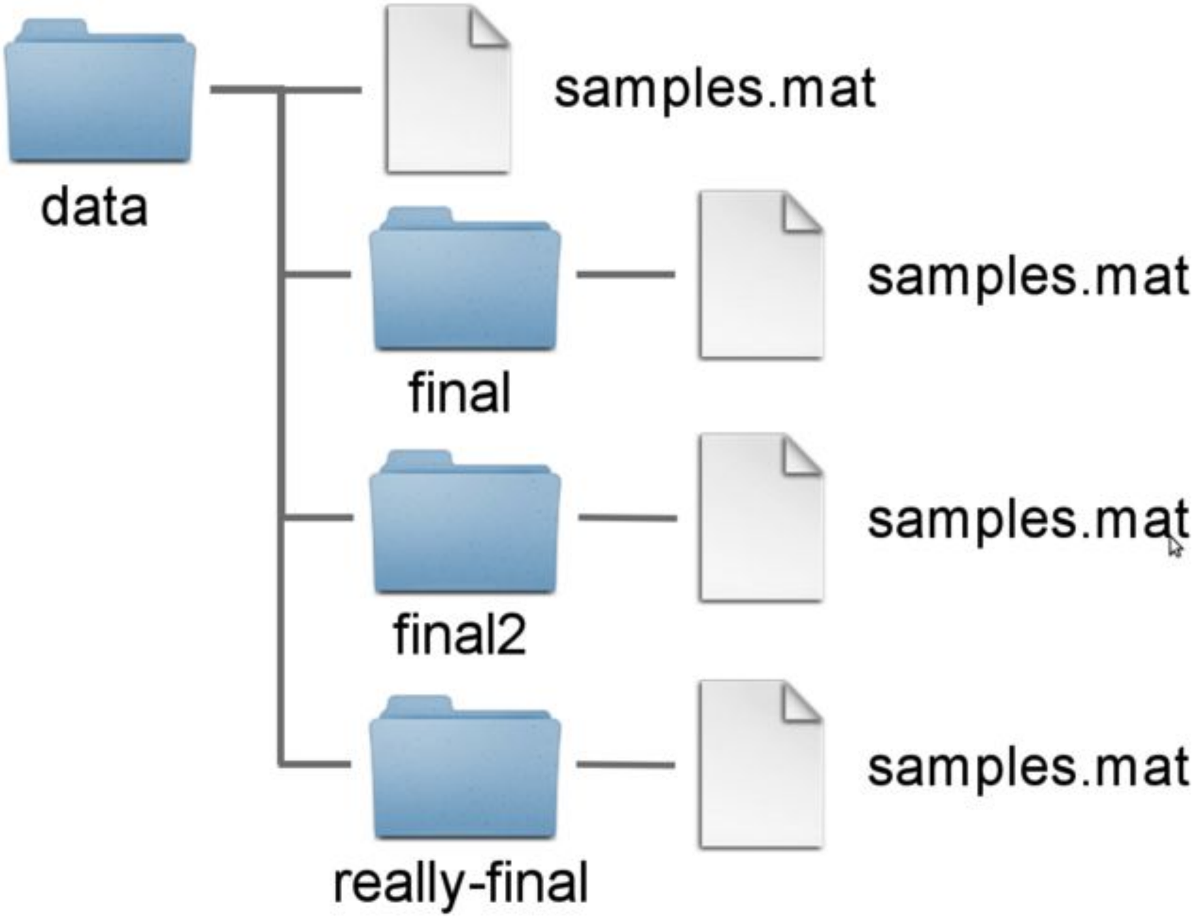
samplesV2.mat

samplesFinal.mat

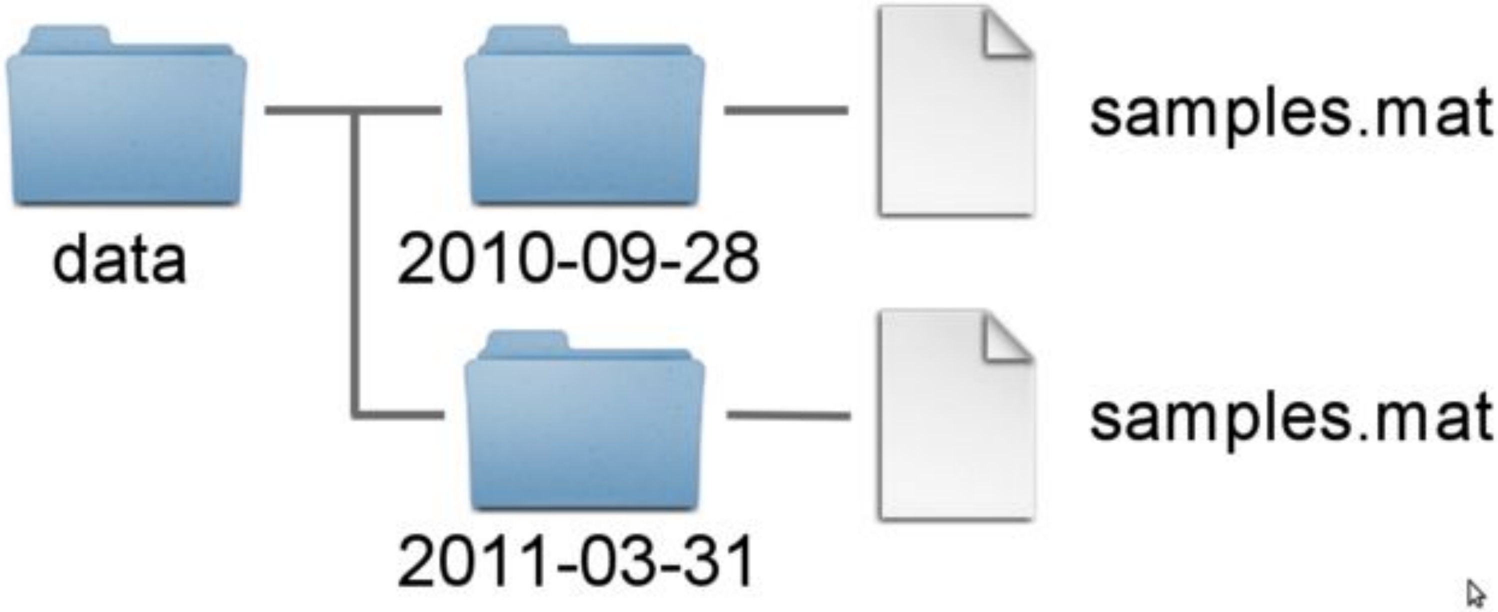
samplesFinalV2.mat

samplesUSE_THIS_ONE.mat

Another (bad) common approach



A possible solution

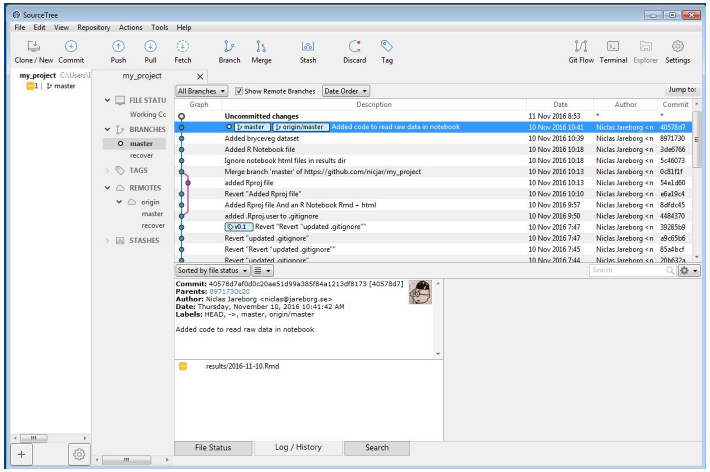


Suggested best practices

- There is a **folder for the raw data**, which do not get altered, or intermixed with data that is the result of manual or programmatic manipulation. I.e., derived data is kept separate from raw data, and **raw data are not duplicated**.
- **Code is kept separate from data**.
- Use a **version control system** (at least for code) – e.g. **git**
- There is a **scratch directory for experimentation**. Everything in the scratch directory can be deleted at any time without negative impact.
- There should be a **README in every directory**, describing the purpose of the directory and its contents.
- Use **file naming schemes** that makes it easy to find files and understand what they are (for humans and machines)
- Use **non-proprietary formats** – *.csv* rather than *.xlsx*
- Etc…

Version control

- What is it?
  - A system that keeps records of your changes
  - Allows for collaborative development
  - Allows you to know who made what changes and when
  - Allows you to revert any changes and go back to a previous state
- Several systems available
  - git, RCS, CVS, SVN, Perforce, Mercurial, Bazaar
  - **git**
    - Command line & GUIs
    - Remote repository hosting
      - GitHub, Bitbucket, etc



Graph	Description	Date	Author	Commit
	Uncommitted changes	11 Nov 2016 8:53	*	*
	master origin/master Added code to read raw data in notebook	10 Nov 2016 10:41	Niclas Jareborg <n	40578d7
	Added bryceveg dataset	10 Nov 2016 10:39	Niclas Jareborg <n	8971730
	Added R Notebook file	10 Nov 2016 10:18	Niclas Jareborg <n	3de6766
	Ignore notebook html files in results dir	10 Nov 2016 10:18	Niclas Jareborg <n	5c46073
	Merge branch 'master' of https://github.com/nicjar/my_project	10 Nov 2016 10:13	Niclas Jareborg <n	0c81f1f
	added Rproj file	10 Nov 2016 10:13	Niclas Jareborg <n	54e1d60
	Revert "Added Rproj file"	10 Nov 2016 10:10	Niclas Jareborg <n	e6a19c4
	Added Rproj file And an R Notebook Rmd + html	10 Nov 2016 9:57	Niclas Jareborg <n	8dfdc45
	added .Rproj.user to .gitignore	10 Nov 2016 9:50	Niclas Jareborg <n	4484370
	v0.1 Revert "Revert "updated .gitignore""	10 Nov 2016 7:47	Niclas Jareborg <n	39285b9
	Revert "updated .gitignore"	10 Nov 2016 7:47	Niclas Jareborg <n	a9c65b6
	Revert "Revert "updated .gitignore""	10 Nov 2016 7:45	Niclas Jareborg <n	85a4bcf
	Revert "updated .gitignore"	10 Nov 2016 7:44	Niclas Jareborg <n	20b632a

Commit: 40578d7af0d0c20ae51d99a385f84a1213df8173 [40578d7]
Parents: 8971730c20
Author: Niclas Jareborg <niclas@jareborg.se>
Date: Thursday, November 10, 2016 10:41:42 AM
Labels: HEAD, ->, master, origin/master

Added code to read raw data in notebook

results/2016-11-10.Rmd

Suggested best practices

- There is a **folder for the raw data**, which do not get altered, or intermixed with data that is the result of manual or programmatic manipulation. I.e., derived data is kept separate from raw data, and **raw data are not duplicated**.
- **Code is kept separate from data**.
- Use a **version control system** (at least for code) – e.g. **git**
- There is a **scratch directory for experimentation**. Everything in the scratch directory can be deleted at any time without negative impact.
- There should be a **README in every directory**, describing the purpose of the directory and its contents.
- Use **file naming schemes** that makes it easy to find files and understand what they are (for humans and machines)
- Use **non-proprietary formats** – *.csv* rather than *.xlsx*
- Etc…

File naming

- Three principles
  1. Machine readable
  2. Human readable
  3. Plays well with default ordering

NO

myabstract.docx
Joe’s Filenames Use Spaces and Punctuation.xlsx
figure 1.png
fig 2.png
JW7d^(2sl@deletethisandyourcareerisoverWx2*.txt

YES

2014-06-08_abstract-for-sla.docx
joes-filenames-are-getting-better.xlsx
fig01_scatterplot-talk-length-vs-interest.png
fig02_histogram-talk-attendance.png
1986-01-28_raw-data-from-challenger-o-rings.txt

Suggested best practices

- There is a **folder for the raw data**, which do not get altered, or intermixed with data that is the result of manual or programmatic manipulation. I.e., derived data is kept separate from raw data, and **raw data are not duplicated**.
- **Code is kept separate from data**.
- Use a **version control system** (at least for code) – e.g. **git**
- There is a **scratch directory for experimentation**. Everything in the scratch directory can be deleted at any time without negative impact.
- There should be a **README in every directory**, describing the purpose of the directory and its contents.
- Use **file naming schemes** that makes it easy to find files and understand what they are (for humans and machines)
- Use **non-proprietary formats** – *.csv* rather than *.xlsx*
- Etc…

Non-proprietary formats

- A text-based format is more future-safe, than a proprietary binary format by a commercial vendor
- ***Markdown*** is a nice way of getting nice output from text.
  - Simple & readable formating
  - Can be converted to lots of different outputs
    - HTML, pdf, MS Word, slides etc

- *Never, never, never use **Excel** for scientific **analysis**!*
  - Script your analysis – bash, python, R, …

Tabular data / Spreadsheets

DO

- Keep your raw data raw; calculations and analyses should be done in a copy of the file
- Put variables in columns and observations in rows
- Give each column a descriptive heading that does not include spaces, numbers, or special characters
- Differentiate between zero and null values
- Validate your data
- Keep a separate txt file with a title and a legend describing your dataset, and outlining any steps you take to tidy your data
- Use a version control system and back up your files
- Export each data file in an open non-proprietary format such as CSV or TAB, with a name that appropriately reflects the content of that file
- Check your data thoroughly. Your data should receive the same care as your publications

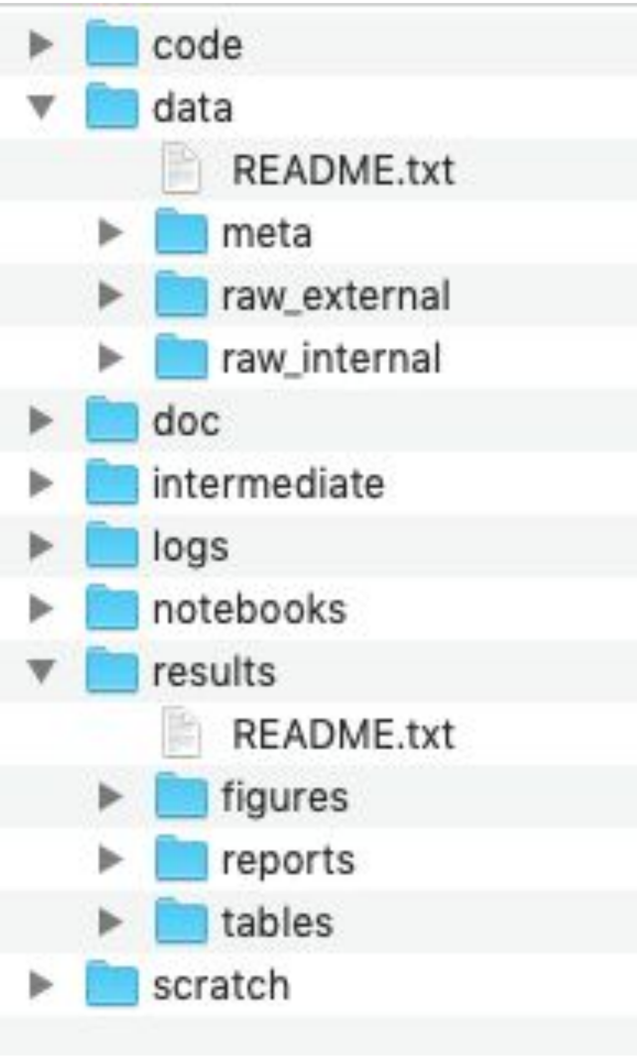
DO NOT

- Put more than 1 piece of information in a cell
- Use colour coding, embedded charts, comments or tables – your spreadsheet is not a lab book
- Include special (i.e. non alphanumeric) characters within the spreadsheet, including commas
- Use merged or blank cells
- Create multiple worksheets within a spreadsheet

F1000

beFAIR beOpen

Directory structure for a sample project



```
▶ code
▼ data
      README.txt
   ▶ meta
   ▶ raw_external
   ▶ raw_internal
▶ doc
▶ intermediate
▶ logs
▶ notebooks
▼ results
      README.txt
   ▶ figures
   ▶ reports
   ▶ tables
▶ scratch
```

all code needed to go from input files to final results

raw and primary data, essentially all input files, **never** edit!

documentation for the study

output files from different analysis steps, *can be deleted*

logs from the different analysis steps

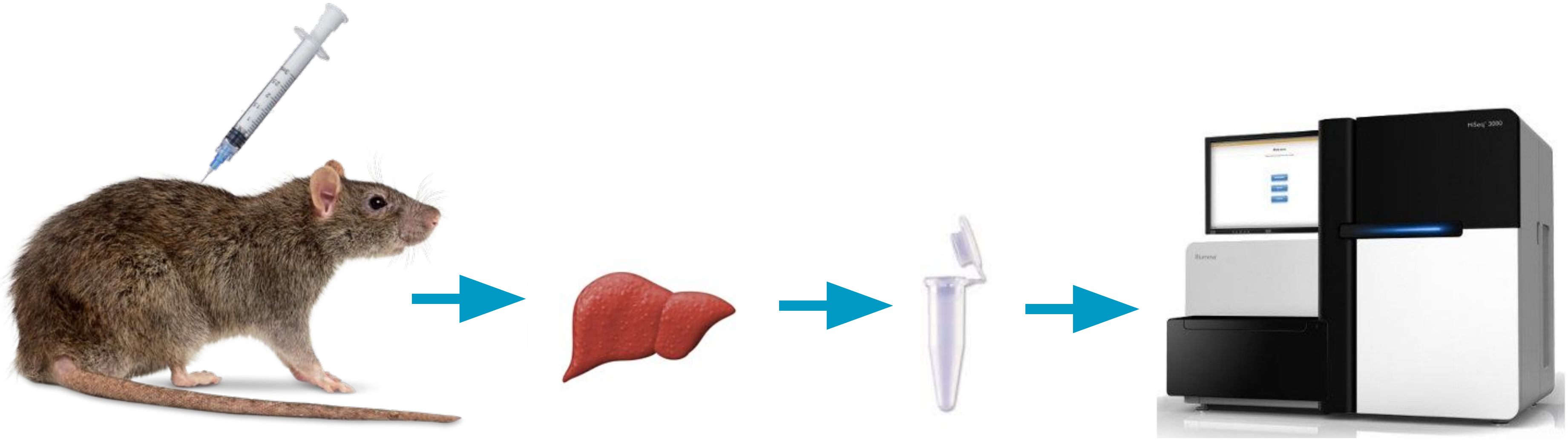
output from workflows and analyses

temporary files that can be safely *deleted or lost*

Noble WS (2009) A Quick Guide to Organizing Computational Biology Projects. PLoS Comput Biol 5(7): e1000424.
http://journals.plos.org/ploscompbiol/article?id=info:doi/10.1371/journal.pcbi.1000424

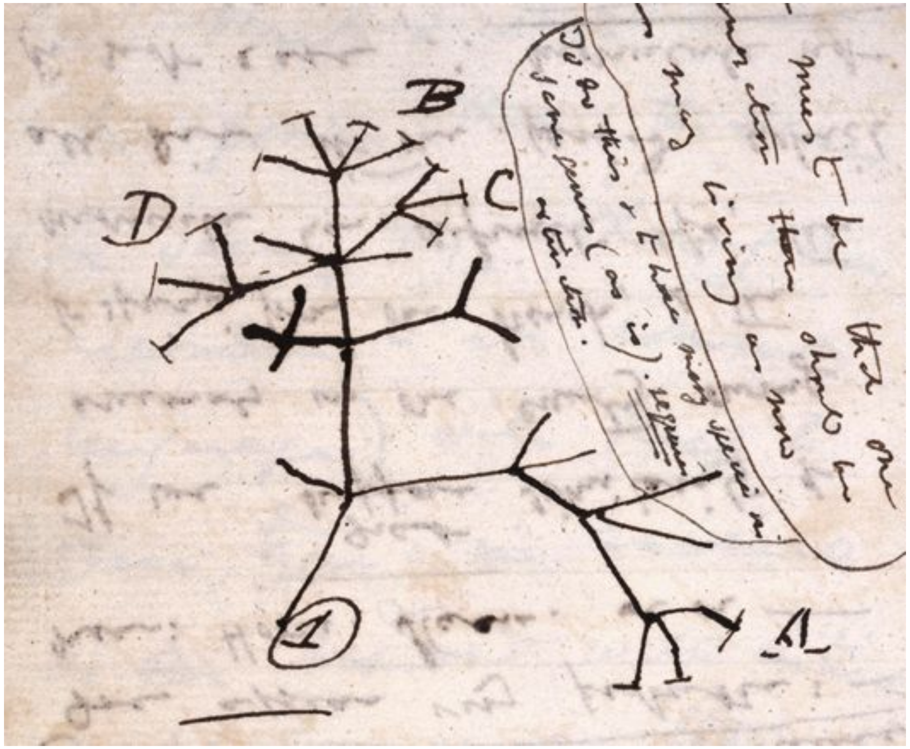
Still missing something

- Need context → document **metadata**
  - From what was the data generated?
  - How do the samples differ?
  - What where the experimental conditions?
  - Etc



Lab notebooks

- Why?
  - You have to understand what you have done
  - **Others should be able to reproduce what you have done**

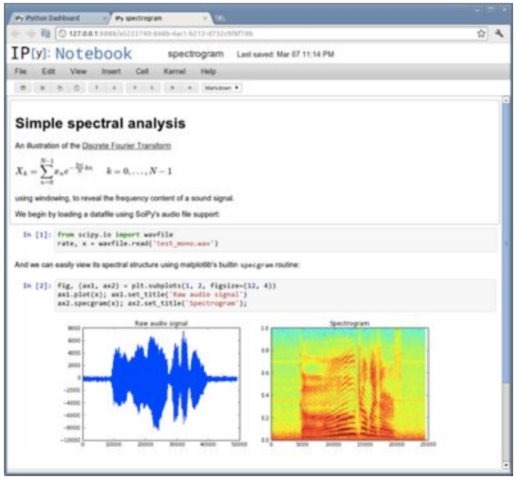

B
C
D
1
A
must be that one
then should be
many living as now
To do this & to have many species
in same genus (as is). requires
extinction.

Lab notes – useful practices

- Put in *separate* directory (e.g. *results*, *documentation*)
- *Dated* entries
- Entries relatively verbose
- Link to *data* and *code* (including versions)
    - Point to commands run and results generated
- Embedded images or tables showing results of analysis done
- Observations, Conclusions, and *ideas* for future work
- Also document analysis that *doesn't* work, so that it can be understood why you choose a particular way of doing the analysis in the end

Where to take down notes

- Paper Notebook
- Word processor program / Text files
- Electronic Lab Notebooks Systems
- 'Interactive' Electronic Notebooks
  - e.g. [jupyther](#), [R Notebooks](#) in RStudio
  - Plain text - work well with version control (Markdown)
  - Embed and execute code
  - Convert to other output formats
    - html, pdf, word



Simple spectral analysis

An illustration of the Discrete Fourier Transform

$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N} kn} \quad k = 0, \ldots, N-1$$

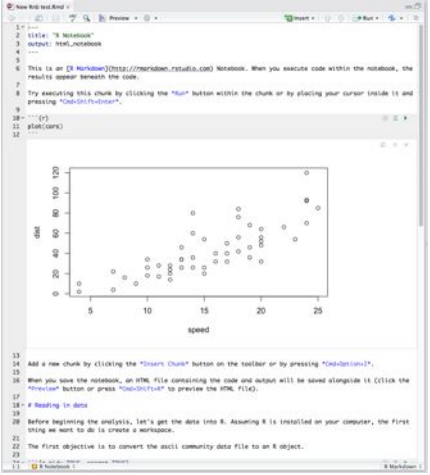
using windowing, to reveal the frequency content of a sound signal.

We begin by loading a datafile using SciPy's audio file support:

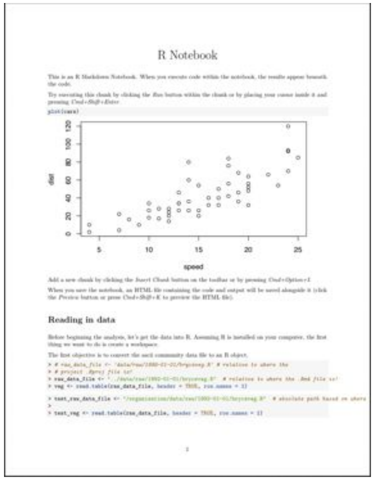
```
In [1]: from scipy.io import wavfile
        rate, x = wavfile.read('test_mono.wav')
```

And we can easily view its spectral structure using matplotlib's builtin specgram routine:

```
In [2]: fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(12, 4))
        ax1.plot(x); ax1.set_title('Raw audio signal')
        ax2.specgram(x); ax2.set_title('Spectrogram');
```



~~~
---
title: "R Notebook"
output: html_notebook
---

This is an [R Markdown](http://rmarkdown.rstudio.com) Notebook. When you execute code within the notebook, the results appear beneath the code.

Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Cmd+Shift+Enter*.

```{r}
plot(cars)
```

Add a new chunk by clicking the *Insert Chunk* button on the toolbar or by pressing *Cmd+Option+I*.

When you save the notebook, an HTML file containing the code and output will be saved alongside it (click the *Preview* button or press *Cmd+Shift+K* to preview the HTML file).

# Reading in data

Before beginning the analysis, let's get the data into R. Assuming R is installed on your computer, the first thing we want to do is create a workspace.

The first objective is to convert the ascii community data file to an R object.
~~~



R Notebook

This is an R Markdown Notebook. When you execute code within the notebook, the results appear beneath the code.

Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Cmd+Shift+Enter*.

```
plot(cars)
```

Add a new chunk by clicking the *Insert Chunk* button on the toolbar or by pressing *Cmd+Option+I*.

When you save the notebook, an HTML file containing the code and output will be saved alongside it (click the *Preview* button or press *Cmd+Shift+K* to preview the HTML file).

Reading in data

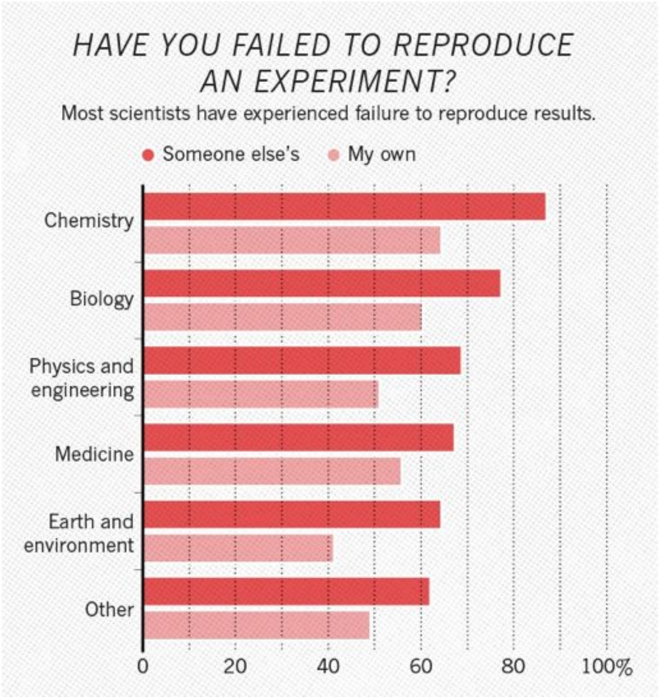
Before beginning the analysis, let's get the data into R. Assuming R is installed on your computer, the first thing we want to do is create a workspace.

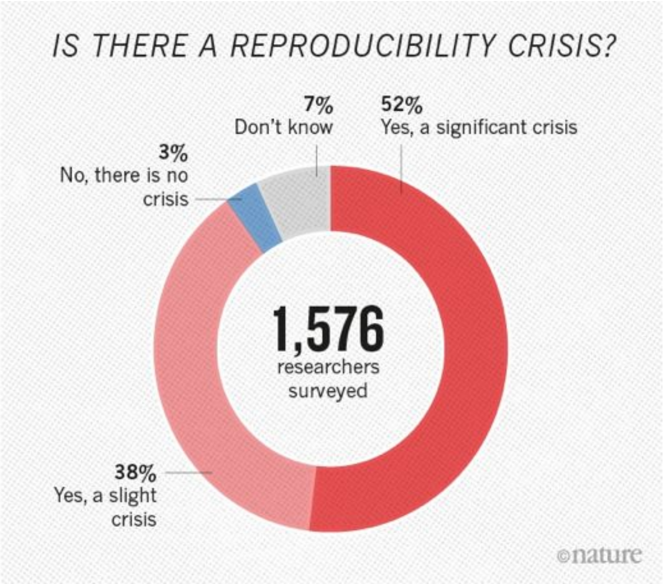
The first objective is to convert the ascii community data file to an R object.

```
> # raw_data_file <- [illegible] # relative to where the
> # project .Rproj file is"
> raw_data_file <- [illegible] # relative to where the .Rmd file is!
> veg <- read.table(raw_data_file, header = TRUE, row.names = 1)

> test_raw_data_file <- [illegible] # absolute path based on where
>
> test_veg <- read.table(raw_data_file, header = TRUE, row.names = 1)
```

1

A reproducibility crisis

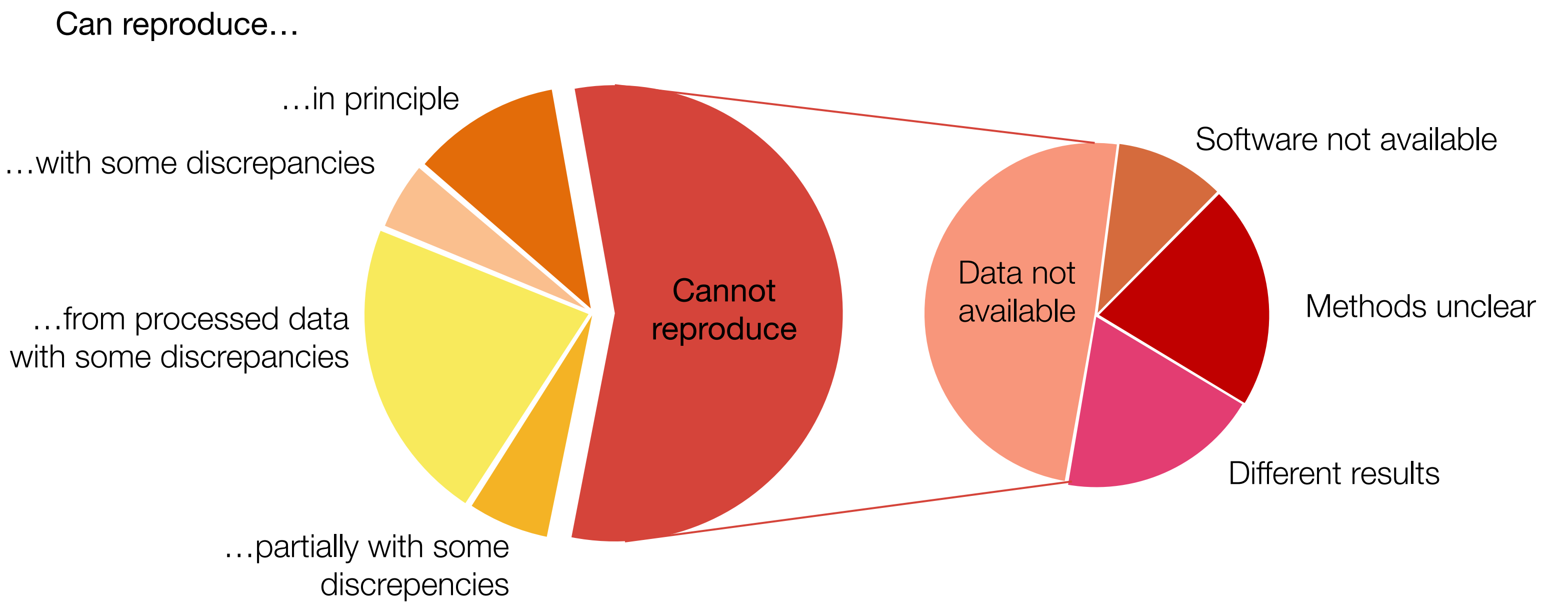




[1] "1,500 scientists lift the lid on reproducibility". Nature. 533: 452–454
[2] Begley, C. G.; Ellis, L. M. (2012). "Drug development: Raise standards for preclinical cancer research". Nature. 483 (7391): 531–533.

A reproducibility crisis

Reproduction of data analyses in 18 articles on microarray-based gene expression profiling published in Nature Genetics in 2005–2006:

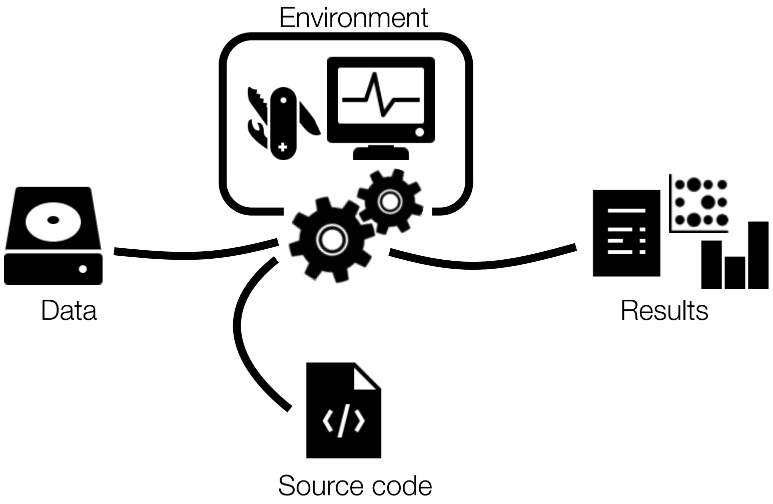


Summary of the efforts to replicate the published analyses.
Adopted from: Ioannidis et al. Repeatability of published microarray gene expression analyses. *Nature Genetics* 41 (2009) doi:10.1038/ng.295

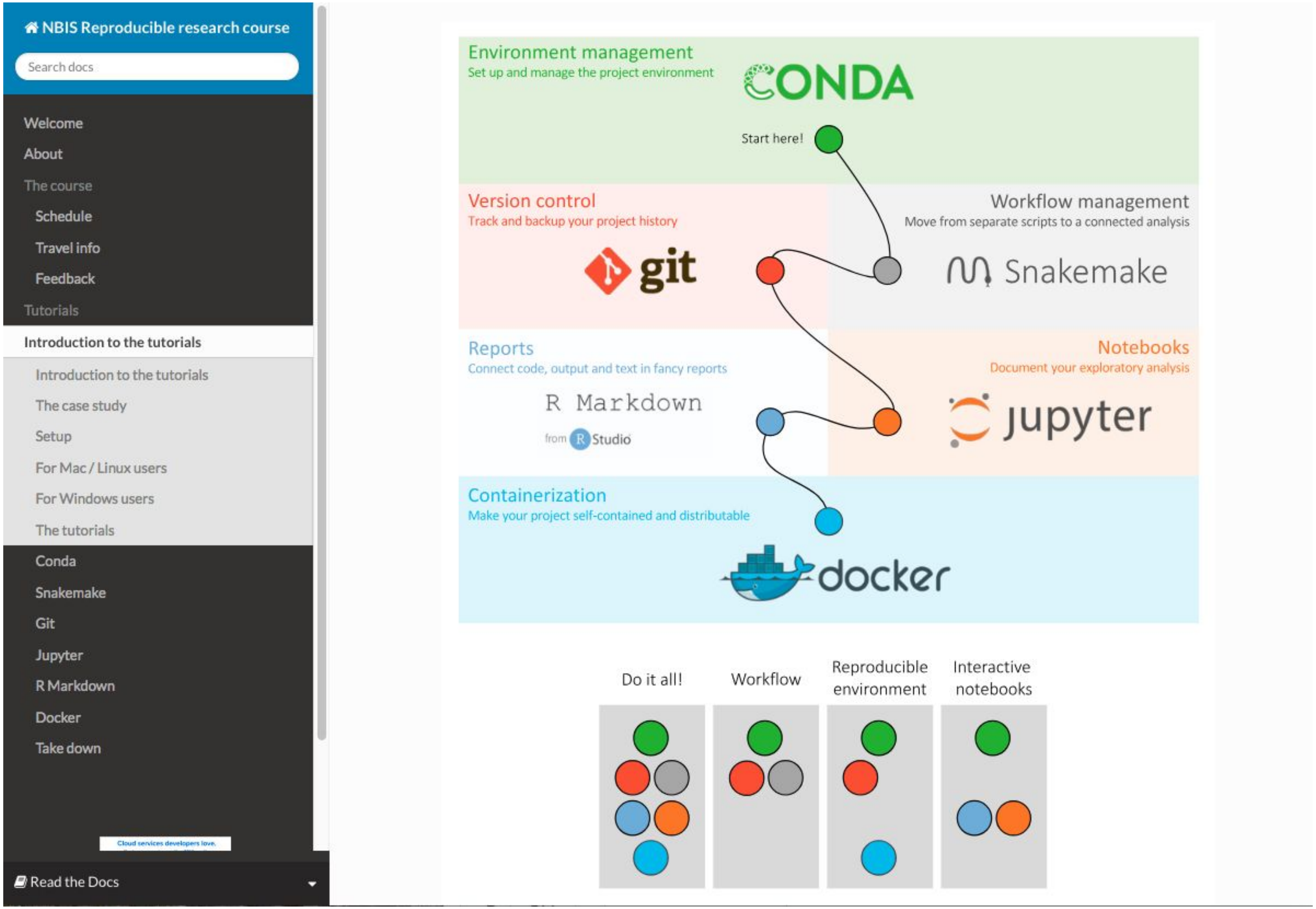
What do we mean by **reproducible research?**

		Data	
		Same	Different
Code	Same	Reproducible	Replicable
	Different	Robust	Generalizable

All parts of a bioinformatics analysis have to be reproducible:



Reproducible Research tutorials



Environment management
Set up and manage the project environment

Version control
Track and backup your project history

Workflow management
Move from separate scripts to a connected analysis

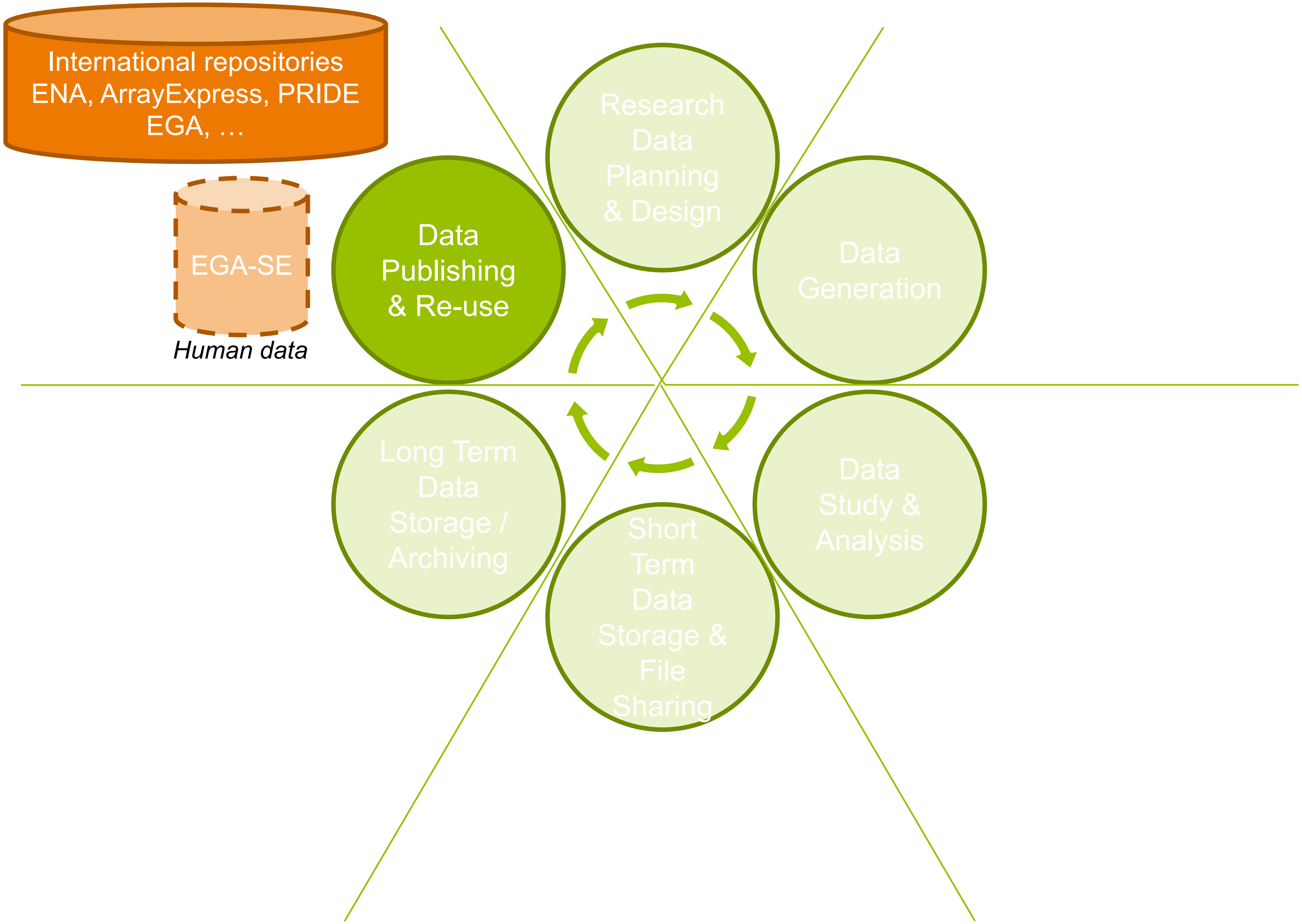
Reports
Connect code, output and text in fancy reports

Notebooks
Document your exploratory analysis

Containerization
Make your project self-contained and distributable

https://nbis-reproducible-research.readthedocs.io/en/course_1803/

Data Publishing & Re-use



International repositories
ENA, ArrayExpress, PRIDE
EGA, …

EGA-SE

Human data

Research Data Planning & Design

Data Generation

Data Study & Analysis

Short Term Data Storage & File Sharing

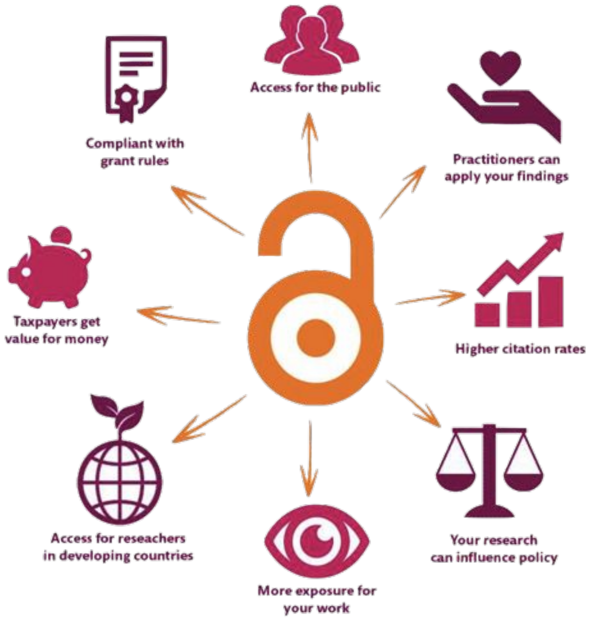
Long Term Data Storage / Archiving

Data Publishing & Re-use

Why should you make research data available for others?

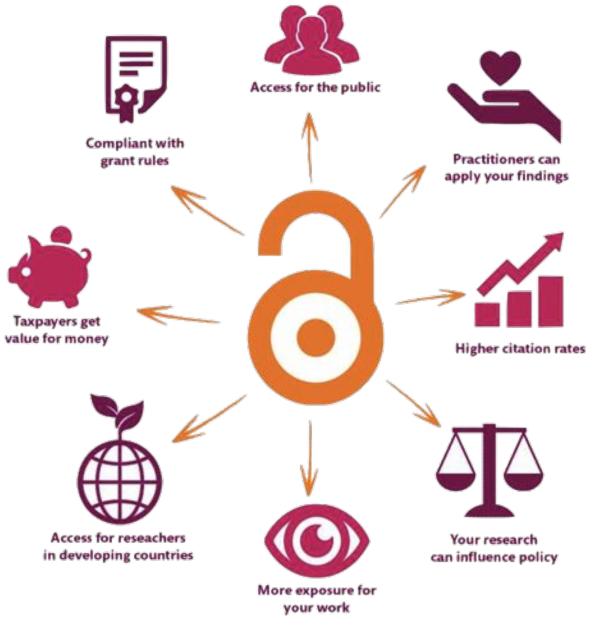
Open Access

*The practice of providing **on-line access** to scientific information that is **free of charge** to the end-user and that is **re-usable**.*


Access for the public
Compliant with grant rules
Practitioners can apply your findings
Taxpayers get value for money
Higher citation rates
Access for reseachers in developing countries
Your research can influence policy
More exposure for your work

Open Access

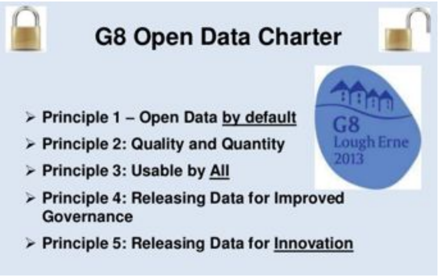
- Democracy and transparency
  - Publicly funded research data should be accessible to all
  - Published results and conclusions should be possible to check by others
- Research
  - Enables others to combine data, address new questions, and develop new analytical methods
  - Reduce duplication and waste
- Innovation and utilization outside research
  - Public authorities, companies, and private persons outside research can make use of the data
- Citation
  - Citation of data will be a merit for the researcher that produced it


Access for the public
Compliant with grant rules
Practitioners can apply your findings
Taxpayers get value for money
Higher citation rates
Access for reseachers in developing countries
More exposure for your work
Your research can influence policy

Open Access to research data

OPEN ACCESS

- Strong international movement towards Open Access (OA)
- European Commission recommended the member states to establish national guidelines for OA
  - Swedish Research Council (VR) submitted proposal to the government Jan 2015
- Research bill 2017–2020 – *28 Nov 2016*
  - "*The aim of the government is that all scientific publications that are the result of publicly funded research should be openly accessible as soon as they are published. Likewise,* ***research data*** *underlying scientific publications should be* ***openly accessible*** *at the time of publication.*" [my translation]
- 2018 – VR assigned by the government to coordinate national efforts to implement open access to research data



G8 Open Data Charter

- ➢ Principle 1 – Open Data by default
- ➢ Principle 2: Quality and Quantity
- ➢ Principle 3: Usable by All
- ➢ Principle 4: Releasing Data for Improved Governance
- ➢ Principle 5: Releasing Data for Innovation

G8
Lough Erne
2013

Vetenskapsrådet

FÖRSLAG TILL NATIONELLA
RIKTLINJER FÖR ÖPPEN TILLGÅNG TILL
VETENSKAPLIG INFORMATION

VETENSKAPSRÅDETS RAPPORTER

Regeringens proposition
2016/17:50

Kunskap i samverkan – för samhällets utmaningar och stärkt konkurrenskraft

Prop.
2016/17:50

Regeringen överlämnar denna proposition till riksdagen.

Stockholm den 24 november 2016

Stefan Löfven

Helene Hellmark Knutsson
(Utbildningsdepartementet)

Propositionens huvudsakliga innehåll

I propositionen presenterar regeringen sin syn på forskningspolitikens inriktning i ett tioårigt perspektiv, med särskilt fokus på satsningar 2017–2020. Syftet är att Sverige ska vara en ledande kunskapsnation och ett av världens främsta forsknings- och innovationsländer.

En utgångspunkt är att värna den fria forskningen samtidigt som forskningspolitiken svarar mot globala och nationella samhällsutmaningar. Prioriterade utmaningar är klimat och miljö, hälsa, ökad digitalisering, ett hållbart samhälle och förbättrade kunskapsresultat i det svenska skol- och utbildningssystemet. Ett antal satsningar som stärker den fria forskningen och svarar mot samhällsutmaningarna presenteras.

Förutsättningarna för universitetens och högskolornas långsiktiga strategiska planering bör stärkas genom ökade anslag för forskning och utbildning på forskarnivå. Även lärosätenas samverkan med det omgivande samhället och sambandet mellan utbildning och forskning förväntas stärkas. Vidare redovisar regeringen ett antal åtgärder för att främja attraktiva villkor för studerande på forskarnivå och unga forskare och för att främja jämställdhet i högskola och forskning.

Regeringen har i budgetpropositionen för 2017 lämnat förslag och redovisat bedömningar om fördelning av medel till forskning och innovation. I denna proposition beskrivs satsningarna närmare. Forskningssatsningarna avser dels långsiktiga nationella forskningsprogram som utgår från de utpekade samhällsutmaningarna, dels vissa särskilda forskningssatsningar.

Satsningarna på innovation avser bl.a. en förstärkning av strategiska innovationsområden, vilka ska kopplas till prioriteringarna i regeringens

1

What is needed for others to be able to re-use your data?

Data Management Snafu

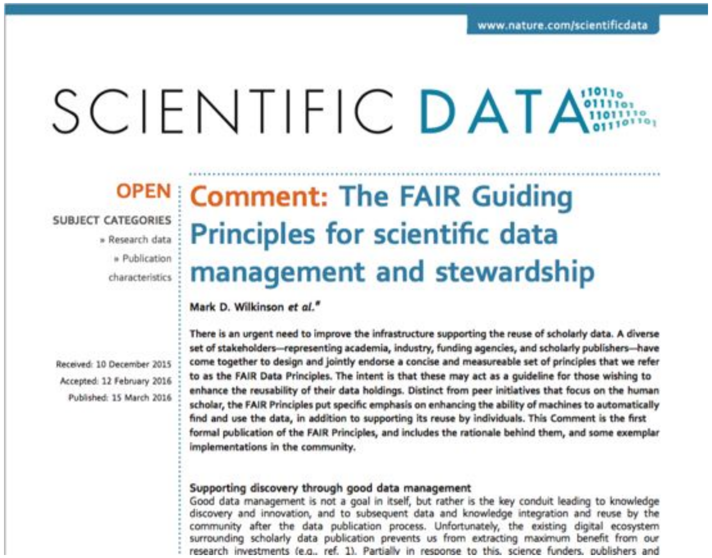
Hello! My name is Dr. Judy Benign, I'm an oncologist at NYU School of Medicine.

https://www.youtube.com/watch?v=N2zK3sAtr-4

FAIR

- To be useful for others data should be
  - ***FAIR*** - Findable, Accessible, Interoperable, and Reusable
    … for both Machines and Humans

Wilkinson, Mark et al. *"The FAIR Guiding Principles for scientific data management and stewardship"*. Scientific Data 3, Article number: 160018 (2016) http://dx.doi.org/10.1038/sdata.2016.18



www.nature.com/scientificdata

SCIENTIFIC DATA

OPEN

SUBJECT CATEGORIES

» Research data

» Publication characteristics

Received: 10 December 2015

Accepted: 12 February 2016

Published: 15 March 2016

Comment: The FAIR Guiding Principles for scientific data management and stewardship

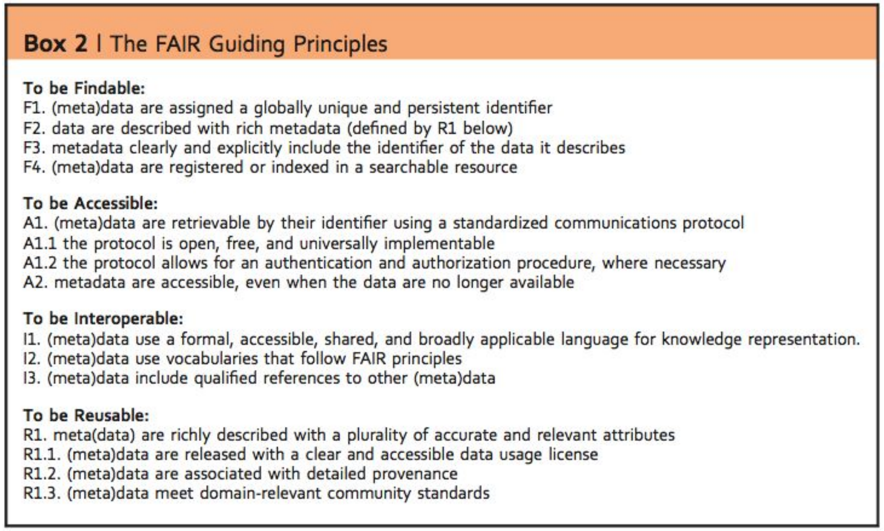
Mark D. Wilkinson *et al.*#

There is an urgent need to improve the infrastructure supporting the reuse of scholarly data. A diverse set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—have come together to design and jointly endorse a concise and measureable set of principles that we refer to as the FAIR Data Principles. The intent is that these may act as a guideline for those wishing to enhance the reusability of their data holdings. Distinct from peer initiatives that focus on the human scholar, the FAIR Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals. This Comment is the first formal publication of the FAIR Principles, and includes the rationale behind them, and some exemplar implementations in the community.

Supporting discovery through good data management

Good data management is not a goal in itself, but rather is the key conduit leading to knowledge discovery and innovation, and to subsequent data and knowledge integration and reuse by the community after the data publication process. Unfortunately, the existing digital ecosystem surrounding scholarly data publication prevents us from extracting maximum benefit from our research investments (e.g., ref. 1). Partially in response to this, science funders, publishers and

DOI: 10.1038/sdata.2016.18



Box 2 | The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

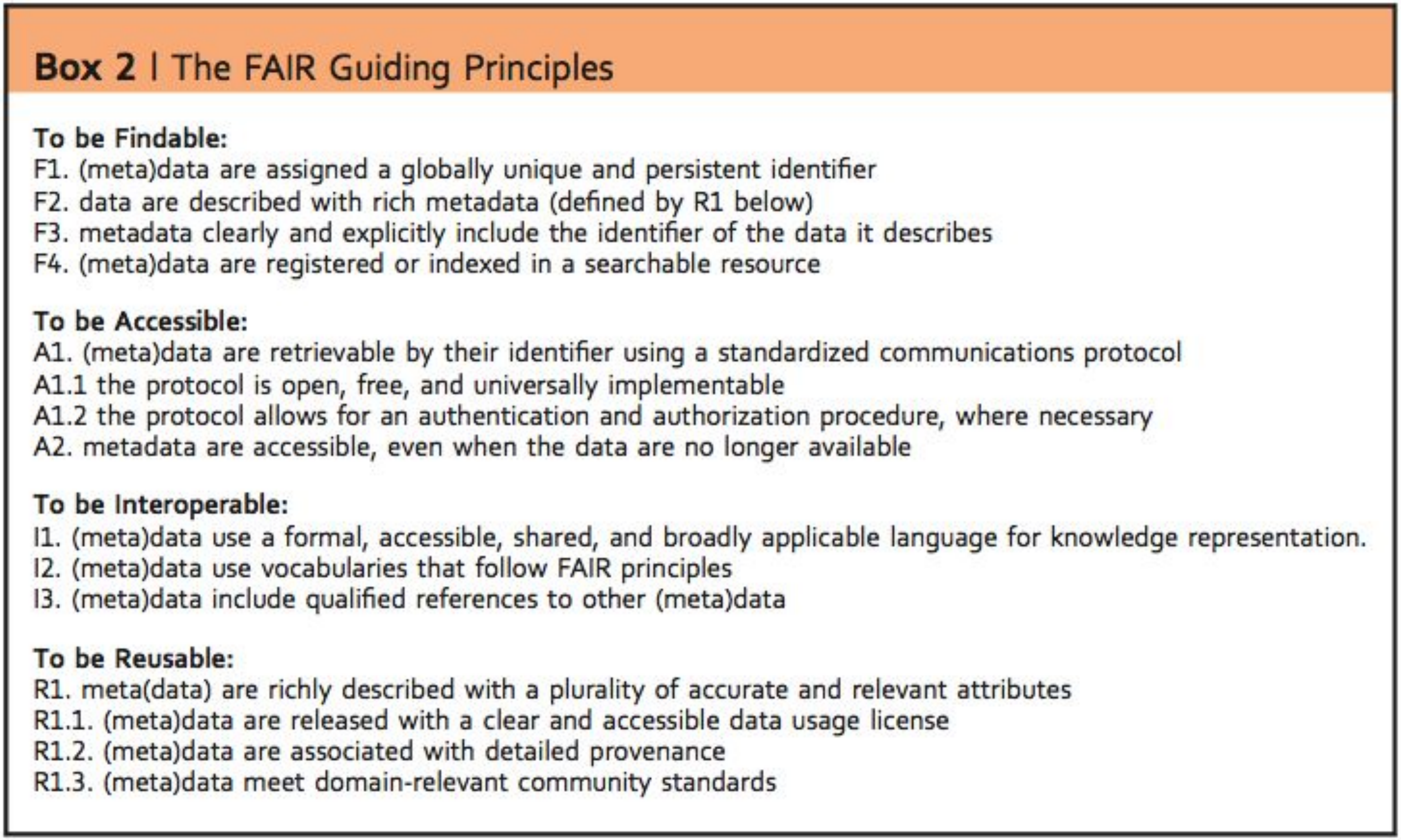
To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

To be Interoperable:
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

To be Reusable:
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

‘We support appropriate efforts to promote open science and facilitate appropriate access to publicly funded research results on findable, accessible, interoperable and reusable (FAIR)’

FAIR



Box 2 | The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

To be Interoperable:
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

To be Reusable:
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

Findable

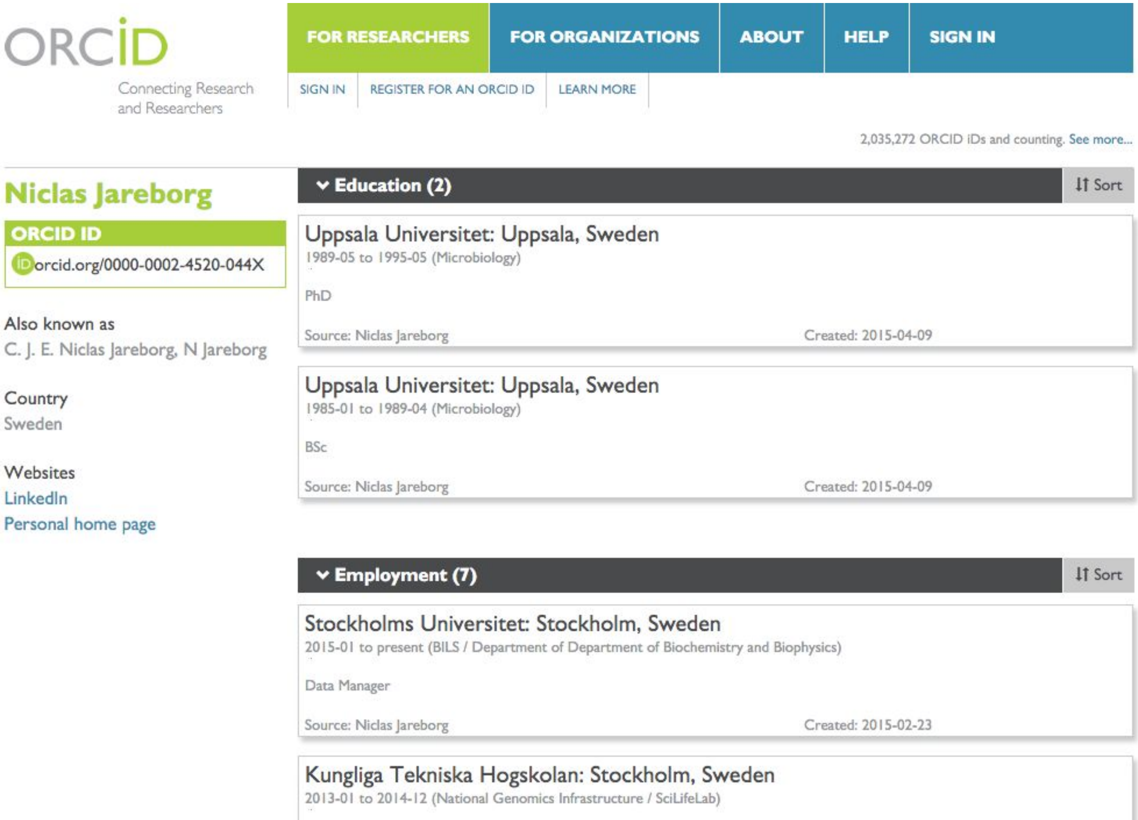
- Long-term storage
  - Data should not disappear
- Persistent identifiers
  - Possibility to refer to a dataset over long periods of time
  - Unique
  - e.g. DOIs (Digital Object Identifiers)
- Discoverability
  - Expose dataset metadata through search functionalities

OPEN ACCESS

doi

Persistent identifier for yourself

- ORCID is an open, non-profit, community-driven effort to create and maintain a registry of unique researcher identifiers and a transparent method of linking research activities and outputs to these identifiers.
- http://orcid.org
- Persistent identifier for you as a researcher



2,035,272 ORCID iDs and counting. See more...

Niclas Jareborg

ORCID ID

orcid.org/0000-0002-4520-044X

Also known as
C. J. E. Niclas Jareborg, N Jareborg

Country
Sweden

Websites
LinkedIn
Personal home page

Education (2)

Uppsala Universitet: Uppsala, Sweden
1989-05 to 1995-05 (Microbiology)
PhD
Source: Niclas Jareborg Created: 2015-04-09

Uppsala Universitet: Uppsala, Sweden
1985-01 to 1989-04 (Microbiology)
BSc
Source: Niclas Jareborg Created: 2015-04-09

Employment (7)

Stockholms Universitet: Stockholm, Sweden
2015-01 to present (BILS / Department of Department of Biochemistry and Biophysics)
Data Manager
Source: Niclas Jareborg Created: 2015-02-23

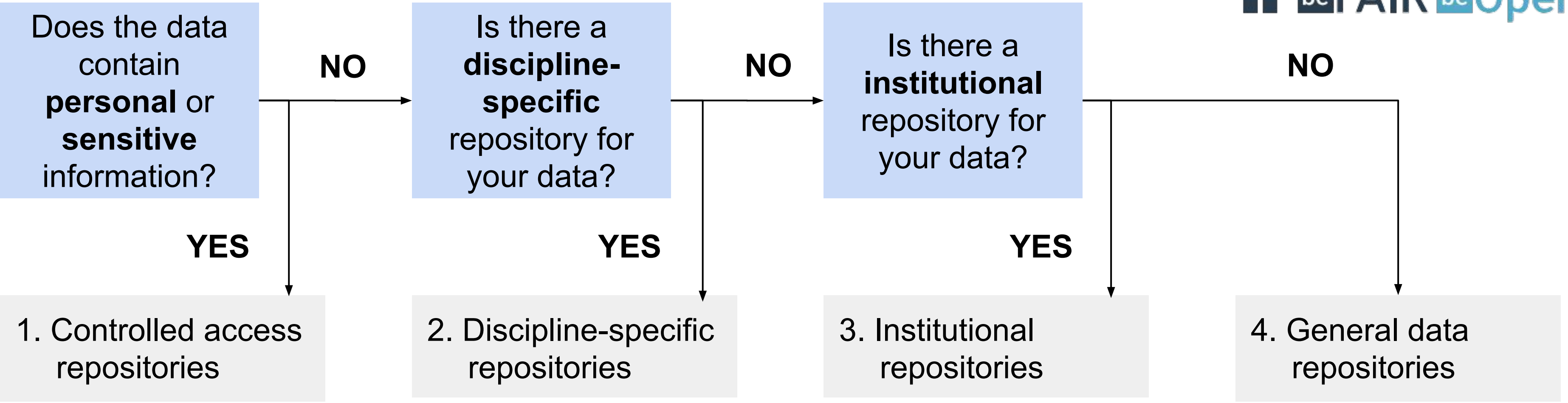
Kungliga Tekniska Hogskolan: Stockholm, Sweden
2013-01 to 2014-12 (National Genomics Infrastructure / SciLifeLab)

NBIS
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

Repositories

SciLifeLab

F1000 beFAIR beOpen



Does the data contain **personal** or **sensitive** information?
NO
Is there a **discipline-specific** repository for your data?
NO
Is there a **institutional** repository for your data?
NO
YES
YES
YES
1. Controlled access repositories
2. Discipline-specific repositories
3. Institutional repositories
4. General data repositories

ENA European Nucleotide Archive

BioSamples

EUROPEAN GENOME-PHENOME ARCHIVE

NCBI

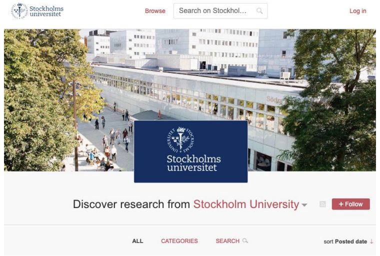
dbSNP Short Genetic Variations

PRIDE

Protein Data Bank in Europe
Bringing Structure to Biology

Etc...

SND Svensk nationell datatjänst



Discover research from Stockholm University

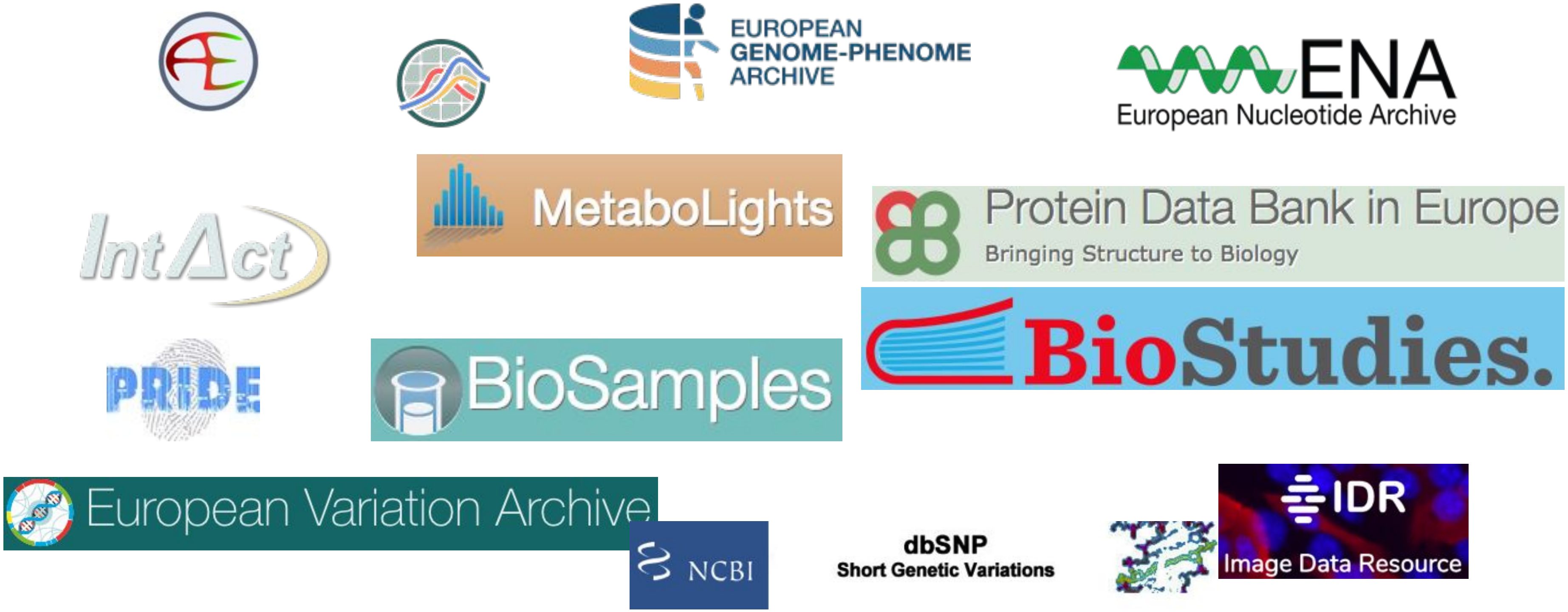
figshare

zenodo

EUDAT

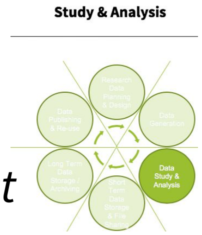
Accessible

International public repositories


EUROPEAN GENOME-PHENOME ARCHIVE
ENA
European Nucleotide Archive
IntAct
MetaboLights
Protein Data Bank in Europe
Bringing Structure to Biology
PRIDE
BioSamples
BioStudies.
European Variation Archive
NCBI
dbSNP
Short Genetic Variations
IDR
Image Data Resource

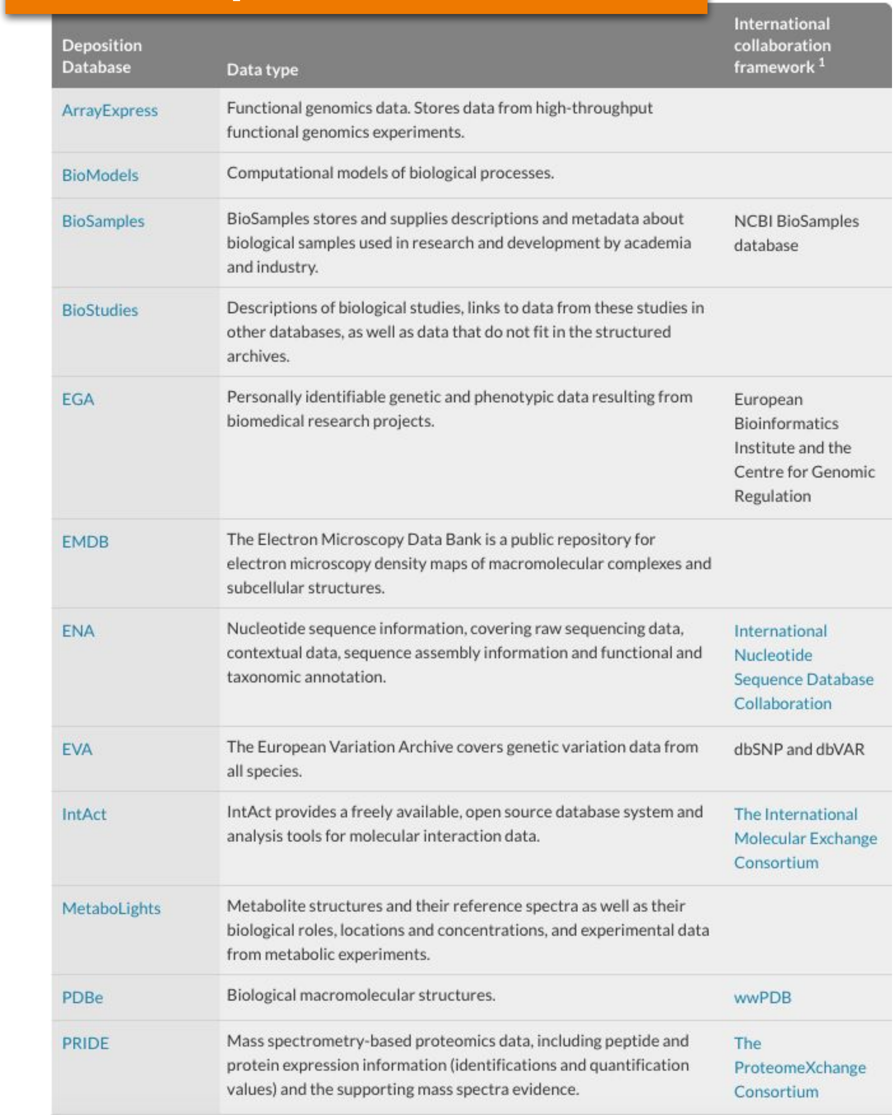
- Best way to make data **FAIR**
- Domain-specific metadata standards

💡 *Strive towards uploading data to its final destination already at the beginning of a project*



Recommended repositories

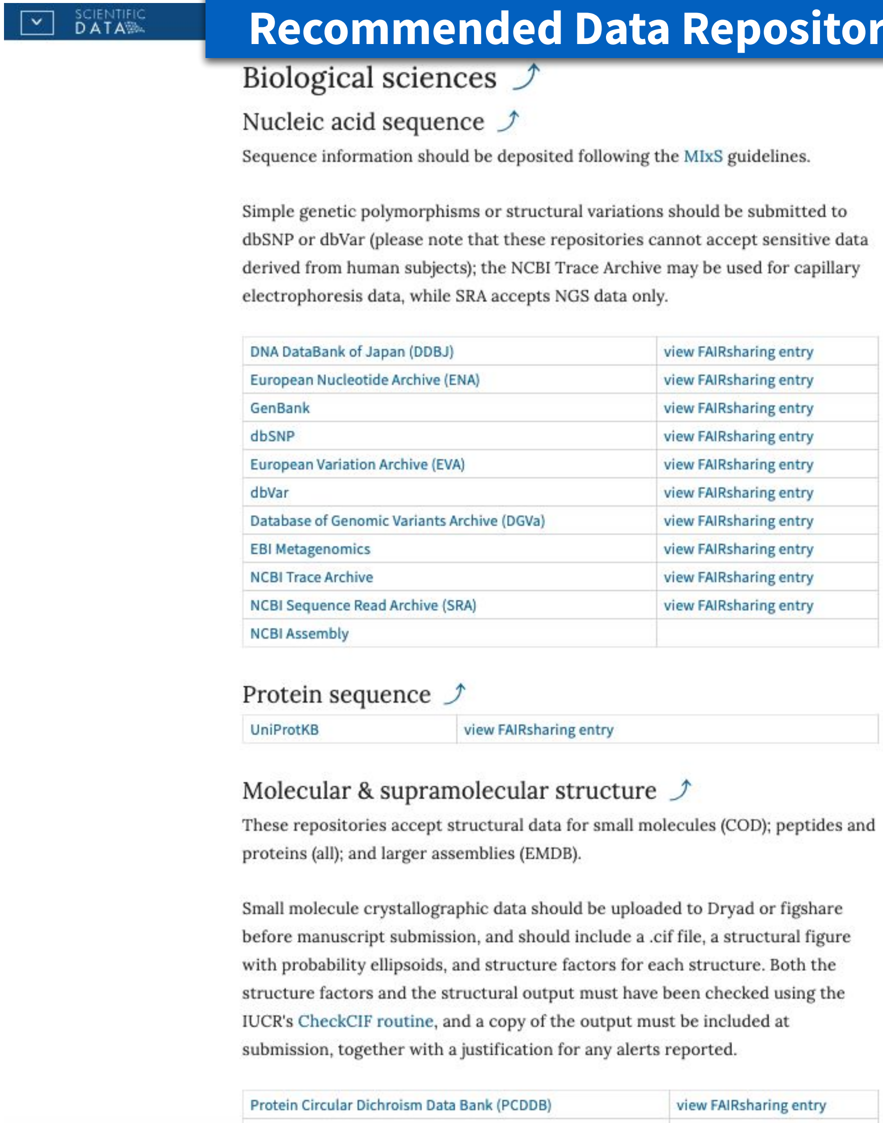
ELIXIR Deposition Database list



Deposition Database	Data type	International collaboration framework [1]
ArrayExpress	Functional genomics data. Stores data from high-throughput functional genomics experiments.	
BioModels	Computational models of biological processes.	
BioSamples	BioSamples stores and supplies descriptions and metadata about biological samples used in research and development by academia and industry.	NCBI BioSamples database
BioStudies	Descriptions of biological studies, links to data from these studies in other databases, as well as data that do not fit in the structured archives.	
EGA	Personally identifiable genetic and phenotypic data resulting from biomedical research projects.	European Bioinformatics Institute and the Centre for Genomic Regulation
EMDB	The Electron Microscopy Data Bank is a public repository for electron microscopy density maps of macromolecular complexes and subcellular structures.	
ENA	Nucleotide sequence information, covering raw sequencing data, contextual data, sequence assembly information and functional and taxonomic annotation.	International Nucleotide Sequence Database Collaboration
EVA	The European Variation Archive covers genetic variation data from all species.	dbSNP and dbVAR
IntAct	IntAct provides a freely available, open source database system and analysis tools for molecular interaction data.	The International Molecular Exchange Consortium
MetaboLights	Metabolite structures and their reference spectra as well as their biological roles, locations and concentrations, and experimental data from metabolic experiments.	
PDBe	Biological macromolecular structures.	wwPDB
PRIDE	Mass spectrometry-based proteomics data, including peptide and protein expression information (identifications and quantification values) and the supporting mass spectra evidence.	The ProteomeXchange Consortium

https://www.elixir-europe.org/platforms/data/elixir-deposition-databases

Scientific Data Recommended Data Repositories



Biological sciences

Nucleic acid sequence

Sequence information should be deposited following the MIxS guidelines.

Simple genetic polymorphisms or structural variations should be submitted to dbSNP or dbVar (please note that these repositories cannot accept sensitive data derived from human subjects); the NCBI Trace Archive may be used for capillary electrophoresis data, while SRA accepts NGS data only.

DNA DataBank of Japan (DDBJ)	view FAIRsharing entry
European Nucleotide Archive (ENA)	view FAIRsharing entry
GenBank	view FAIRsharing entry
dbSNP	view FAIRsharing entry
European Variation Archive (EVA)	view FAIRsharing entry
dbVar	view FAIRsharing entry
Database of Genomic Variants Archive (DGVa)	view FAIRsharing entry
EBI Metagenomics	view FAIRsharing entry
NCBI Trace Archive	view FAIRsharing entry
NCBI Sequence Read Archive (SRA)	view FAIRsharing entry
NCBI Assembly	

Protein sequence

UniProtKB	view FAIRsharing entry

Molecular & supramolecular structure

These repositories accept structural data for small molecules (COD); peptides and proteins (all); and larger assemblies (EMDB).

Small molecule crystallographic data should be uploaded to Dryad or figshare before manuscript submission, and should include a .cif file, a structural figure with probability ellipsoids, and structure factors for each structure. Both the structure factors and the structural output must have been checked using the IUCR's CheckCIF routine, and a copy of the output must be included at submission, together with a justification for any alerts reported.

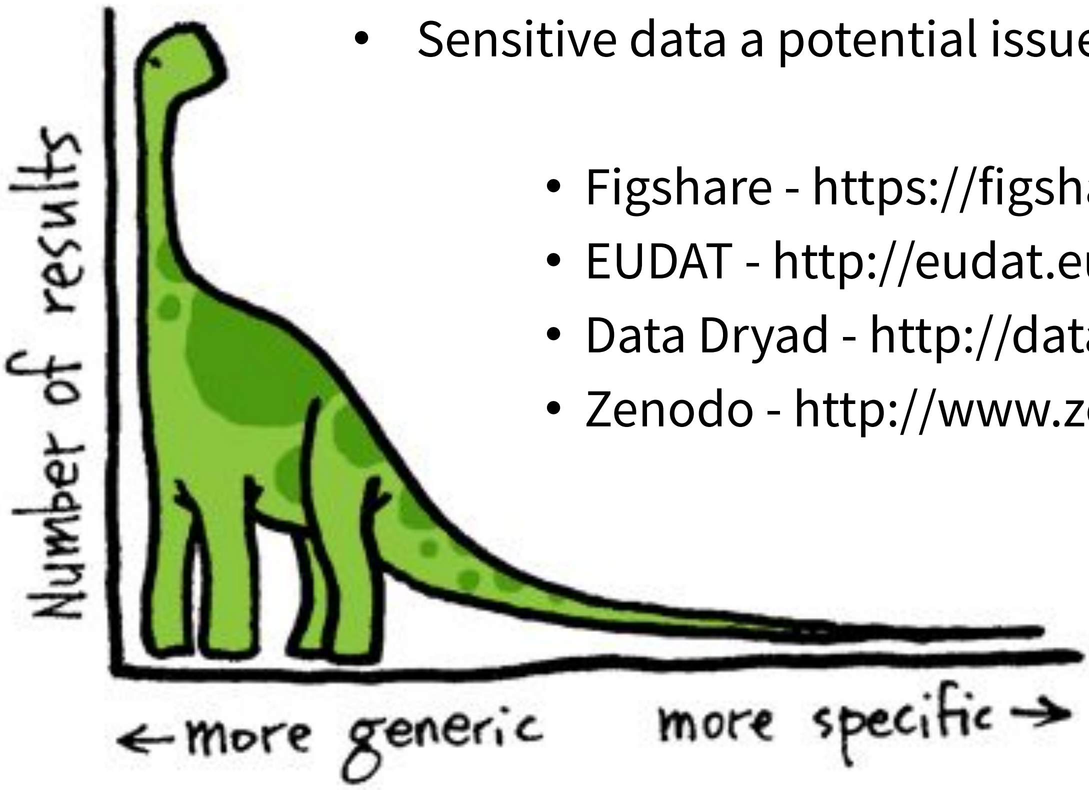
Protein Circular Dichroism Data Bank (PCDDB)	view FAIRsharing entry

https://www.nature.com/sdata/policies/repositories#life

"Long-tail data" repositories

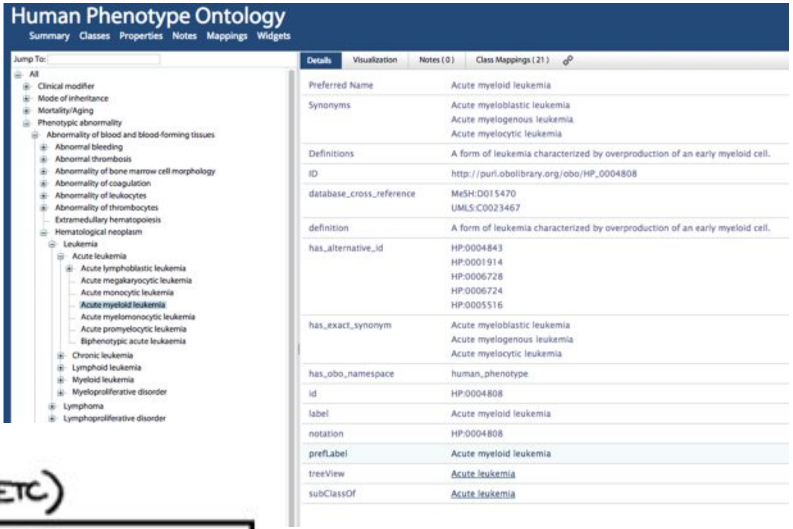
- Research data that doesn't fit in structured data repositories
- Data publication – persistent identifiers
- Metadata submission – not tailored to Life Science
  - *Affects discoverability*
  - *(Less) FAIR*
- Sensitive data a potential issue

- Figshare - https://figshare.com/
- EUDAT - http://eudat.eu/
- Data Dryad - http://datadryad.org/
- Zenodo - http://www.zenodo.org/



Interoperable & Re-useable *Metadata*

- Standards
  - Controlled vocabularies / Ontologies
    - Agreed terms for different phenomena



Human Phenotype Ontology

- All
  - Clinical modifier
  - Mode of inheritance
  - Mortality/Aging
  - Phenotypic abnormality
    - Abnormality of blood and blood-forming tissues
      - Abnormal bleeding
      - Abnormal thrombosis
      - Abnormality of bone marrow cell morphology
      - Abnormality of coagulation
      - Abnormality of leukocytes
      - Abnormality of thrombocytes
      - Extramedullary hematopoiesis
      - Hematological neoplasm
        - Leukemia
          - Acute leukemia
            - Acute lymphoblastic leukemia
            - Acute megakaryocytic leukemia
            - Acute monocytic leukemia
            - Acute myeloid leukemia
            - Acute myelomonocytic leukemia
            - Acute promyelocytic leukemia
            - Biphenotypic acute leukaemia
          - Chronic leukemia
          - Lymphoid leukemia
          - Myeloid leukemia
          - Myeloproliferative disorder
        - Lymphoma
        - Lymphoproliferative disorder

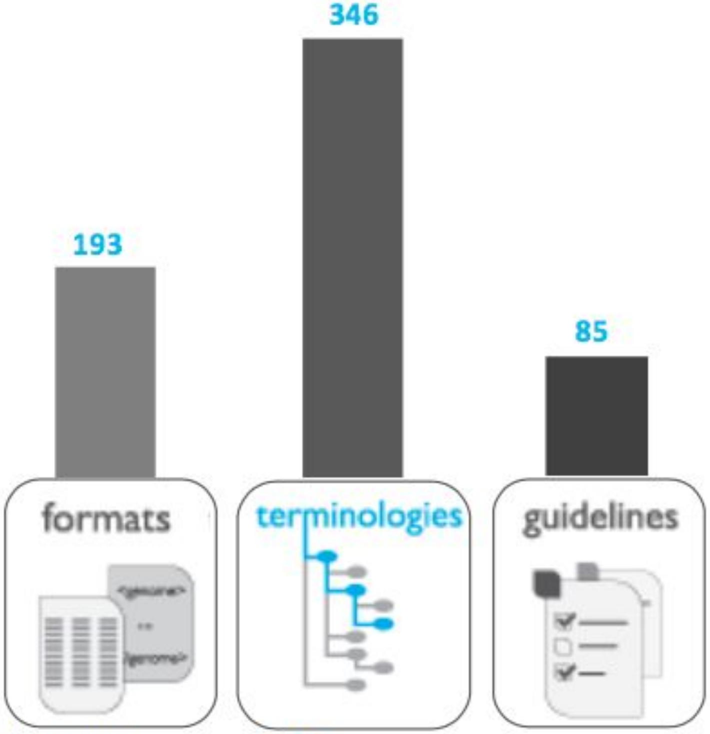
Preferred Name	Acute myeloid leukemia
Synonyms	Acute myeloblastic leukemia Acute myelogenous leukemia Acute myelocytic leukemia
Definitions	A form of leukemia characterized by overproduction of an early myeloid cell.
ID	http://purl.obolibrary.org/obo/HP_0004808
database_cross_reference	MeSH:D015470 UMLS:C0023467
definition	A form of leukemia characterized by overproduction of an early myeloid cell.
has_alternative_id	HP:0004843 HP:0001914 HP:0006728 HP:0006724 HP:0005516
has_exact_synonym	Acute myeloblastic leukemia Acute myelogenous leukemia Acute myelocytic leukemia
has_obo_namespace	human_phenotype
id	HP:0004808
label	Acute myeloid leukemia
notation	HP:0004808
prefLabel	Acute myeloid leukemia
treeView	Acute leukemia
subClassOf	Acute leukemia

HOW STANDARDS PROLIFERATE:
(SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC)

SITUATION: THERE ARE 14 COMPETING STANDARDS.
14?! RIDICULOUS! WE NEED TO DEVELOP ONE UNIVERSAL STANDARD THAT COVERS EVERYONE'S USE CASES.
YEAH!
SOON:
SITUATION: THERE ARE 15 COMPETING STANDARDS.

FAIRsharing.org

In the life sciences there are >600 *content standards*



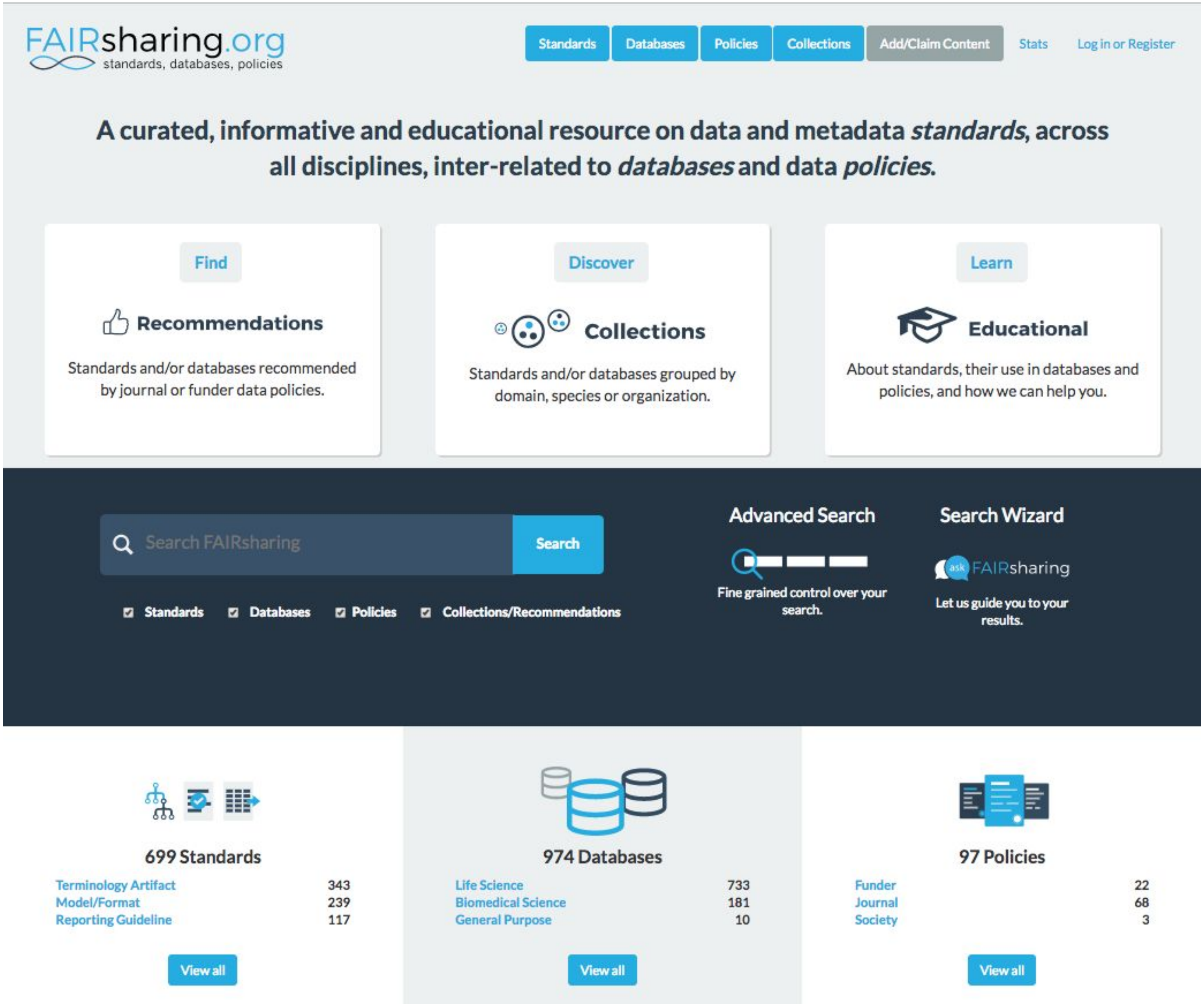
MAGE-Tab
GCDML
SRAxml
SOFT
FASTA
CML
DICOM
GELML
SBRML
MITAB
MzML
ISA-Tab
SEDML…

AAO
CHEBI
OBI
VO
PATO
ENVO
MOD
TEDDY
XAO
BTO
DO
PRO
IDO…

miame
MIAPA
MIRIAM
MIQAS
MIX
REMARK
MIGEN
MIQE
MIAPE
ARRIVE
CONSORT
MIASE
MISFISHIE….

FAIRsharing.org

(was biosharing.org)



FAIRsharing.org
standards, databases, policies

Standards | Databases | Policies | Collections | Add/Claim Content | Stats | Log in or Register

A curated, informative and educational resource on data and metadata *standards*, across all disciplines, inter-related to *databases* and data *policies*.

Find

Recommendations

Standards and/or databases recommended by journal or funder data policies.

Discover

Collections

Standards and/or databases grouped by domain, species or organization.

Learn

Educational

About standards, their use in databases and policies, and how we can help you.

Search FAIRsharing | Search

☑ Standards ☑ Databases ☑ Policies ☑ Collections/Recommendations

Advanced Search

Fine grained control over your search.

Search Wizard

ask FAIRsharing

Let us guide you to your results.

699 Standards

Terminology Artifact	343
Model/Format	239
Reporting Guideline	117

View all

974 Databases

Life Science	733
Biomedical Science	181
General Purpose	10

View all

97 Policies

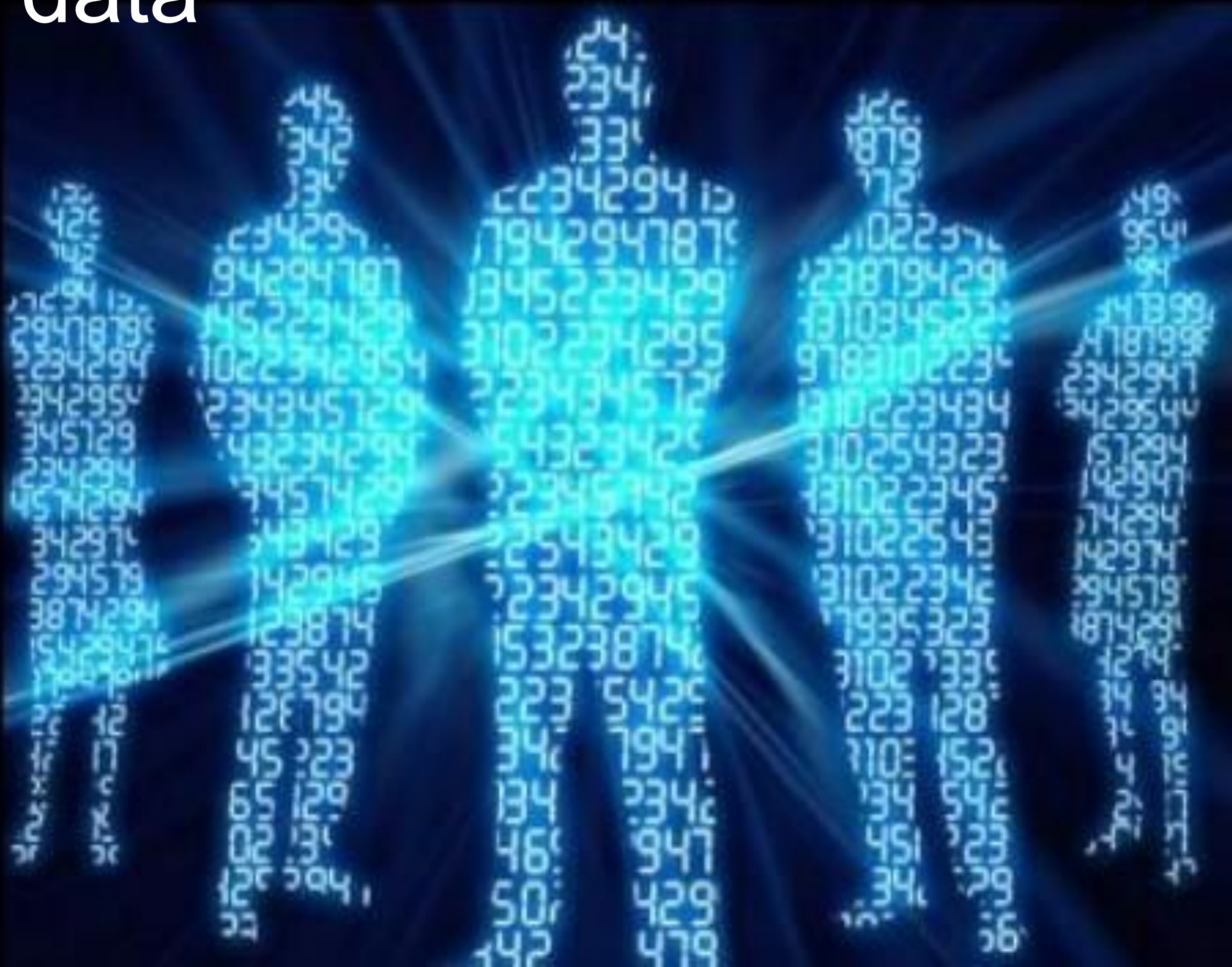
Funder	22
Journal	68
Society	3

View all

Can you share all types of data publicly?

If not, what would be the reasons?

Personal data



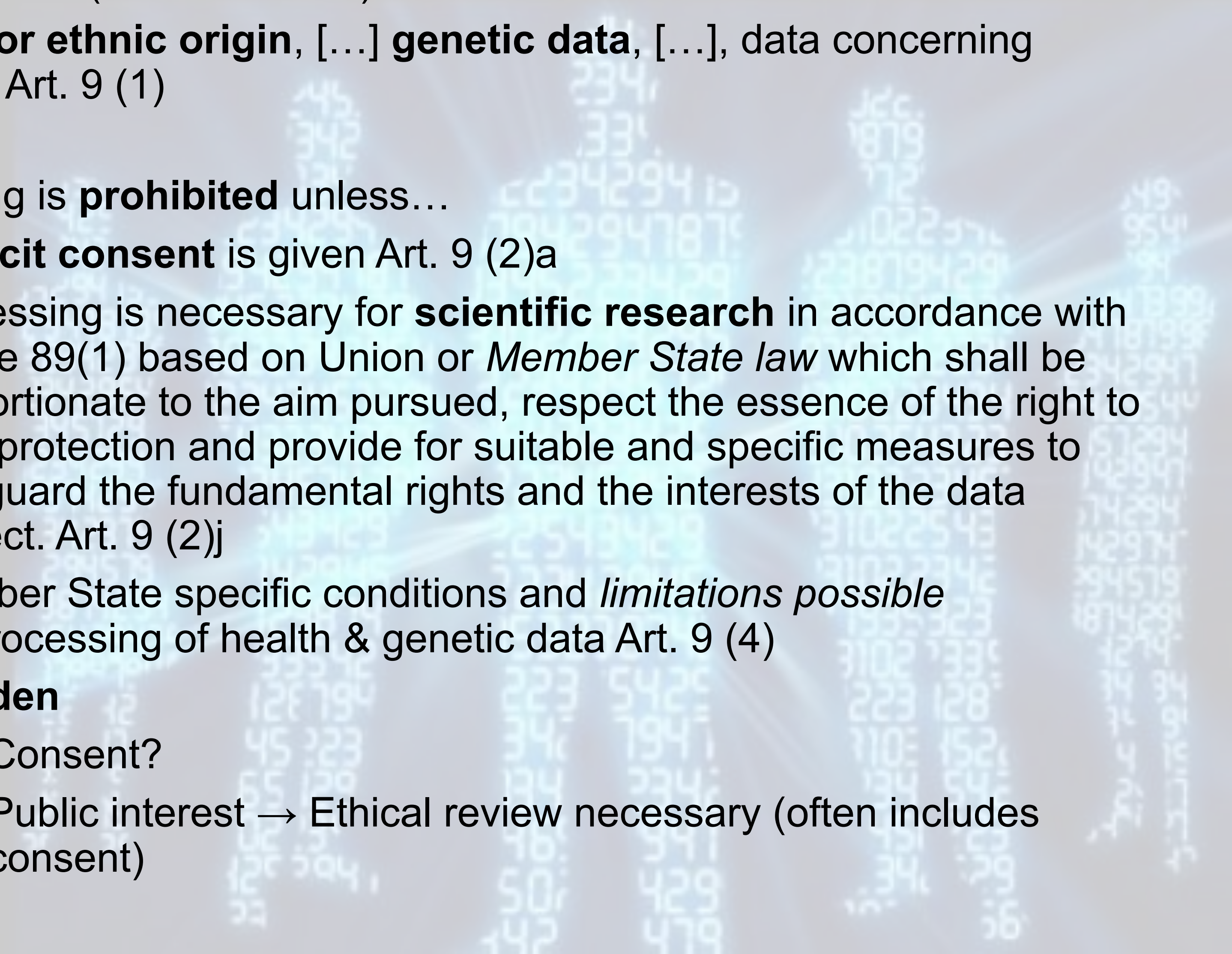
Personal data - Legislation

- GDPR – General Data Protection Regulation (*Dataskyddsförordningen*) + others
- Act concerning the Ethical Review of Research Involving Humans (*Lag om etikprövning av forskning som avser människor*)

GDPR

- All kinds of information that is directly or indirectly referable to a natural person who is alive constitute personal data
- To process personal data:
  - *All processing of personal data must fulfil the **fundamental principles** defined in the Regulation, among them are:*
    - Decide a **purpose** and stick to it
    - Identify the **legal basis** for data processing before it starts
- *Have you defined the **purpose** and **legal basis** for handling personal data in your project?*

GDPR – Sensitive data



- Special categories (*Sensitive data*)
  - … **racial or ethnic origin**, […] **genetic data**, […], data concerning **health** … Art. 9 (1)

  - Processing is **prohibited** unless…
    - **explicit consent** is given Art. 9 (2)a
    - processing is necessary for **scientific research** in accordance with Article 89(1) based on Union or *Member State law* which shall be proportionate to the aim pursued, respect the essence of the right to data protection and provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject. Art. 9 (2)j
    - Member State specific conditions and *limitations possible* for processing of health & genetic data Art. 9 (4)
    - **Sweden**
      - Consent?
      - Public interest → Ethical review necessary (often includes consent)

GDPR – Help

- A **Data Protection Officer** (*dataskyddssombud*)
  - The natural person that is responsible for ensuring that the organization/company adheres to the GDPR
  - Educate
  - Audit
  - Contact point between organization and Data Protection Agency

GU

https://medarbetarportalen.gu.se/projekt-process/aktuella-projekt/dataskyddsforordning

KI

https://ki.se/medarbetare/gdpr-pa-karolinska-institutet

KTH

https://intra.kth.se/anstallning/anstallningsvillkor/att-vara-statligt-an/behandling-av-person/dataskyddsforordningen-gdpr-1.800623

LiU

https://insidan.liu.se/dataskyddsforordningen/anmalan-av-personuppgiftsbehandling?l=sv

LU

https://personuppgifter.blogg.lu.se

SU

https://www.su.se/medarbetare/organisation-styrning/juridik/personuppgifter/dataskyddsf%C3%B6rordningen

UmU

https://www.aurora.umu.se/regler-och-riktlinjer/juridik/personuppgifter/

UU

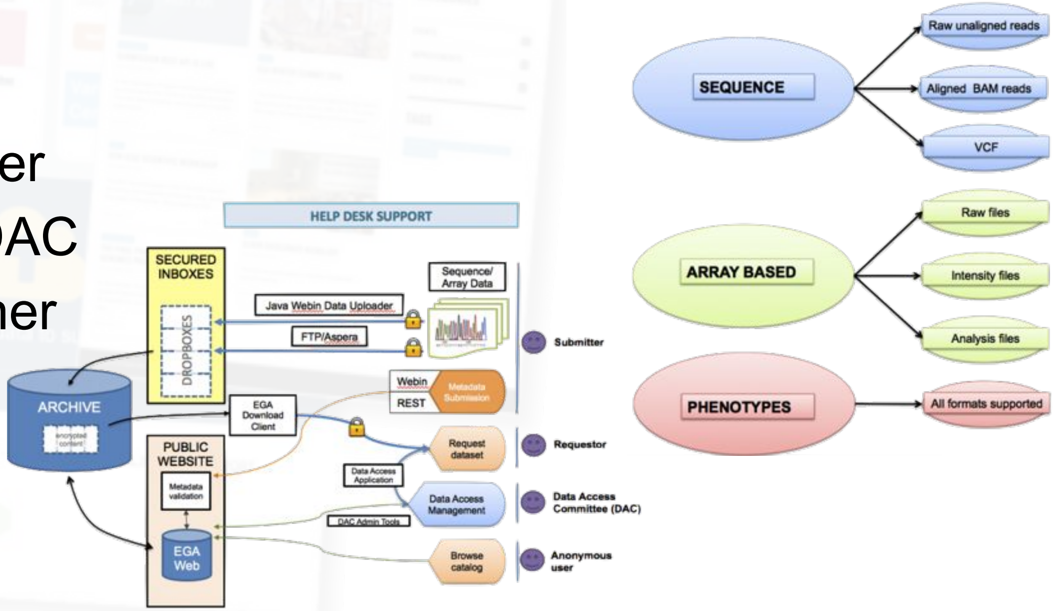
https://mp.uu.se/web/info/stod/dataskyddsforordningen

What about publishing sensitive data?

"As open as possible, as closed as necessary"

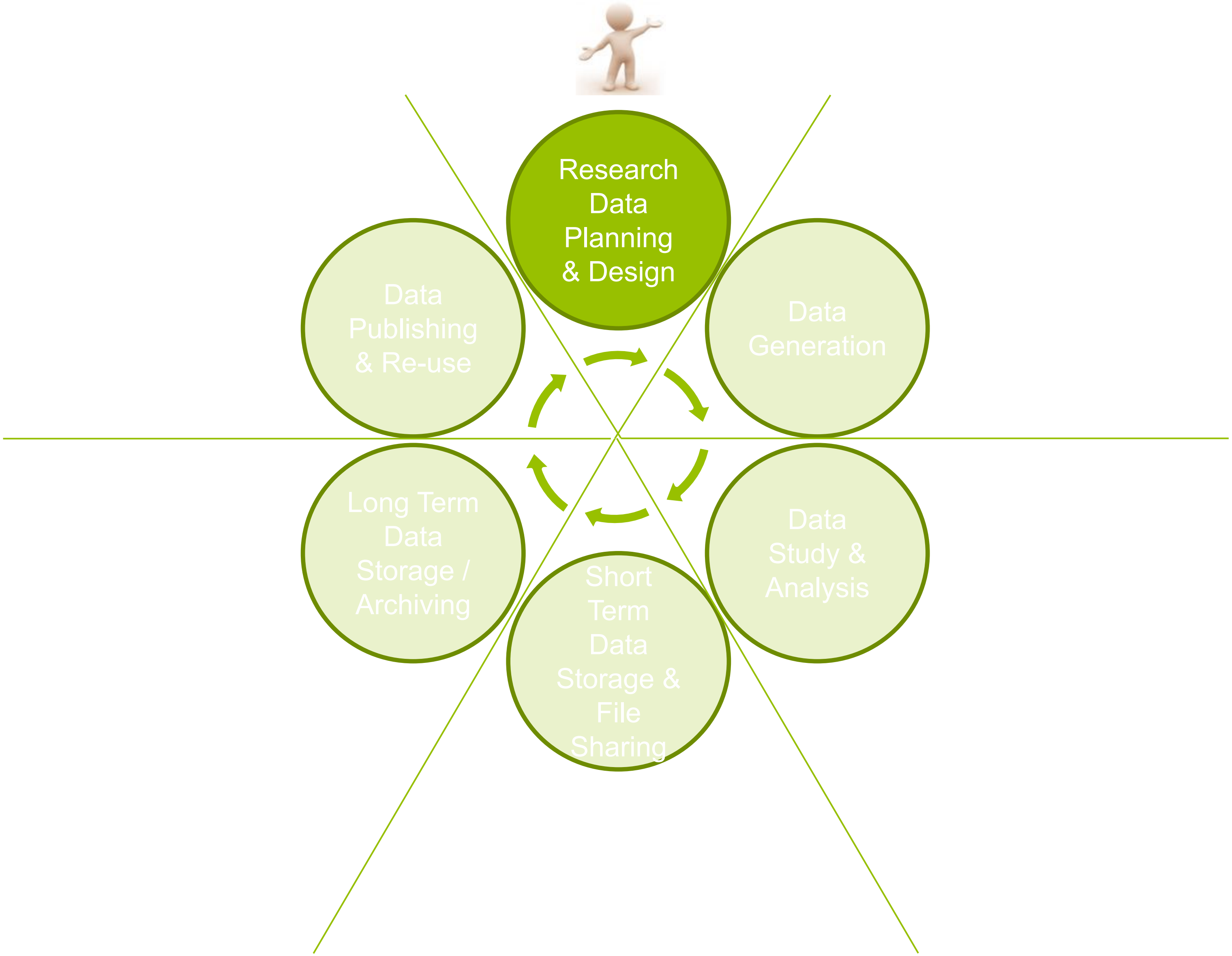
- **EGA** – European Genome-phenome Archive
  - Repository that promotes the distribution and sharing of **genetic and phenotypic data** consented for specific approved uses but **not fully open, public distribution**.
  - All types of sequence and genotype experiments, including case-control, population, and family studies.

- Data Access Agreement
  - Defined by the data owner
- Data Access Committee – DAC
  - Decided by the data owner



When should you start planning for how to manage you data?

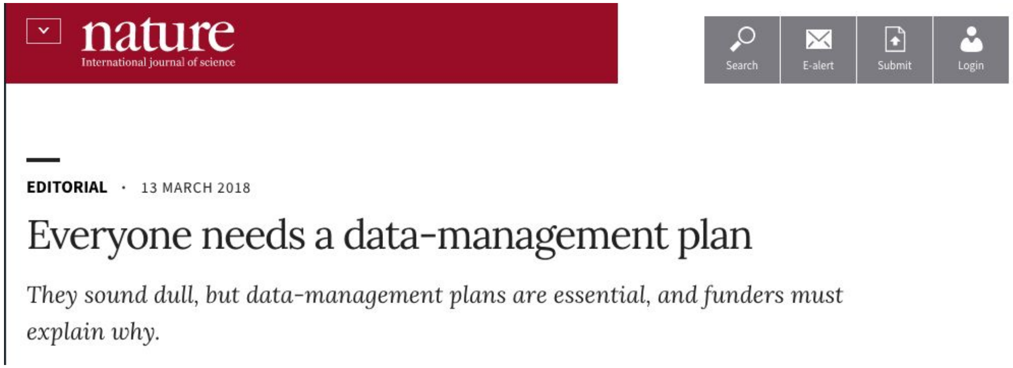
Planning & Design


Research Data Planning & Design
Data Generation
Data Study & Analysis
Short Term Data Storage & File Sharing
Long Term Data Storage / Archiving
Data Publishing & Re-use

Data Management Plans

Will become a standard part of the research funding application process

- **Data collection** - data types and volumes, analysis code
- **Data organization** - folder and file structure, and naming
- **Data documentation** - data and analysis, metadata standards
- **Data storage** - storage/backup/protection & time lines
- **Data policies** - conditions/licences for using data & legal/ethical issues
- **Data sharing** - *When* and *How* will *What* data (and code) be shared
- **Roles and responsibilities** - who's responsible for what & is competence available
- **Budget** - People & Hardware/Software



EDITORIAL · 13 MARCH 2018

Everyone needs a data-management plan

They sound dull, but data-management plans are essential, and funders must explain why.

Planning ahead

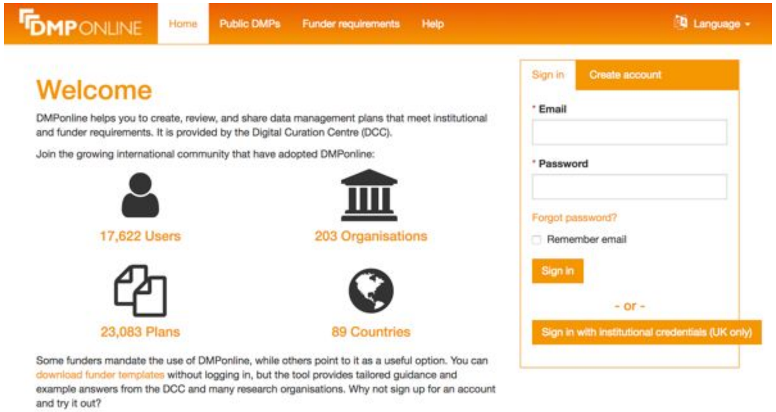
💡 *Consider structuring metadata in the format needed by the repository already at planning stage*

NBIS
NATIONAL BIOINFORMATICS
INFRASTRUCTURE SWEDEN

DMP tools

SciLifeLab

DMPonline



Welcome

DMPonline helps you to create, review, and share data management plans that meet institutional and funder requirements. It is provided by the Digital Curation Centre (DCC).

Join the growing international community that have adopted DMPonline:

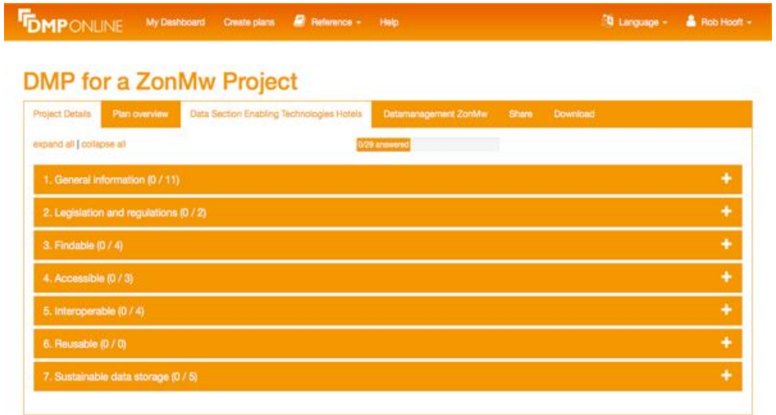
17,622 Users

203 Organisations

23,083 Plans

89 Countries

Some funders mandate the use of DMPonline, while others point to it as a useful option. You can download funder templates without logging in, but the tool provides tailored guidance and example answers from the DCC and many research organisations. Why not sign up for an account and try it out?



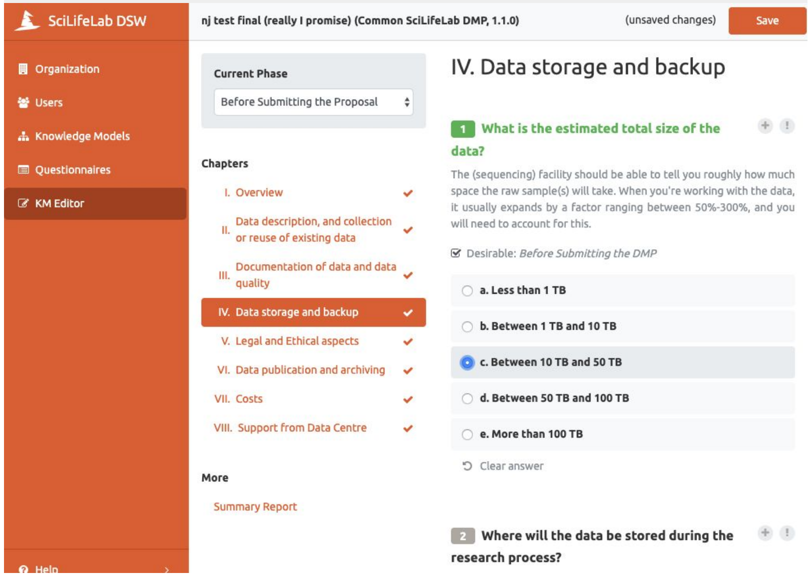
DMP for a ZonMw Project

1. General information (0 / 11) +
2. Legislation and regulations (0 / 2) +
3. Findable (0 / 4) +
4. Accessible (0 / 3) +
5. Interoperable (0 / 4) +
6. Reusable (0 / 0) +
7. Sustainable data storage (0 / 5) +

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

ELIXIR Data Stewardship Wizard

Smart Data Management Plans for FAIR Open Science
For Serious Researchers and Data Stewards



nj test final (really I promise) (Common SciLifeLab DMP, 1.1.0)

Current Phase: Before Submitting the Proposal

Chapters

- I. Overview ✓
- II. Data description, and collection or reuse of existing data ✓
- III. Documentation of data and data quality ✓
- IV. Data storage and backup ✓
- V. Legal and Ethical aspects ✓
- VI. Data publication and archiving ✓
- VII. Costs ✓
- VIII. Support from Data Centre ✓

More

Summary Report

IV. Data storage and backup

1 What is the estimated total size of the data?

The (sequencing) facility should be able to tell you roughly how much space the raw sample(s) will take. When you're working with the data, it usually expands by a factor ranging between 50%-300%, and you will need to account for this.

☑ Desirable: *Before Submitting the DMP*

○ a. Less than 1 TB
○ b. Between 1 TB and 10 TB
● c. Between 10 TB and 50 TB
○ d. Between 50 TB and 100 TB
○ e. More than 100 TB

Clear answer

2 Where will the data be stored during the research process?

https://ds-wizard.org/

https://dsw.scilifelab.se

NBIS Data Management support

- Project planning
  - Metadata
  - File formats
  - Licensing
  - *Data Management Plans*

- Data analysis

- Data publication and submission
  - Support submissions to public repositories
  - Metadata
  - DOIs to dataset (if needed)

Take home messages

- Consider doing a Data Management Plan for your project
  - How do you ensure that your research output is FAIR?
- Plan for submitting ”raw data” to public repositories as early as possible
- Organize project metadata from the start
  - In ways that makes it easy to submit to public repositories
  - Use available standards
- Pick a thought-through file and folder structure organization for your computational analyses
- Strive for reproducibility
  - Data & Code
- Be aware that there are legal aspects to processing human data
- *Ask for help if you need it!*

Source Acknowledgements

- Research Data Management, EUDAT - http://hdl.handle.net/11304/79db27e2-c12a-11e5-9bb4-2b0aad496318
- Noble WS (2009) A Quick Guide to Organizing Computational Biology Projects. PLoS Comput Biol 5(7): e1000424. doi:10.1371/journal.pcbi.1000424
- Reproducible research
    - Reproducible Science Curriculum – https://github.com/Reproducible-Science-Curriculum/rr-init
    - Leif Väremo & Rasmus Ågren
        - https://bitbucket.org/scilifelab-lts/reproducible_research_example/src
        - https://nbis-reproducible-research.readthedocs.io/en/course_1803
- GDPR
    - Datainspektionen – https://www.datainspektionen.se/lagar--regler/dataskyddsforordningen/
- … and probably others I have forgotten