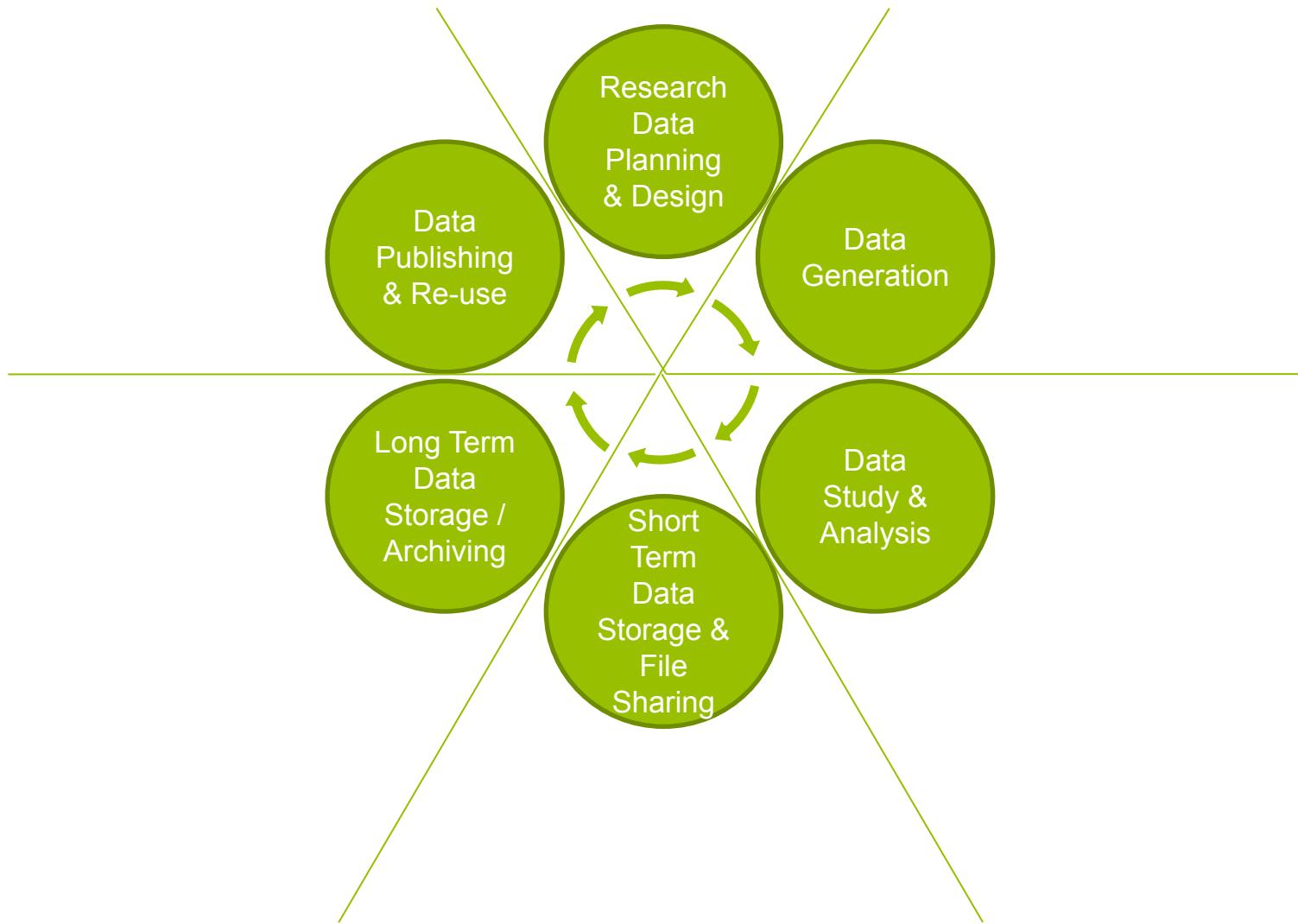

Data Management

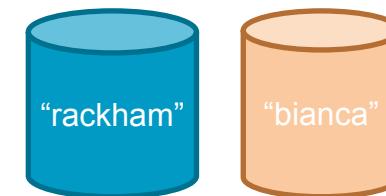
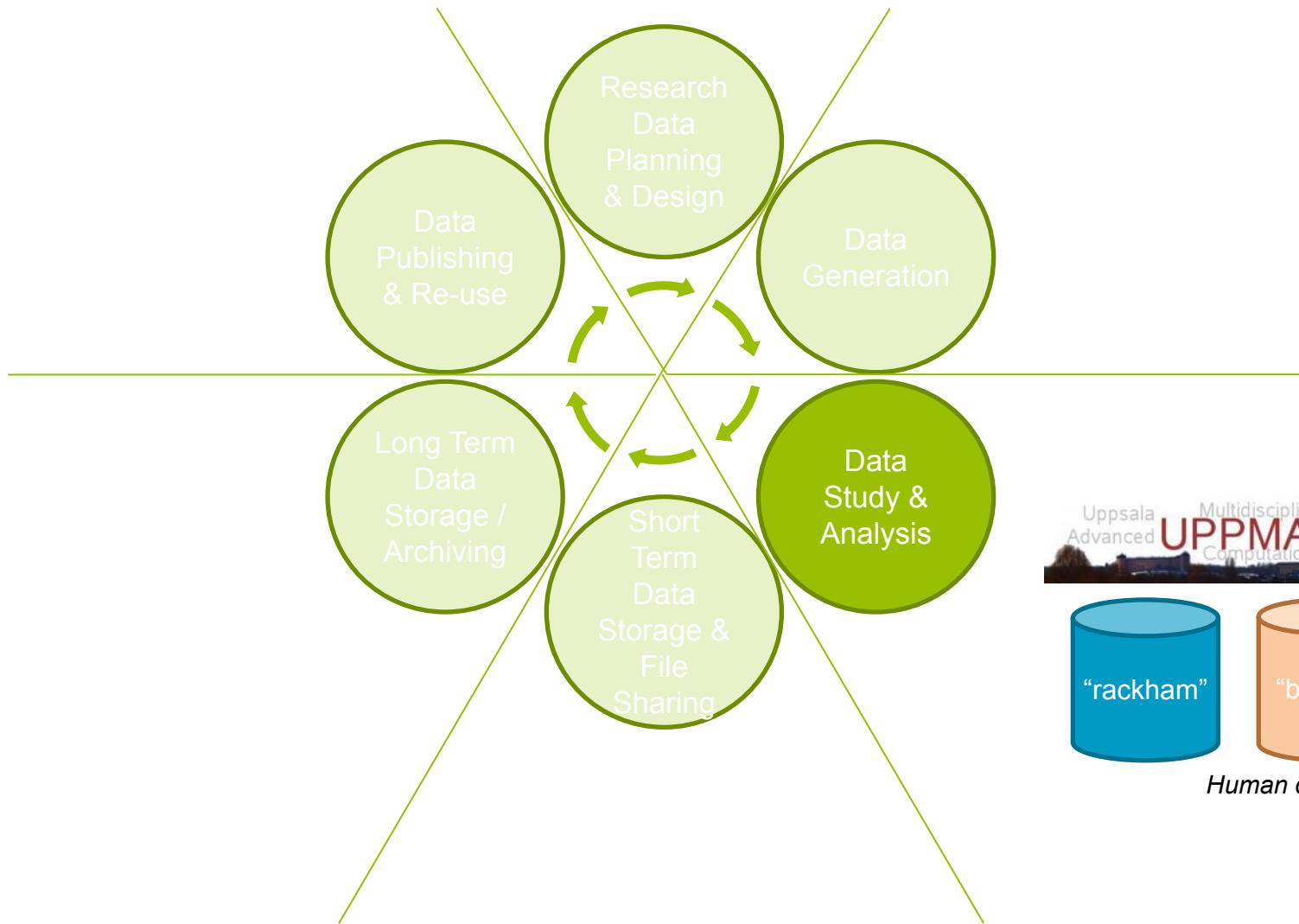
Niclas Jareborg & Yvonne Kallberg, NBIS
niclas.jareborg@nbis.se

2019-09-18

Bioinformatics for Principal Investigators course

**Do you think data
management is
important?**





Human derived data

How do you know how an old result was generated?

- 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

Follow

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



Poor organizational choices lead to significantly slower research progress

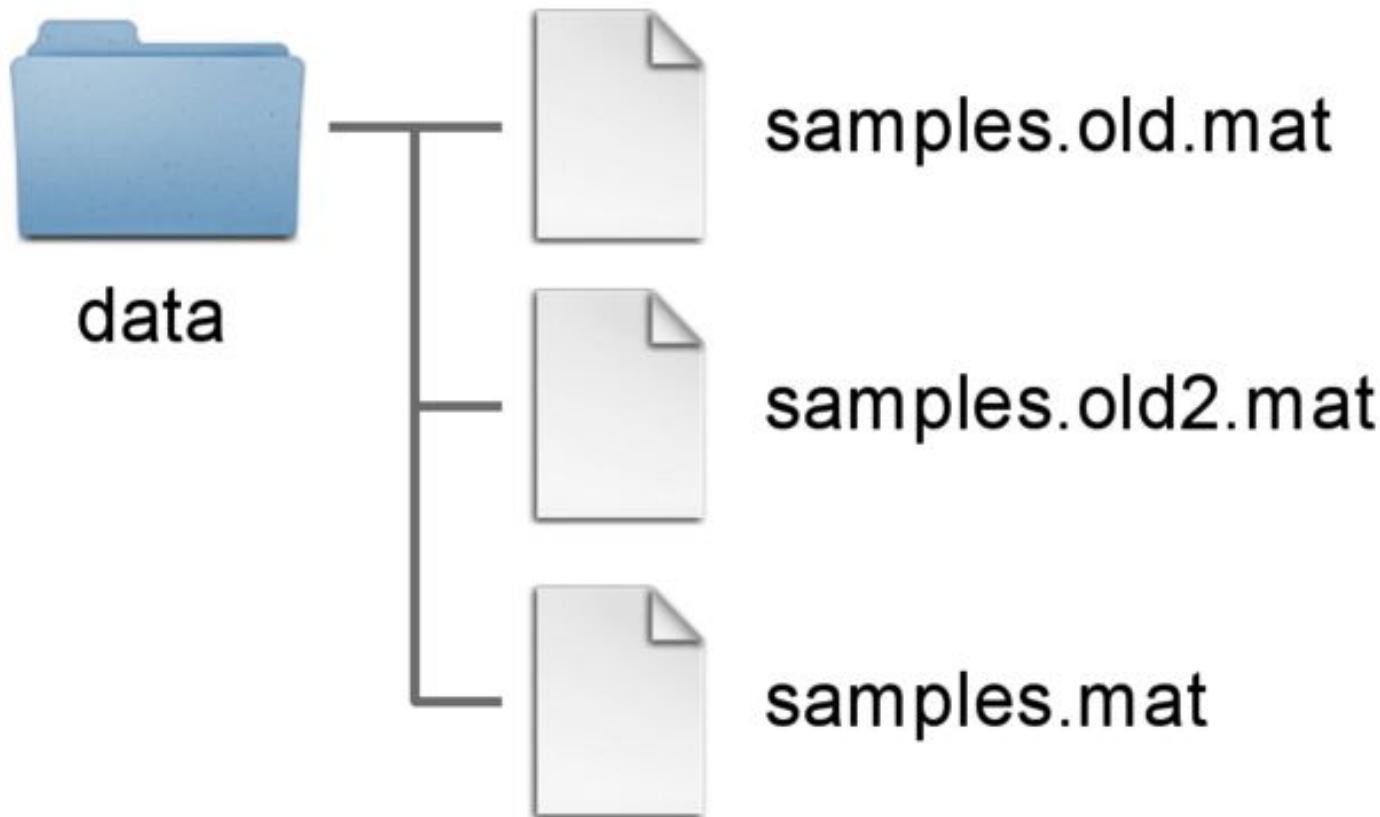
“Your primary collaborator is yourself six months from now, and your past self doesn’t answer e-mails.”

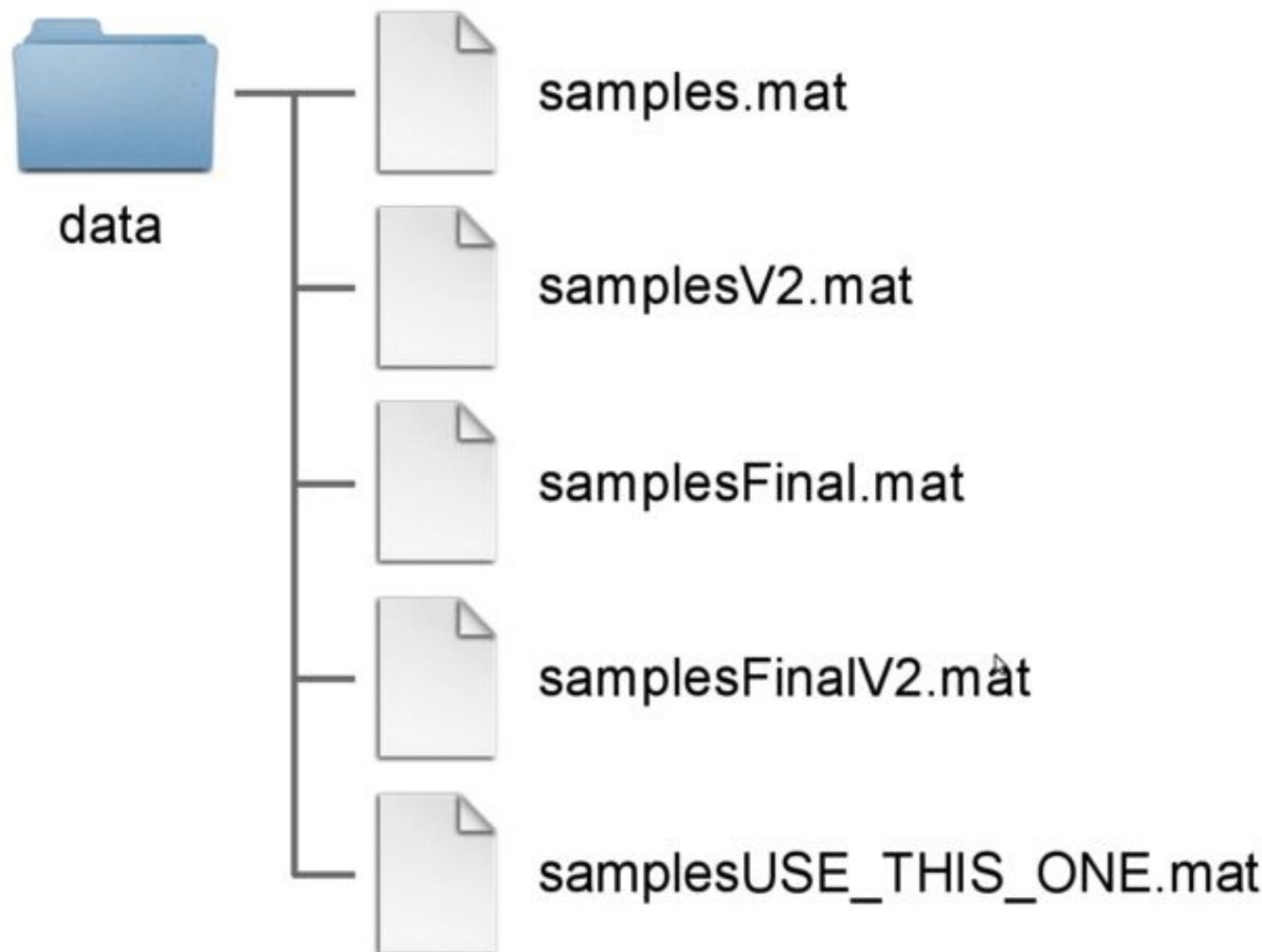


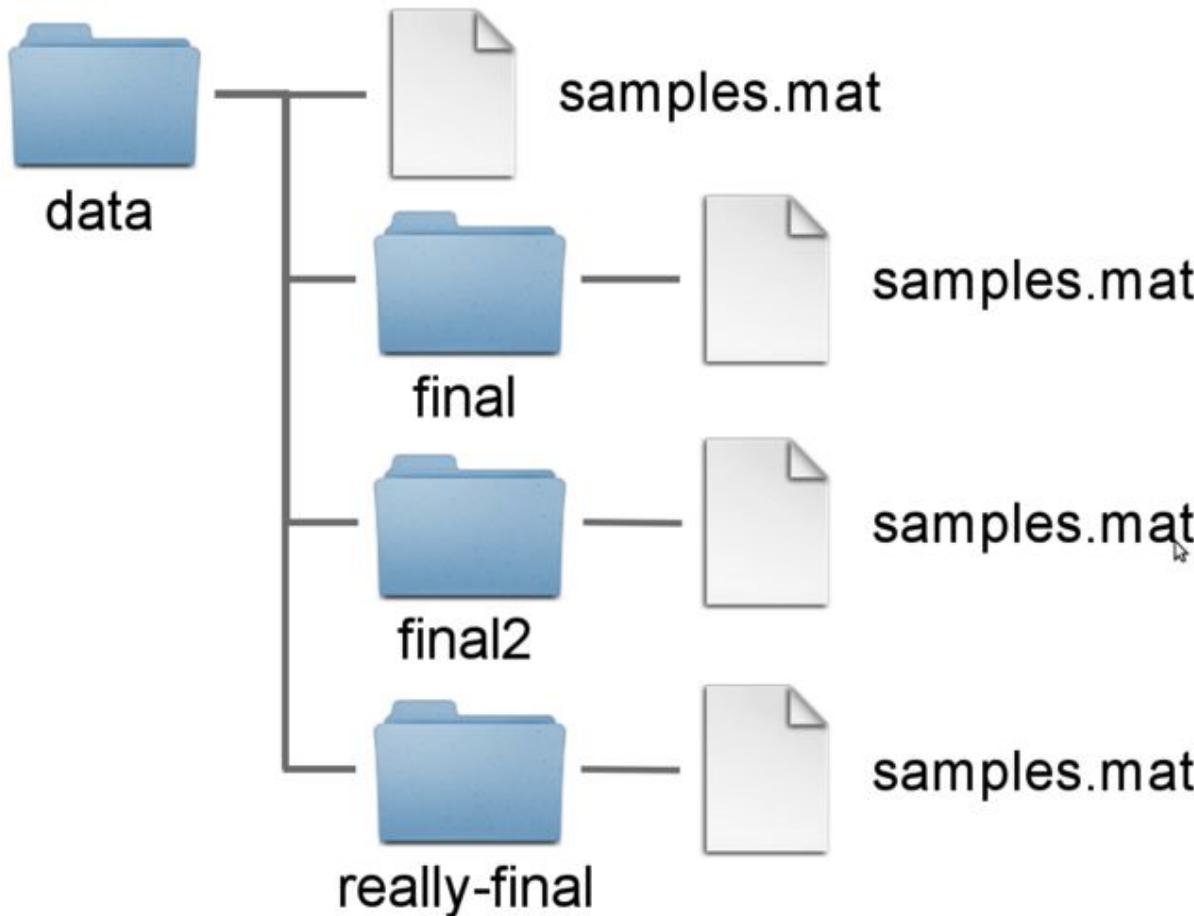
data

samples.mat









A possible solution



- Directory Structure
 - 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.**
 - 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 a **version control system** (at least for code) – e.g. **git**
- Use **non-proprietary formats** for files– .csv rather than .xlsx
- Etc...

▶  code	all code needed to go from input files to final results
▼  data	raw and primary data, essentially all input files, never edit!
 README.txt	
▶  meta	
▶  raw_external	
▶  raw_internal	
▶  doc	documentation for the study
▶  intermediate	output files from different analysis steps, <i>can be deleted</i>
▶  logs	logs from the different analysis steps
▶  notebooks	
▼  results	output from workflows and analyses
 README.txt	
▶  figures	
▶  reports	
▶  tables	
▶  scratch	temporary files that can be safely <i>deleted or lost</i>

- Directory Structure
 - 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.**
 - 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 a **version control system** (at least for code) – e.g. **git**
- Use **non-proprietary formats** for files– .csv rather than .xlsx
- Etc...

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

NO

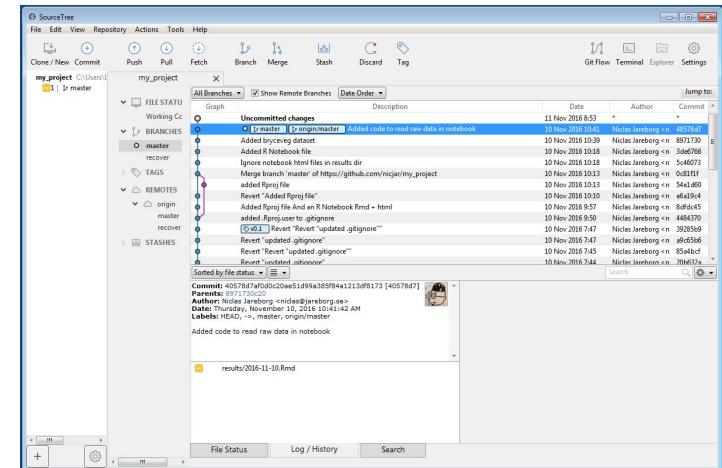
myabstract.docx
Joe's Filenames Use Spaces and Punctuation.xlsx
figure 1.png
fig 2.png
JW7d^(2sl@deletethisandyourcareerisoverWx2*.txt

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

- Directory Structure
 - 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.**
 - 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 a **version control system** (at least for code) – e.g. **git**
- Use **non-proprietary formats** for files– .csv rather than .xlsx
- Etc...

- 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



- Directory Structure
 - 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.**
 - 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 a **version control system** (at least for code) – e.g. **git**
- Use **non-proprietary formats** for files– .csv rather than .xlsx
- Etc...

- 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, ...



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

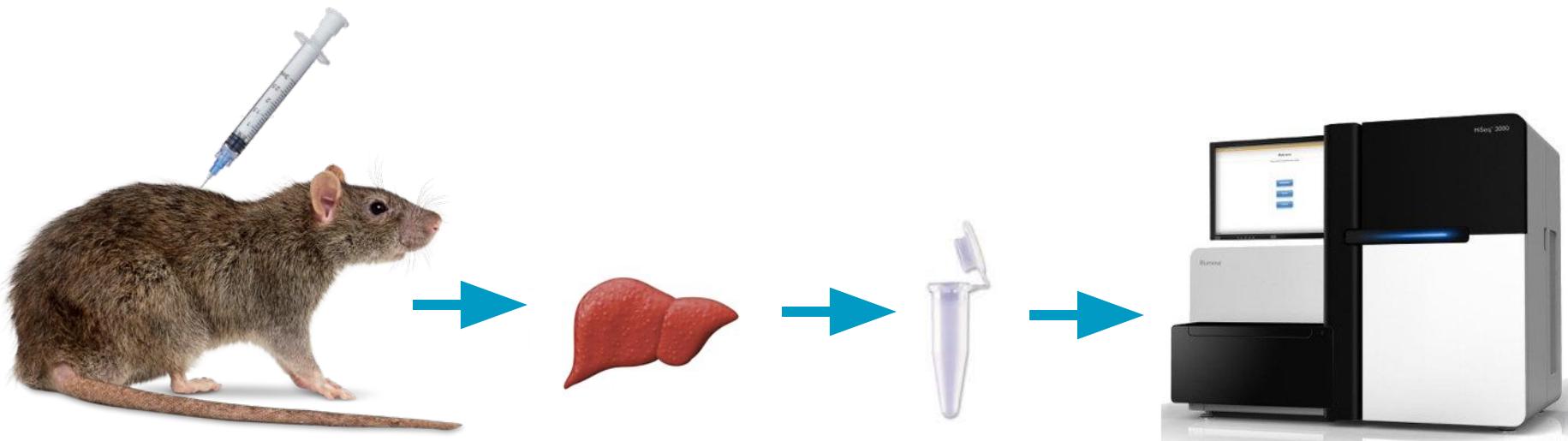
DO NOT

- 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

 be FAIR  be Open

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



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

Human Phenotype Ontology

Summary Classes Properties Notes Mappings Widgets

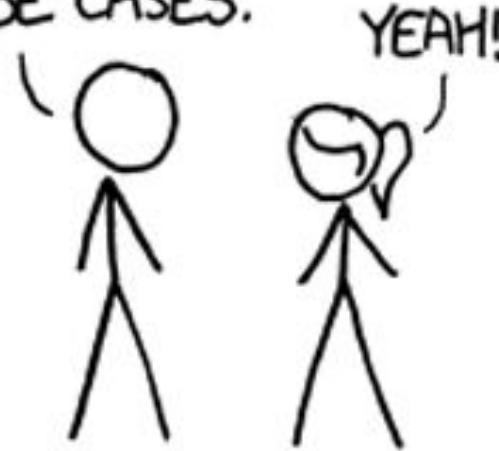
Details	Visualization	Notes (0)	Class Mappings (21)
Preferred Name	Acute myeloid leukemia		
Synonyms	Acute myeloblastic leukemia Acute myelogenous leukemia Acute myelocytic leukemia		
Definitions	A form of leukemia characterized by overproduction of an early myeloid cell.		
ID	http://purl.obolibrary.org/obo/HP_0004808		
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		

Jump To: 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 Malignant eosinophil proliferation Multiple myeloma Myelodysplasia Plasmacytoma Abnormality of connective tissue Abnormality of head or neck Abnormality of limbs Abnormality of metabolism/homeostasis

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

SITUATION:
THERE ARE
14 COMPETING
STANDARDS.

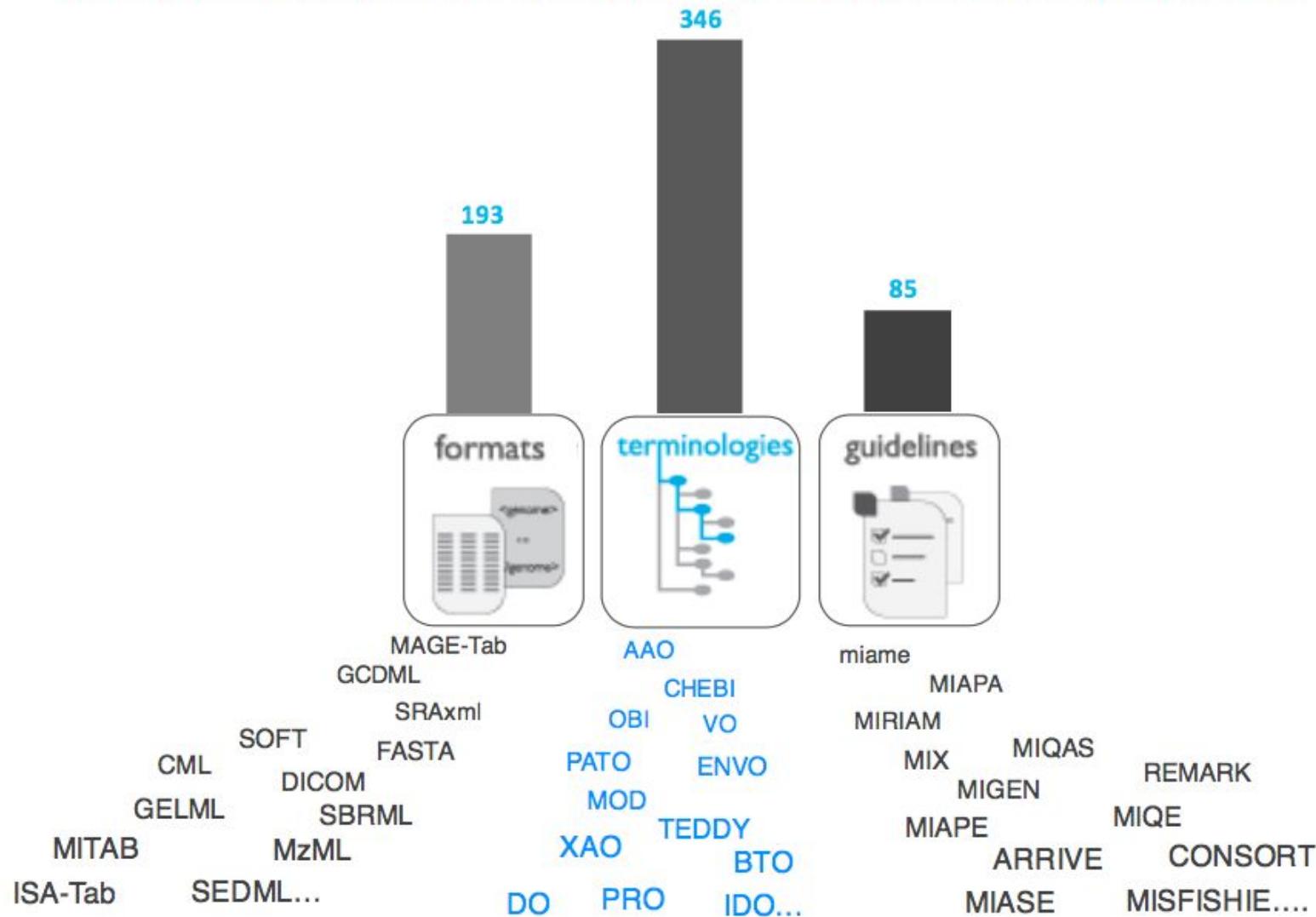
14?! RIDICULOUS!
WE NEED TO DEVELOP
ONE UNIVERSAL STANDARD
THAT COVERS EVERYONE'S
USE CASES.



SOON:

SITUATION:
THERE ARE
15 COMPETING
STANDARDS.

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



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

Standards Databases Policies Collections/Recommendations

Advanced Search 
Fine grained control over your search.

Search Wizard 
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](#)

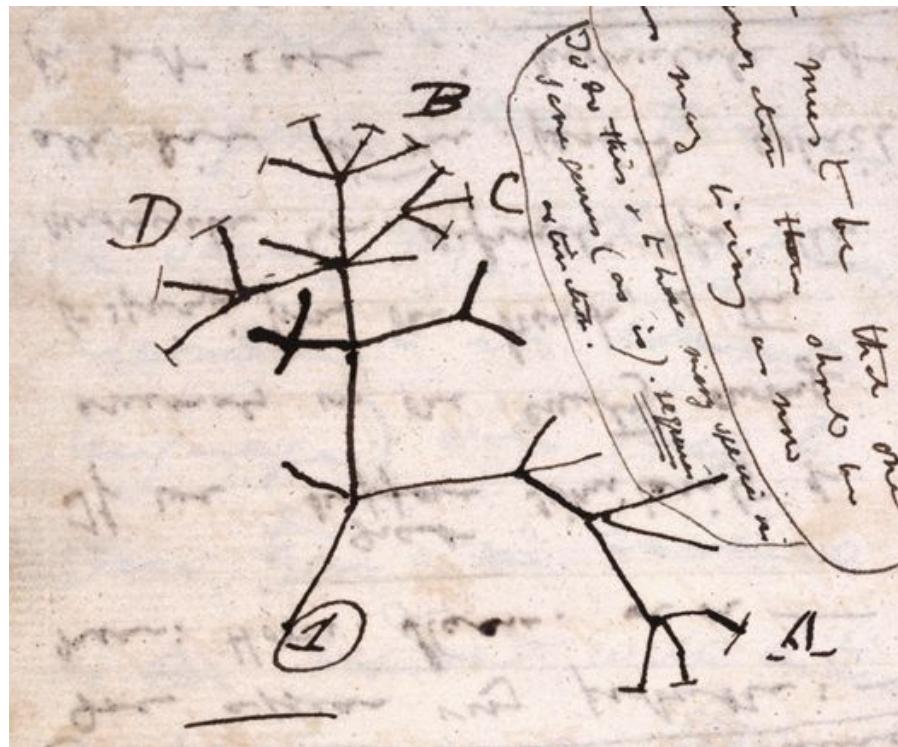


97 Policies

Funder	22
Journal	68
Society	3

[View all](#)

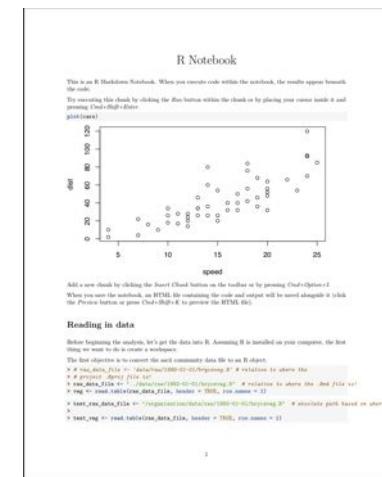
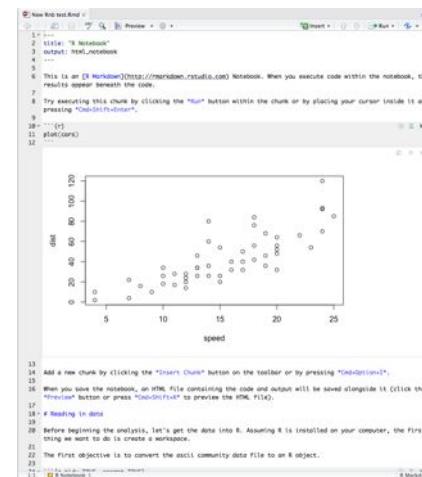
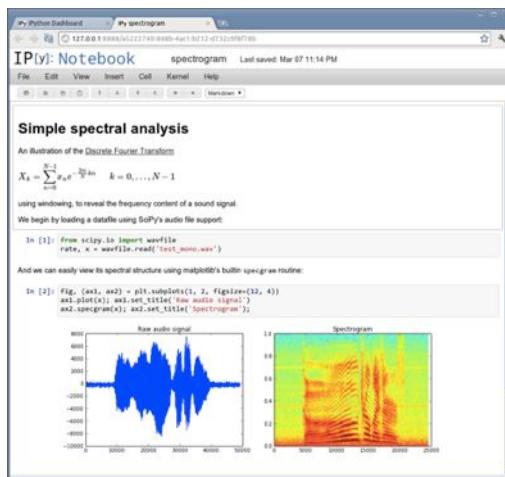
- Why?
 - You have to understand what you have done
 - **Others should be able to reproduce what you have done**



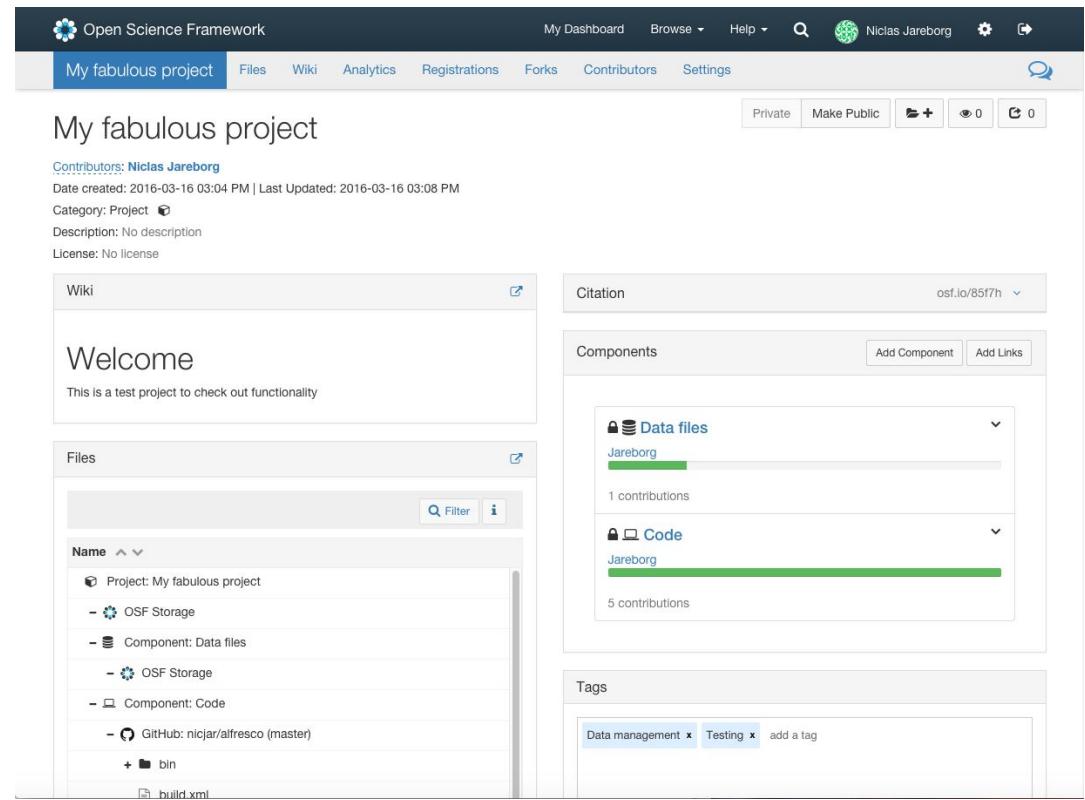
- Put in *separate* directory (e.g. *results, documentation*)
- *Dated* entries
- Entries relatively verbose
- Link to *data* and *code* (including versions)
 - Point to commands run and results generated
- Embedded images or tables showing results of analysis done
- Observations, Conclusions, and *ideas* 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. [jupyter](#), [R Notebooks](#) in RStudio
 - Plain text - work well with version control (Markdown)
 - Embed and execute code
 - Convert to other output formats
 - html, pdf, word



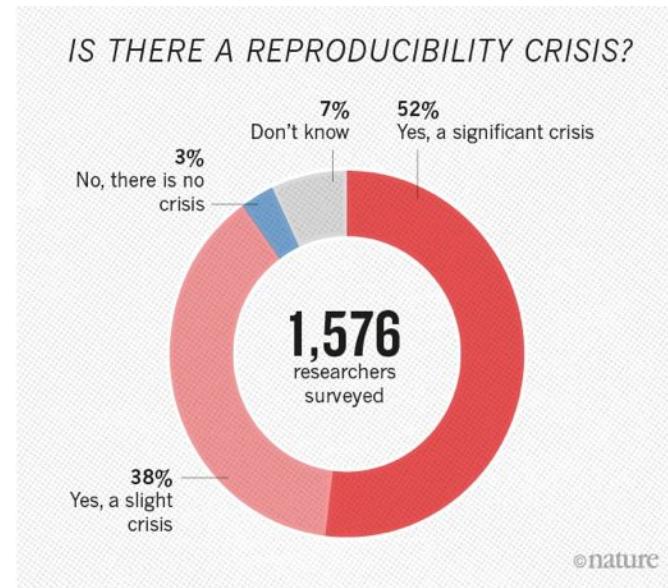
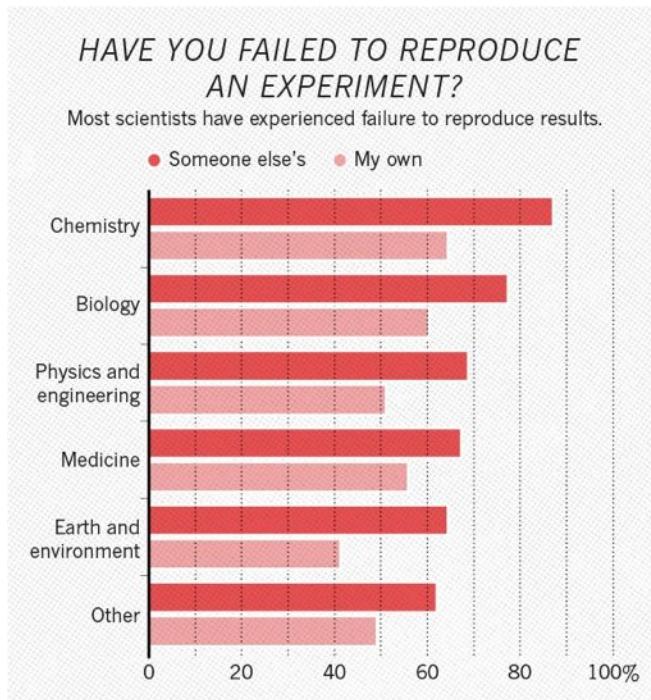
- Open Science Framework – <http://osf.io>
 - Organize research project documentation and outputs
 - Control access for collaboration
 - 3rd party integrations
 - Google Drive
 - Dropbox
 - GitHub
 - External links
 - Etc
 - Persistent identifiers
 - Publish article preprints



The screenshot shows the OSF project dashboard for "My fabulous project".

- Header:** Open Science Framework, My Dashboard, Browse, Help, Niclas Jareborg, Settings.
- Project Information:**
 - My fabulous project
 - Contributors: Niclas Jareborg
 - Date created: 2016-03-16 03:04 PM | Last Updated: 2016-03-16 03:08 PM
 - Category: Project
 - Description: No description
 - License: No license
- Wiki:** Welcome, This is a test project to check out functionality.
- Files:**
 - Project: My fabulous project
 - OSF Storage
 - Component: Data files
 - OSF Storage
 - Component: Code
 - Github: nicjar/alfresco (master)
 - bin
 - build.xml
- Citation:** osf.io/85f7h
- Components:**
 - Data files: Jareborg (1 contributions)
 - Code: Jareborg (5 contributions)
- Tags:** Data management, Testing, add a tag.

A reproducibility crisis

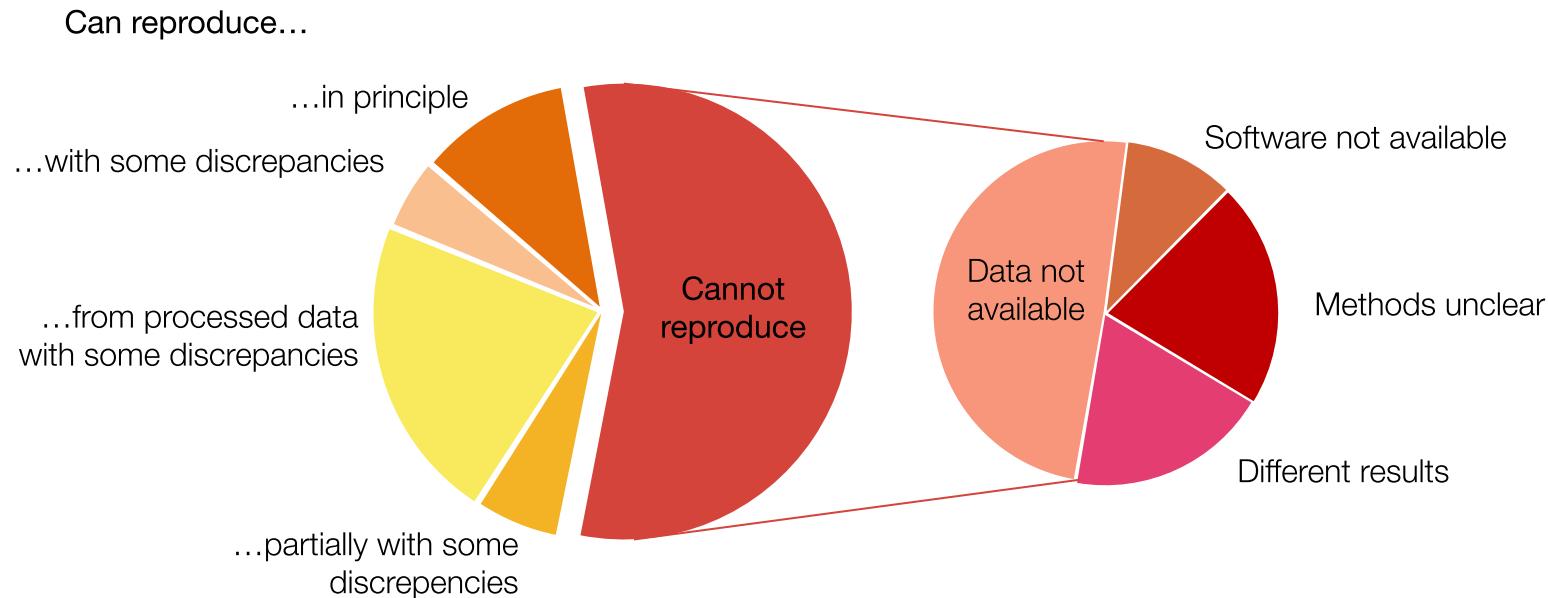


[1] "1,500 scientists lift the lid on reproducibility". Nature. 533: 452–454

[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:



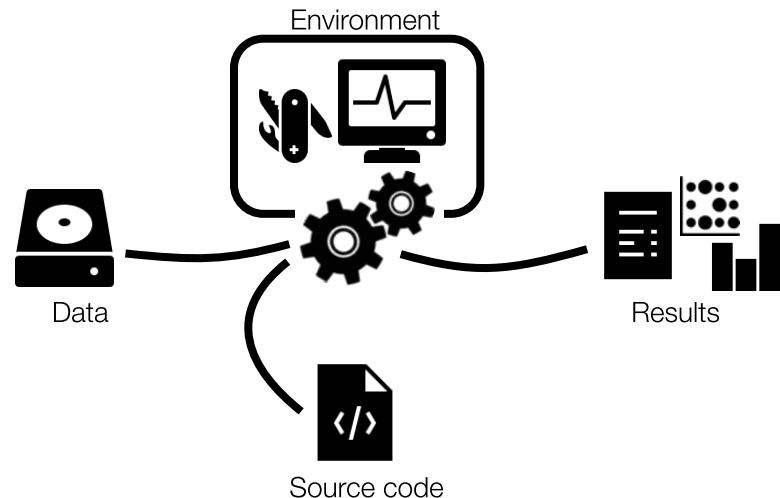
Summary of the efforts to replicate the published analyses.

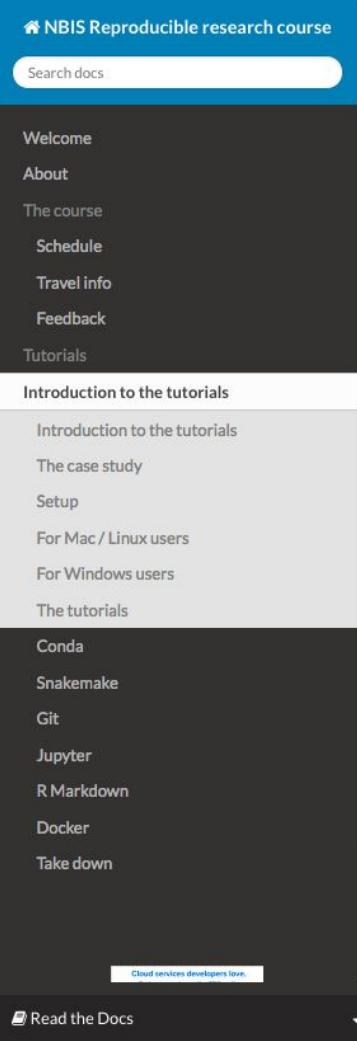
Adopted from: Ioannidis et al. Repeatability of published microarray gene expression analyses.
Nature Genetics 41 (2009) doi:10.1038/ng.295

What do we mean by reproducible research?

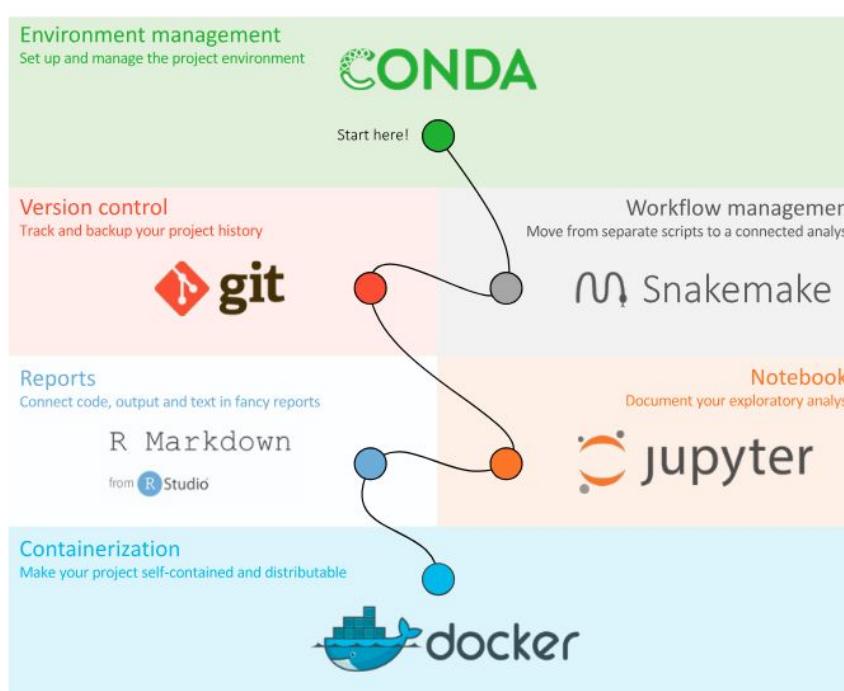
		Data	
		Same	Different
Code	Same	Reproducible	Replicable
	Different	Robust	Generalizable

All parts of a bioinformatics analysis have to be reproducible:





The screenshot shows the documentation for the NBIS Reproducible research course. It includes a sidebar with links to Welcome, About, The course, Schedule, Travel info, Feedback, and Tutorials. Under Tutorials, there are sections for Introduction to the tutorials (with links to Introduction to the tutorials, The case study, Setup, For Mac / Linux users, For Windows users, and The tutorials), Conda, Snakemake, Git, Jupyter, R Markdown, Docker, and Take down. A footer at the bottom right says "Cloud services developers love" followed by a "Read the Docs" button.

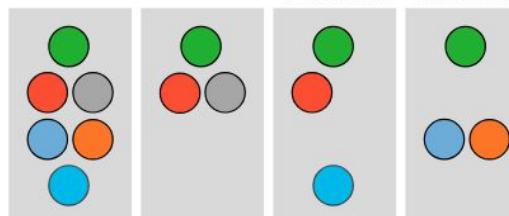


The diagram illustrates the interconnected nature of various reproducible research tools:

- Environment management:** Set up and manage the project environment. This is represented by the **CONDA** logo.
- Version control:** Track and backup your project history. This is represented by the **git** logo.
- Reports:** Connect code, output and text in fancy reports. This is represented by the **R Markdown** logo, with a note "from R Studio".
- Containerization:** Make your project self-contained and distributable. This is represented by the **Docker** logo.
- Workflow management:** Move from separate scripts to a connected analysis. This is represented by the **Snakemake** logo.
- Notebooks:** Document your exploratory analysis. This is represented by the **Jupyter** logo.

A central "Start here!" button is connected by lines to each of the tool logos.

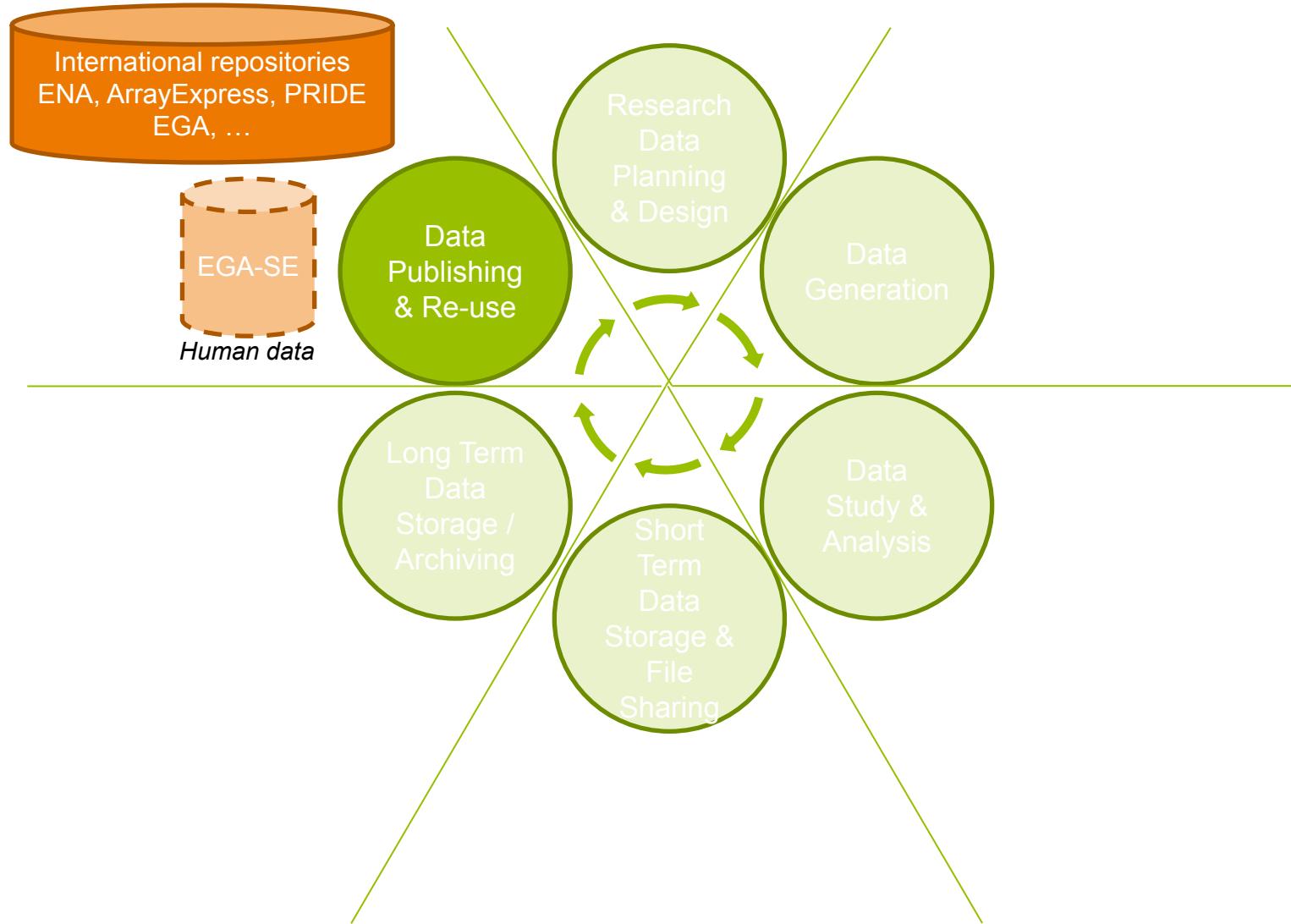
Do it all! Workflow Reproducible environment Interactive notebooks



The icon grid shows combinations of tools:

- Do it all!: CONDA + git + R Markdown + Docker
- Workflow: CONDA + Snakemake
- Reproducible environment: CONDA + Docker
- Interactive notebooks: Jupyter

<https://nbis-reproducible-research.readthedocs.io/en/latest/>





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

Reflections?

- 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



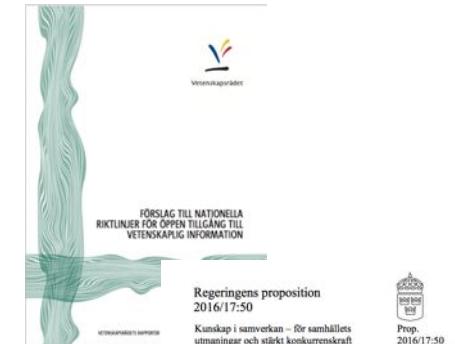
- Strong international movement towards Open Access (OA)
 - European Commission recommended the member states to establish national guidelines for OA
 - Swedish Research Council (VR) submitted proposal to the government Jan 2015
 - Research bill 2017–2020 – 28 Nov 2016
 - *The aim of the government is 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



Propositionens huvudsakliga innehåll

I propositionen presenteras regelgångar om av- och förkändningsrättskrivning i en teknisk projekt, med särskilt fokus på semester 2012. Syftet är att ge veta om den huvudsakliga konkretiseringen och att utvärdera dess tillämpningsvärde.

En utveckling är i vissa av den forskning som förekommer i litteraturen om teknisk projekt. Primitivare studier är kända i både mkt. hushåll, diskurs, etikett och teknik. I denna artikel har dock författaren utvärderat utvecklingen i teknisk projekt.

Ett annat område som är viktigt att finnas för forskningen och också varit omdebatterat är personer som deltar i teknisk projekt.

- 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>



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

Mark D. Wilkinson et al.*

Received: 10 December 2015
 Accepted: 12 February 2016
 Published: 15 March 2016

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

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

G20 HANGZHOU SUMMIT

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

HANGZHOU, CHINA 4-5 SEPTE



Findable:

- 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 it describes;
- F4. (meta)data are registered or indexed in a **searchable resource**;

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;

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;

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**;

<https://www.nature.com/articles/sdata201618>

1. Start with a management plan
2. Describe and document your data for humans and machines
3. Preserve your data

F1000



Your go-to guide to making your data **Findable, Accessible, Interoperable, and Reusable (FAIR)**

So that you and others can get the most out of your data, it is important that you adhere to the [FAIR principles](#) to ensure your data are **Findable, Accessible, Interoperable, and Reusable** – whilst making your data openly available where it is safe to do so. This is no small task, so here are some ideas to help you get started:

1



Start with a management plan

An output management plan (OMP) is a useful starting point for collecting or creating data, software, research materials, and intellectual property. Creating an OMP before you begin your research, and updating it throughout the research cycle, will help ensure that your outputs are as open and FAIR as possible when your project is complete.

Some funders require grant-holders to produce a plan as part of their application for funding, and/or after funding has been secured.

You should consider:

- What outputs you will be creating or collecting, and how these will be documented
- What ethical or legal requirements, if any, apply to the outputs
- How you will organise, store, secure, and share the outputs
- What resources are required and who is responsible

2



Describe and document your data for humans and machines

Describing how your data were created, how they are structured, and what they mean is crucial to making your data **reusable**. As a general rule, someone who is not familiar with your data should be able to understand what it is about using only the metadata and documentation provided.

Good metadata is clearly associated with the dataset it describes and is available in a machine-readable format, such as text or RDF. Depending on your field of study, there may already be standards in place that will help guide how your data and metadata should be structured, formatted, and annotated.

F1000



https://f1000.com/resources/FAIR_Open_Guide.pdf

- **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

[More later...](#)

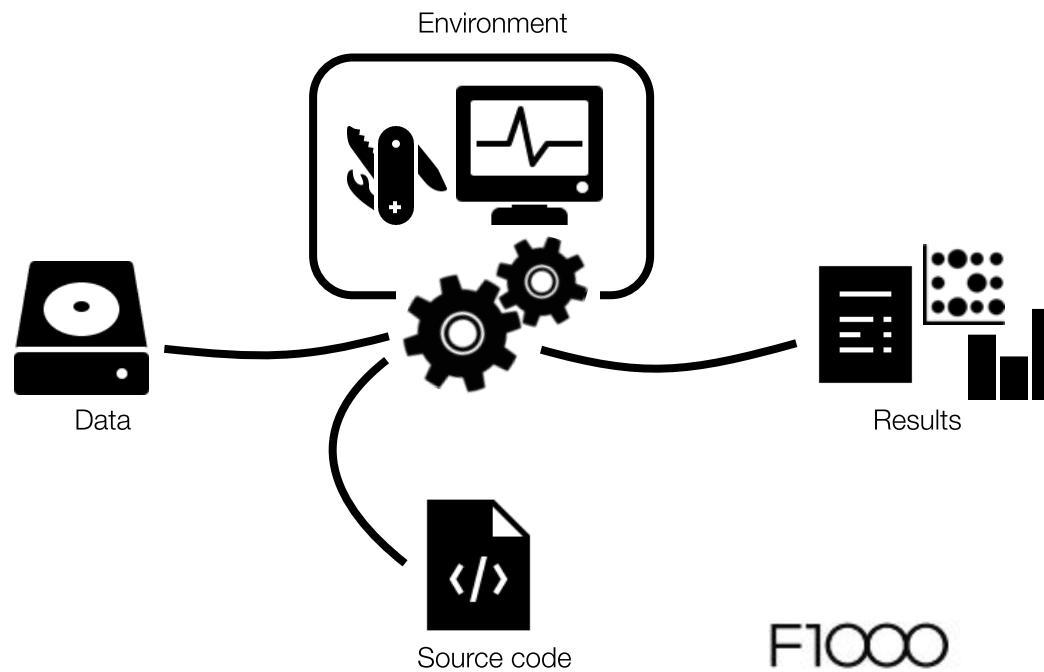
F1000

 beFAIR  beOpen

2. Describe and document

Reusability

As a general rule, someone who is not familiar with your data should be able to understand what it is about using only the metadata and documentation provided.



F1000

beFAIR beOpen

3. Preserve your data

*Data preservation helps ensure that your data will be **accessible** and **reusable** in the future.*

Best practices include:

- **Backing up** data files regularly
- Storing master copies of data files in **open formats**
- **Validating** preserved data files regularly
- Using **more than one form of storage** for data files
- Appropriately **securing data physically**, and/or on any network or computer on which they are held

Data



CC0

[Creative Commons Zero](#) is ideal for openly sharing data – it has no restrictions on Reuse whatsoever. CC0 datasets are widely accepted and expected in science.

Other CC licenses

[CC-BY](#) – Prevents others from applying legal restrictions beyond the terms of the license to the licensed dataset.

[CC BY-SA](#) – Requires outputs derived from licensed dataset to also be licensed as CC BY-SA.

[CC BY-NC](#) – Prevents the licensed data from being used for commercial purpose.

[CC BY-ND](#) – Prevents the licensed data from being modified.

[CC BY-NC-ND](#) – Prevents the licensed data from being used for commercial purposes or modified.

[CC BY-NC-SA](#) – Prevents the licensed data from being used for commercial purposes, and requires outputs derived from licensed dataset to also be licensed as CC BY-SA.

Code

To allow your code to be freely used, modified, and shared by others it's recommended to use a licence approved by the [Open Source Initiative](#), such as

- [MIT](#)
- [GNU General Public License](#)
- [Apache License 2.0](#)

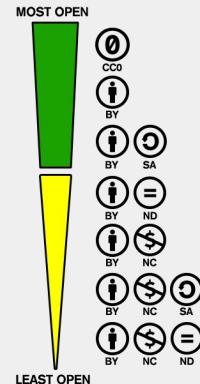
→ [Choose an Open Source License](#)

→ [CC License Chooser](#)

→ [How to License Research Data](#)

Note!

- NC, ND and SA licenses have implications for reuse and *interoperability*
- Be aware of any *licensing restrictions* where your dataset contains *data derived from a 3rd party*.



- *Research Data Publishing is a cornerstone of Open Access*
- 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



- 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

The screenshot shows the ORCID profile page for Niclas Jareborg. At the top, there's a navigation bar with tabs for 'FOR RESEARCHERS' (highlighted in green), 'FOR ORGANIZATIONS', 'ABOUT', 'HELP', and 'SIGN IN'. Below the navigation bar, there are links for 'SIGN IN', 'REGISTER FOR AN ORCID ID', and 'LEARN MORE'. A statistic at the top right says '2,035,272 ORCID IDs and counting. See more...'. The main content area displays Niclas Jareborg's profile information, including his ORCID ID (0000-0002-4520-044X) and a link to his ORCID profile. It also lists his education and employment history.

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

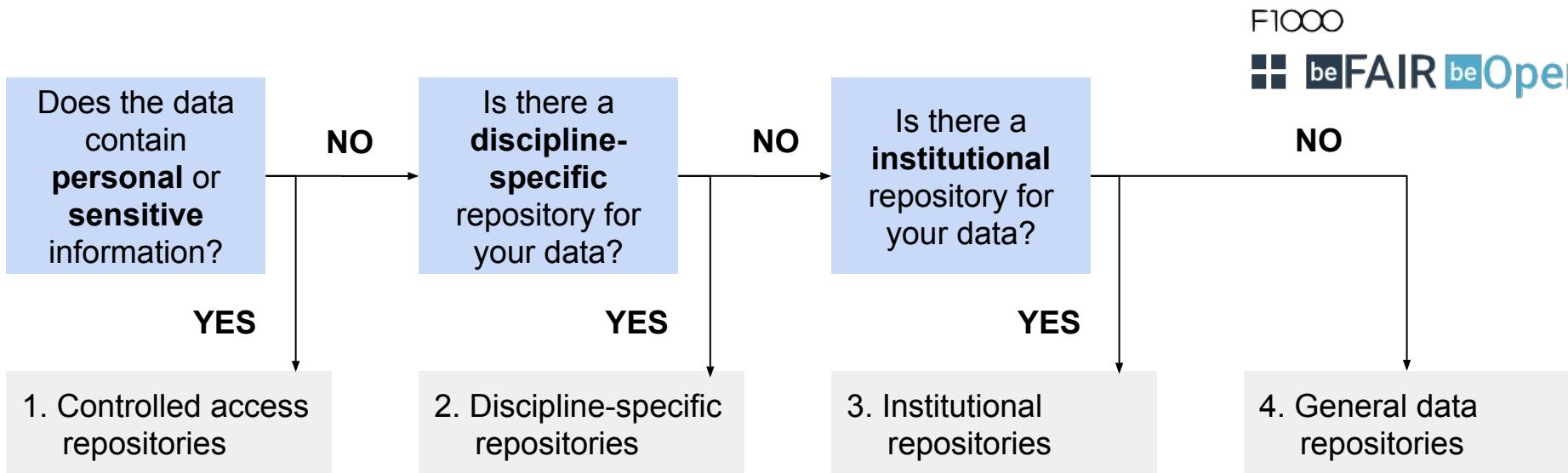
Uppsala Universitet: Uppsala, Sweden
1985-01 to 1989-04 (Microbiology)
BSc
Source: Niclas Jareborg
Created: 2015-04-09

Employment (7)

Stockholms Universitet: Stockholm, Sweden
2015-01 to present (BILS / Department of 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)

Repositories



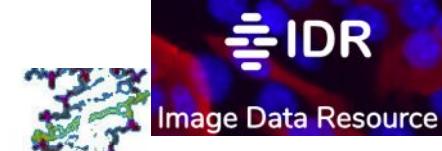
dbSNP
Short Genetic Variations



Etc...

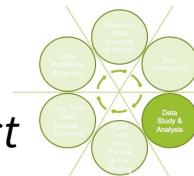


International public repositories



- 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-database>



Scientific Data Recommended Data Repositories

Biological sciences ↗

Nucleic acid sequence ↗

Sequence information should be deposited following the [MiS 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 ↗

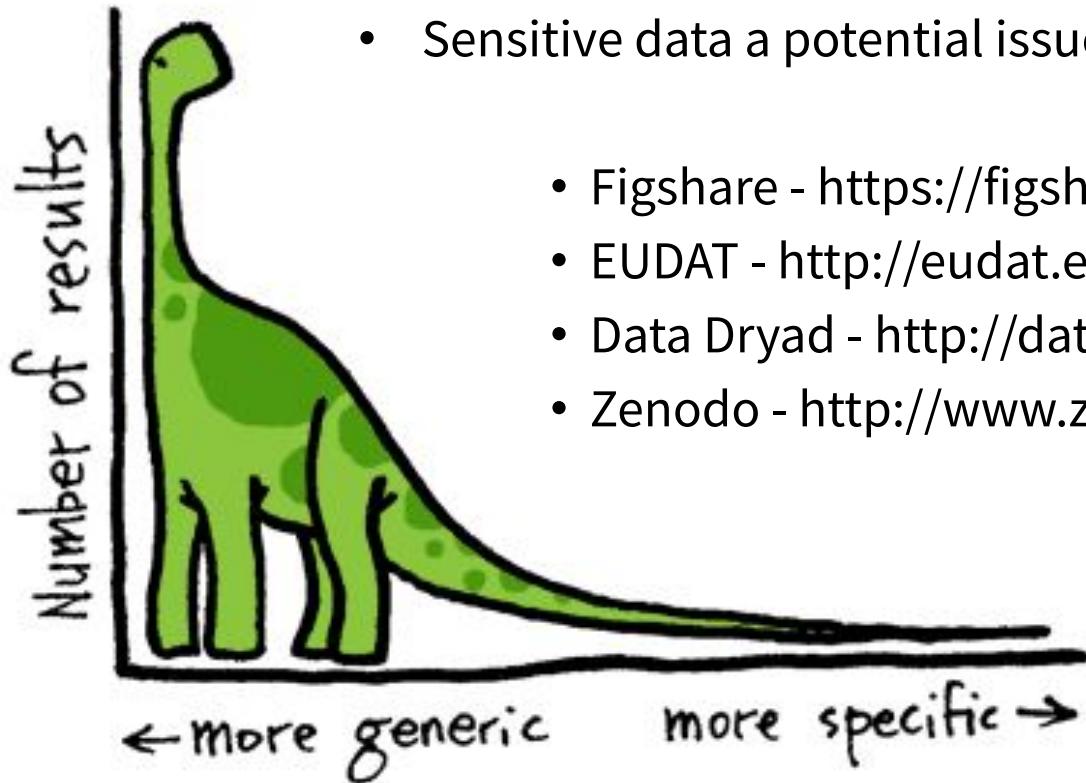
These repositories accept structural data for small molecules (COD); peptides and proteins (all); and larger assemblies (EMDB).

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>

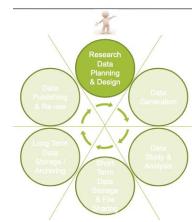
- Research data that doesn't fit in structured data repositories
- Data publication – persistent identifiers
- Metadata submission – not tailored to Life Science
 - *Affects discoverability*
 - *(Less) FAIR*
- Sensitive data a potential issue



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

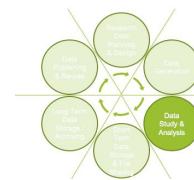
💡 Consider structuring metadata in the format needed by the repository already at **planning** stage

Planning & Design

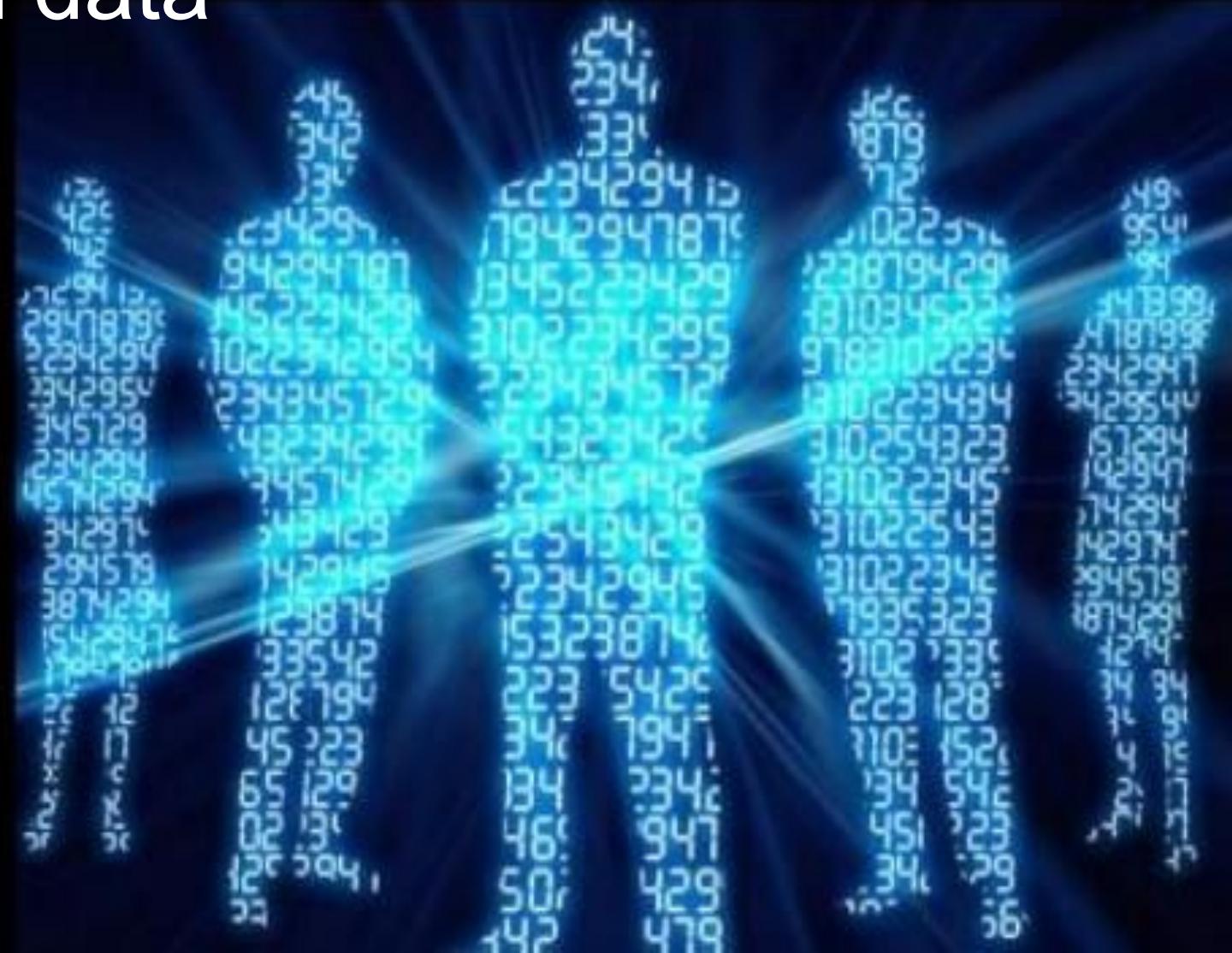


💡 Strive towards uploading data to its final destination already at when it has been **generated**

Study & Analysis



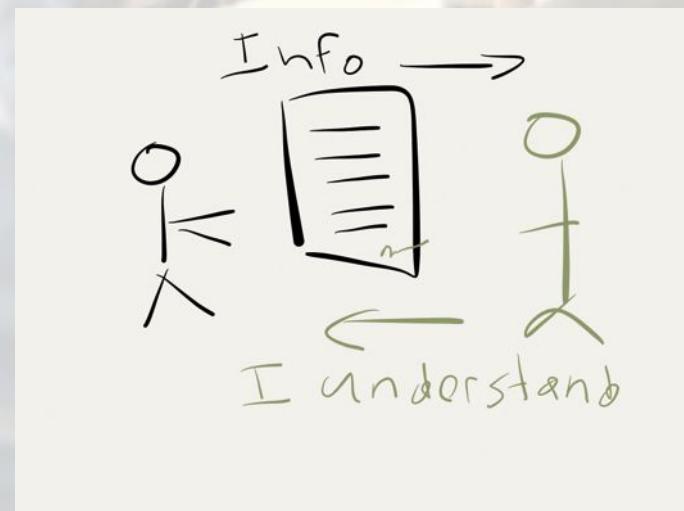
Personal data



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



- Research that concerns studies of biological material that has been taken from a living person and that can be traced back to that person may only be conducted if it has been approved subsequent to an ethical vetting
- Informed consent
 - The subject must be informed about the purpose or the research and the consequences and risks that the research might entail
 - The subject must consent



- 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.*
 - Decide a **purpose** and stick to it
 - Only collect data that is needed
 - Don't collect more data than necessary
 - Don't use data for another incompatible purpose
 - Erase data when no longer needed
 - Ensure that data is correct and updated
 - Protect collected data – confidential and intact
 - Identify the **legal basis** for data processing before it starts
 - Inform in a transparent and honest way
- The Data Inspection Board (*Datainspektionen*) is the Swedish Data Protection Authority



Exemptions for some of
these possible for research

- **Consent**
- To be able to fulfil contract with data subject
- Legal obligation
- Necessary in order to protect the vital interests of the data subject
- **Public interest**
- Necessary for the purposes of the legitimate interests pursued by the controller

- 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)

- The (legal) person that decides why and how personal data should be processed is called the **Controller** (*personuppgiftsansvarig*)
 - e.g. the employing university
 - Controller responsible for
 - Has to ensure the **rights of the individuals**
 - **Take measures** to ensure that the Regulation is followed, and be able to **show** that it is
 - **Privacy by Design** as standard
 - Keep a **register of processing**
 - Apply **security measures** when processing data
 - **Report** personal data **breaches** to the Data Protection Authority
 - Perform *Impact Assessments* and consult Data Protection Authority (when necessary)
 - Appoint **Data Protection Officer**

- The controller of personal data can delegate processing of personal data to a **Processor** (*personuppgiftsbiträde*)
 - e.g. UPPMAX/Uppsala university
 - Joint responsibility with Controller
 - A legal agreement **must** be established ([Art. 28.3](#))
 - Instruction from Controller to Processor on how data shall be processed

Data Protection Impact Assessment ([Art. 35.1](#))

- A personal data controller must carry out an impact assessment if a type of processing is **likely to result in a high risk to the rights and freedoms** of natural persons. The impact assessment must be carried out **prior** to the processing.
- The Swedish Data Protection Authority has adopted a list of where an impact assessment is required ([Swedish](#), [English](#))
 - 9 criteria
- Example processings requiring a DPIA
 - Processing, including storage for archiving purposes, of pseudonymised sensitive personal data that refers to data subjects from research projects or clinical trials. (Criteria 4 and 7)
 - Organisations that collect and store sensitive data in order to serve as a basis for future research purposes. (Criteria 4 and 7)

- Open source tool for DPIAs from CNIL (French DPA)

Version 1.6.3

Pia | Privacy impact assessment

DASHBOARD

The Project

- CONTEXT**
 - Overview
 - Data, processes and supporting ...
- FUNDAMENTAL PRINCIPLES**
 - Proportionality and necessity
 - Controls to protect the personal ...
- RISKS**
 - Planned or existing measures
 - Illegitimate access to data
 - Unwanted modification of data
 - Data disappearance
 - Risks overview
- VALIDATION**
 - Risk mapping
 - Action plan
 - DPO and concerned persons opi...

Validate PIA

ATTACHMENTS

+ Add

Context
This section gives you a clear view of the treatment(s) of personal data in question.

OVERVIEW
This part allows you to identify and present the object of the study.

Which is the processing under consideration?
Ingesting, storing and dissemination to authorized data requestors, on behalf of the data Controllers, of human genetic and phenotypic data collected for research purposes.

1 comment(s)
29/08/2018 Comment

What are the responsibilities linked to the processing?
Data provider is Controller
EGA-SE (NBIS) i.e. Uppsala university is Processor

0 comment(s)
03/09/2018 Comment

Are there standards applicable to the processing?
GA4GH Security Technology Infrastructure / Standards and practices...

Knowledge base

Principle
Processing's description

Definition
Controller

Definition
Processor

<https://www.cnil.fr/en/open-source-pia-software-helps-carry-out-data-protection-impact-assesment>

- **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 Authority

GU

https://medarbetarportalen.gu.se/diarieforing-arkiverin_g-och-personuppgiftsbehandling/rutiner-for-behandlin_g-av-personuppgifter/personuppgifter-i-forskning/?languageld=100001

KI

<https://staff.ki.se/gdpr-at-ki>

KTH

https://intra.kth.se/en/anstallning/anstallningsvillkor/att_vara_statligt_an/behandling_av_person/dataskyddsforordningen-gdpr-1.800623

LiU

<https://insidan.liu.se/dataskyddsforordningen?l=en>

LU

<https://staff.lu.se/dataprotection>

SU

<http://su.se/english/gdpr>

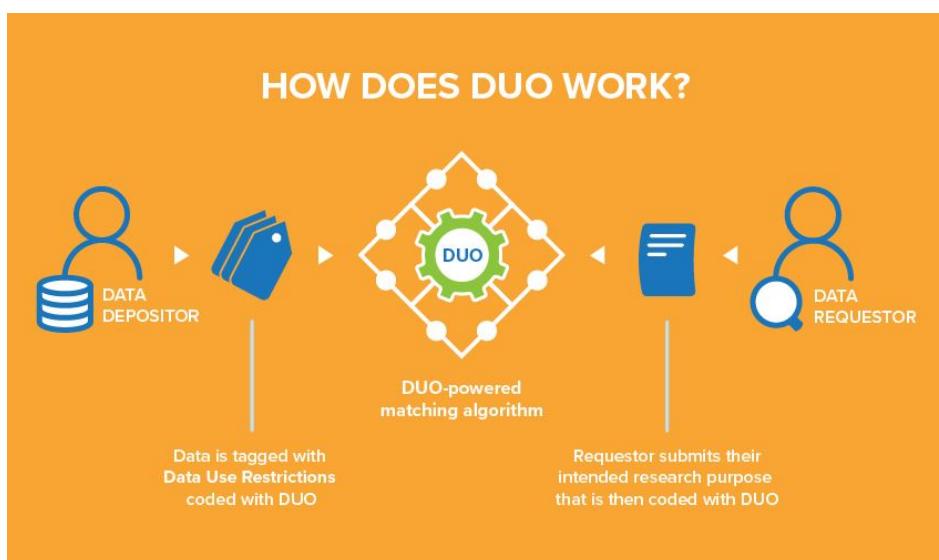
UmU

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

UU

<https://mp.uu.se/en/web/info/stod/dataskyddsforordningen>

- The genetic information of an individual is personal data
 - **Sensitive** personal data (as it relates to health)
 - Explicitly defining in GDPR
 - Even if *anonymized / pseudonymized*
 - In principle, **no** difference between WGS, Exome, Transcriptome or GWAS data
- *Theoretically* possible to identify the individual person from which the sequence was derived from the sequence itself
 - The more associated metadata there is, the easier this gets
 - Gymrek et al. “Identifying Personal Genomes by Surname Inference”. Science 339, 321 (2013); DOI:10.1126/science.1229566
- *Apply technical and organizational measures to protect the sensitive data, e.g.*
 - Strong IT security and procedures to limit access to data
 - Separate from other personal data
 - Pseudonymization
 - Encryption.



Consent Codes

Name	Abbreviation	Description
Primary Categories (I^o)		
No restrictions	NRES	No restrictions on data use.
General research use and clinical care	GRU(CC)	For health/medical/biomedical purposes, including the study of population origins or ancestry.
Health/medical/biomedical research and clinical care	HMB(CC)	Use of the data is limited to health/medical/biomedical purposes; does not include the study of population origins or ancestry.
Disease-specific research and clinical care	DS-[XX](CC)	Use of the data must be related to [disease].
Population origins/ancestry research	POA	Use of the data is limited to the study of population origins or ancestry.
Secondary Categories (II^o) (can be one or more extra conditions, in addition to I ^o category)		
Other research-specific restrictions	RS-[XX]	Use of the data is limited to studies of [research type] (e.g., pediatric research).
Research use only	RUO	Use of data is limited for research purposes (e.g., does not include its use in clinical care).
No "general methods" research	NMDS	Use of the data includes methods development research (e.g., development of software or algorithms) ONLY within the bounds of other data use limitations.
Genetic studies only	GSO	Use of the data is limited to genetic studies only (i.e., no "phenotype-only" research).
Requirements		
Not-for-profit use only	NPU	Use of the data is limited to not-for-profit organizations.
Publication required	PUB	Requestor agrees to make results of studies using the data available to the larger scientific community.
Collaboration required	COL-[XX]	Requestor must agree to collaboration with the primary study investigator(s).
Ethics approval required	IRB	Requestor must provide documentation of local IRB/REC approval.
Geographical restrictions	GS-[XX]	Use of the data is limited to within [geographic region].
Publication moratorium/embargo	MOR-[XX]	Requestor agrees not to publish results of studies until [date].
Time limits on use	TS-[XX]	Use of data is approved for [x months].
User-specific restrictions	US	Use of data is limited to use by approved users.
Project-specific restrictions	PS	Use of data is limited to use within an approved project.
Institution-specific restrictions	IS	Use of data is limited to use within an approved institution.

doi:10.1371/journal.pgen.1005772.t001



Global Alliance
for Genomics & Health
Collaborate. Innovate. Accelerate.



Nordic Collaboration for Sensitive Data



NordForsk



<https://neic.no/tryggve/>

Tryggve vision

Tryggve2 develops and facilitates access to secure e-infrastructure for sensitive data, suitable for hosting large-scale cross-border biomedical research studies

<https://neic.no/tryggve/>



Tryggve2 major deliverables

2017 – 2020

1. **Sensitive data archiving**
2. Production quality processing services
3. Homogenized user experience
 - User mobility
 - Workflow mobility
 - Data mobility
4. **Nordic use cases** <https://neic.no/tryggve/usecase/>
 - Research
 - Infrastructure development
5. ELIXIR AAI
6. IT Security
7. **ELSI Topics**

<https://neic.no/tryggve/>



Tryggve ELSI checklist

Tryggve Checklist on ELSI issues and GDPR compliance

Cohorts / Datasets	
List and number the cohorts/datasets that will be used in the project	
1.	
2.	
3.	
4.	
5.	
Ethical reviews and informed consents	
Has the project (or parts of the project) undergone ethical review ?	
<input type="checkbox"/> Yes <input type="checkbox"/> Yes, parts <input type="checkbox"/> No <input type="checkbox"/> Needs to be confirmed	
<ul style="list-style-type: none"> What are the <i>limitations of use</i> in the ethics approval, if any? List per cohort/dataset <ul style="list-style-type: none"> e.g. only for research on certain types of diseases, sharing only within certain geographical boundaries, etc 	
1.	
2.	
3.	
4.	
5.	
Have informed consents been collected from the research subjects?	
<input type="checkbox"/> Yes <input type="checkbox"/> Yes, for some cohorts <input type="checkbox"/> No <input type="checkbox"/> Needs to be confirmed	
<ul style="list-style-type: none"> What are the <i>limitations of use</i> defined in the informed consent, if any? List per cohort/dataset <ul style="list-style-type: none"> e.g. only for research on certain types of diseases, sharing only within certain geographical boundaries etc 	
1.	
2.	
3.	
4.	
5.	
State the intended research purpose	

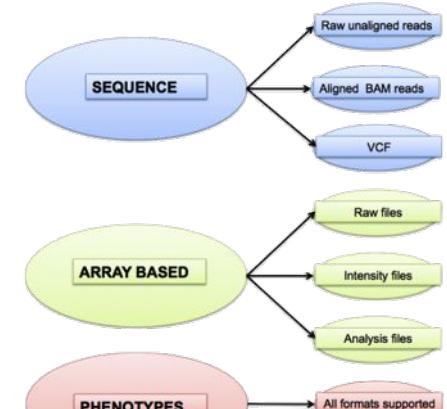
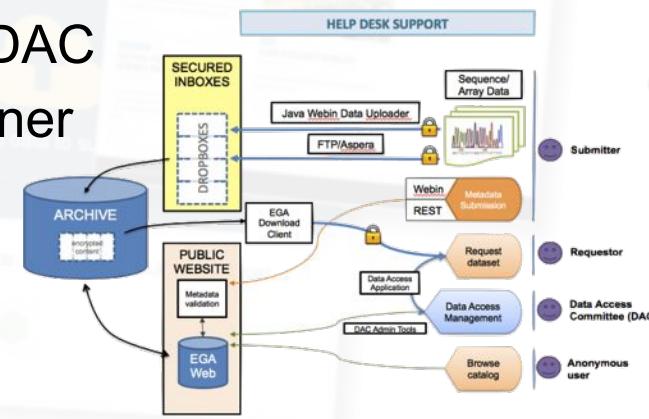
GDPR	
Is the intended research purpose within the scope of the <i>limitations of use</i> that is defined in the ethics approval(s) and/or the informed consent(s)?	
<input type="checkbox"/> Yes <input type="checkbox"/> No <input type="checkbox"/> Needs to be confirmed	
State the purpose of processing the personal data	
Who are the data controllers of the personal data processed in the project?	
<ol style="list-style-type: none"> 1. 2. 3. 4. 5. 	
<ul style="list-style-type: none"> If there are more than one controller of the personal data processed in the project, will the parties be joint controllers? <ul style="list-style-type: none"> Has a joint controllership agreement between the parties been established? 	
<input type="checkbox"/> Yes <input type="checkbox"/> No, separate for each cohort <input type="checkbox"/> Needs to be investigated	
<input type="checkbox"/> Yes <input type="checkbox"/> No, but it should be <input type="checkbox"/> Needs to be investigated	
What is the legal basis for processing the personal data?	
State cohort/dataset for each type of legal basis	
<ul style="list-style-type: none"> Public interest Cohorts: Consent Cohorts: Are consents in compliance with the GDPR? Cohorts: <input type="checkbox"/> Other, which? 	
What are the exemptions for the prohibition for processing of special categories of data (such as health and genetic data) under Art. 9 GDPR used?	
State cohort/dataset for each type of exemption.	
<input type="checkbox"/> Scientific research Cohorts: <input type="checkbox"/> Consent Cohorts: <input type="checkbox"/> Needs to be investigated	
Have data processing agreements been established between the data controller(s) and any data processors ? List processors and agreements established for each of these	
<p>Note: A data processing agreement has to contain the obligatory clauses specified in Art 28.3 of the GDPR. The agreement should also regulate the use of any sub-processors.</p> <input type="checkbox"/> Needs to be investigated	

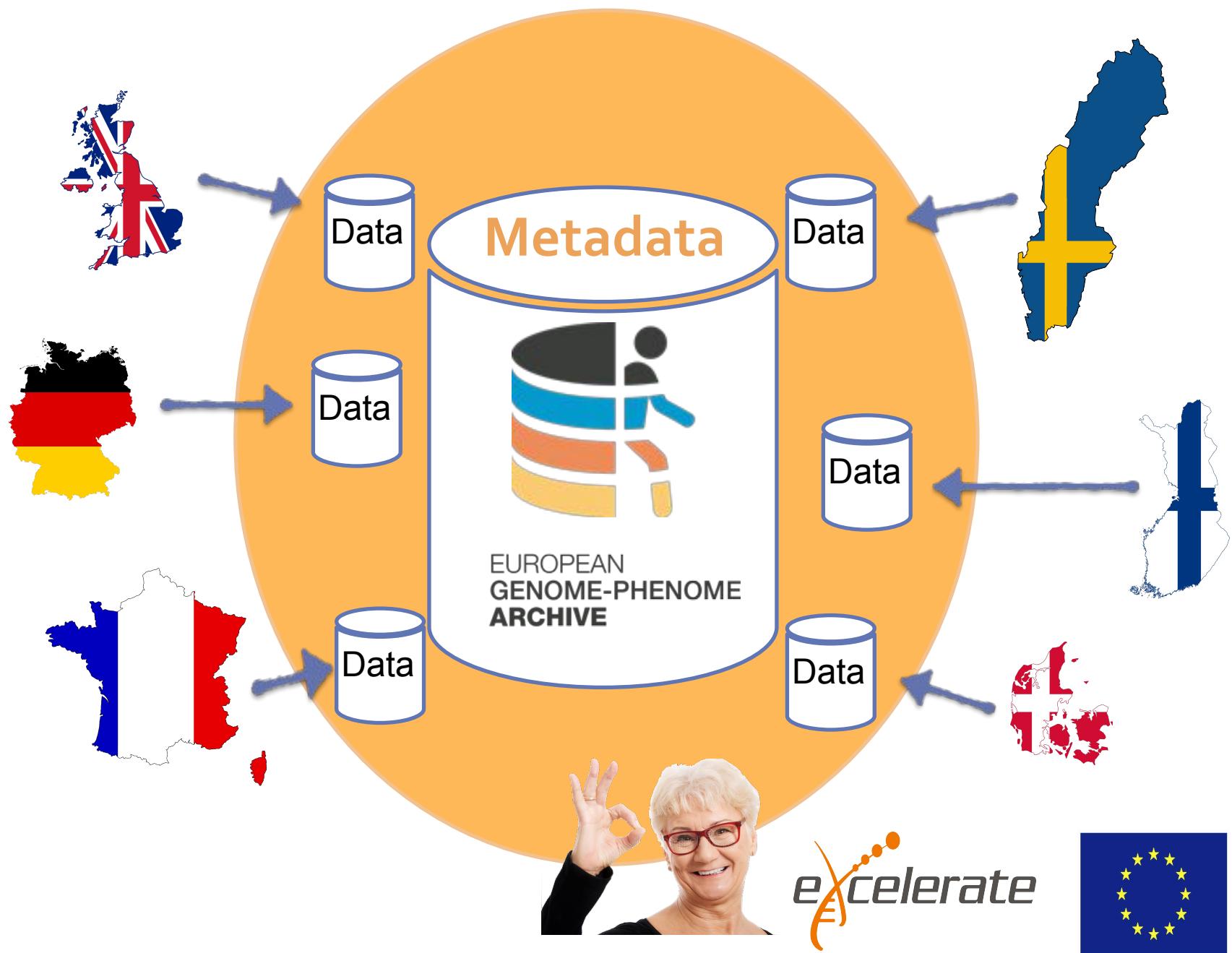
<input type="checkbox"/> Not relevant Processors: <ul style="list-style-type: none"> Agreements: o Agreements: o Agreements: o 	
Have Data Protection Impact Assessments (DPIA) been performed for the personal data? List DPIAs done and for which parts of the data	
<input type="checkbox"/> Needs to be investigated <input type="checkbox"/> Not needed • •	
What technical and procedural safeguards have been established for processing the data?	
Other considerations	
Are there other relevant national legislation considerations that has to be taken into account? <ul style="list-style-type: none"> e.g. regarding public access to information (in particular SE?), biobank acts, etc. 	
<input type="checkbox"/> Needs to be investigated <input type="checkbox"/> No <input type="checkbox"/> Yes	
Are there other Terms & conditions for data access (in particular if presenting obstacles for cross-border processing of health data)?	
<ul style="list-style-type: none"> e.g. register data access policies (requirement of PI in the same country, moving data to other secure services) 	
<input type="checkbox"/> Needs to be investigated <input type="checkbox"/> No <input type="checkbox"/> Yes	
Are there other legal agreements between use case parties that should be considered? <ul style="list-style-type: none"> e.g. conditions regarding data reuse 	
<input type="checkbox"/> Needs to be investigated <input type="checkbox"/> No <input type="checkbox"/> Yes	

“As open as possible, as closed as necessary”

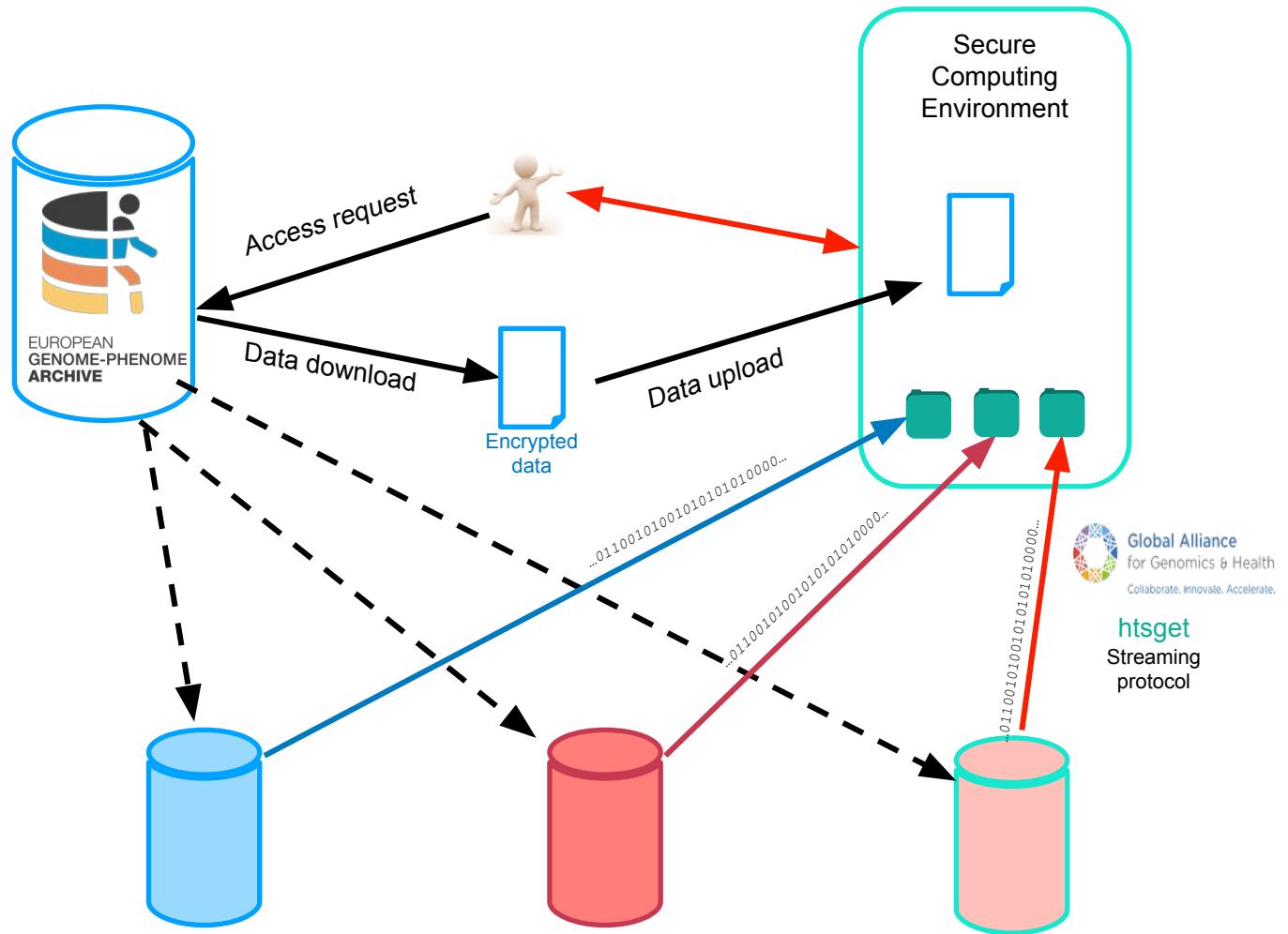


- **EGA – European Genome-phenome Archive**
 - Repository that promotes the distribution and sharing of **genetic and phenotypic data** consented for specific approved uses but **not fully open, public distribution**.
 - All types of sequence and genotype experiments, including case-control, population,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

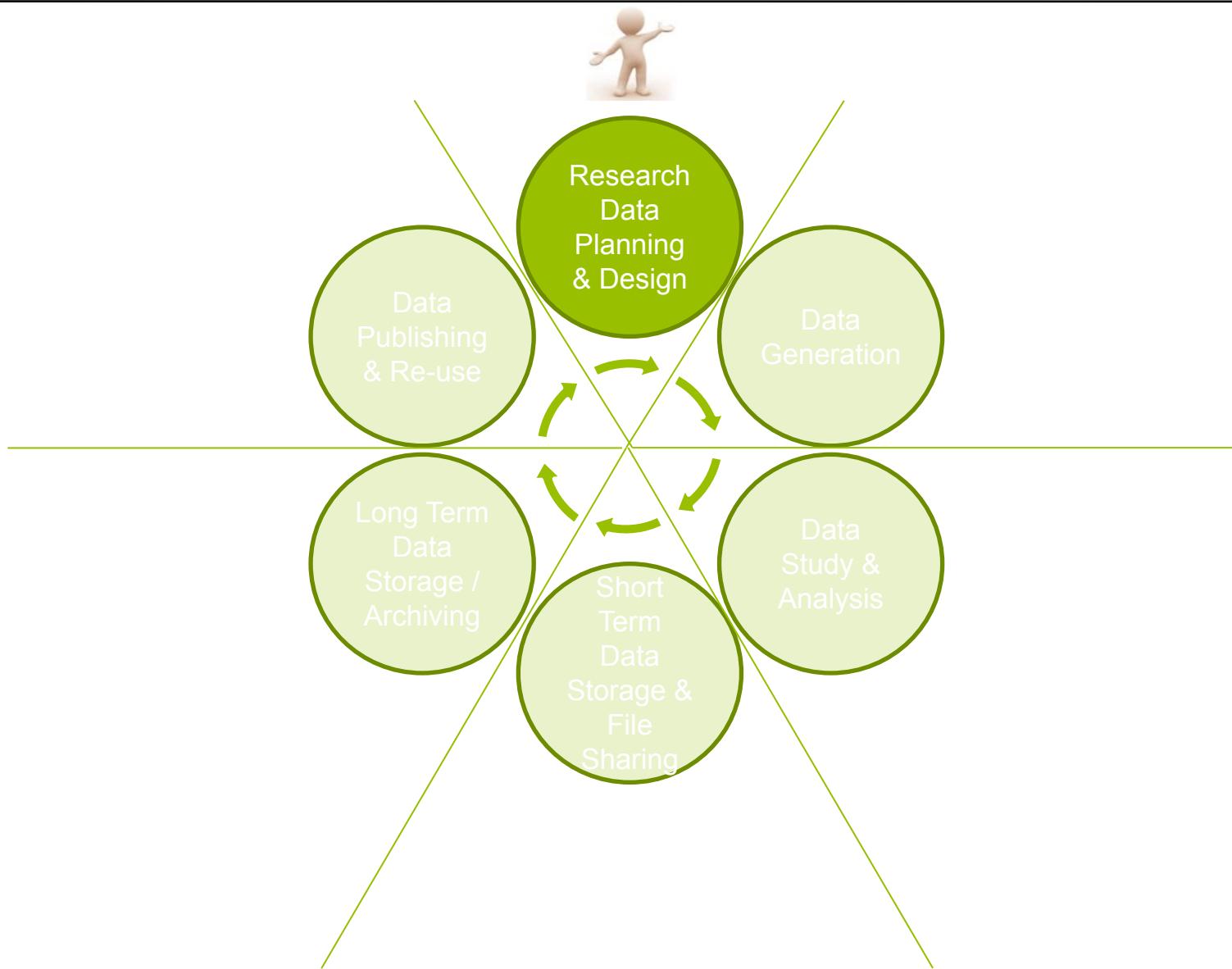


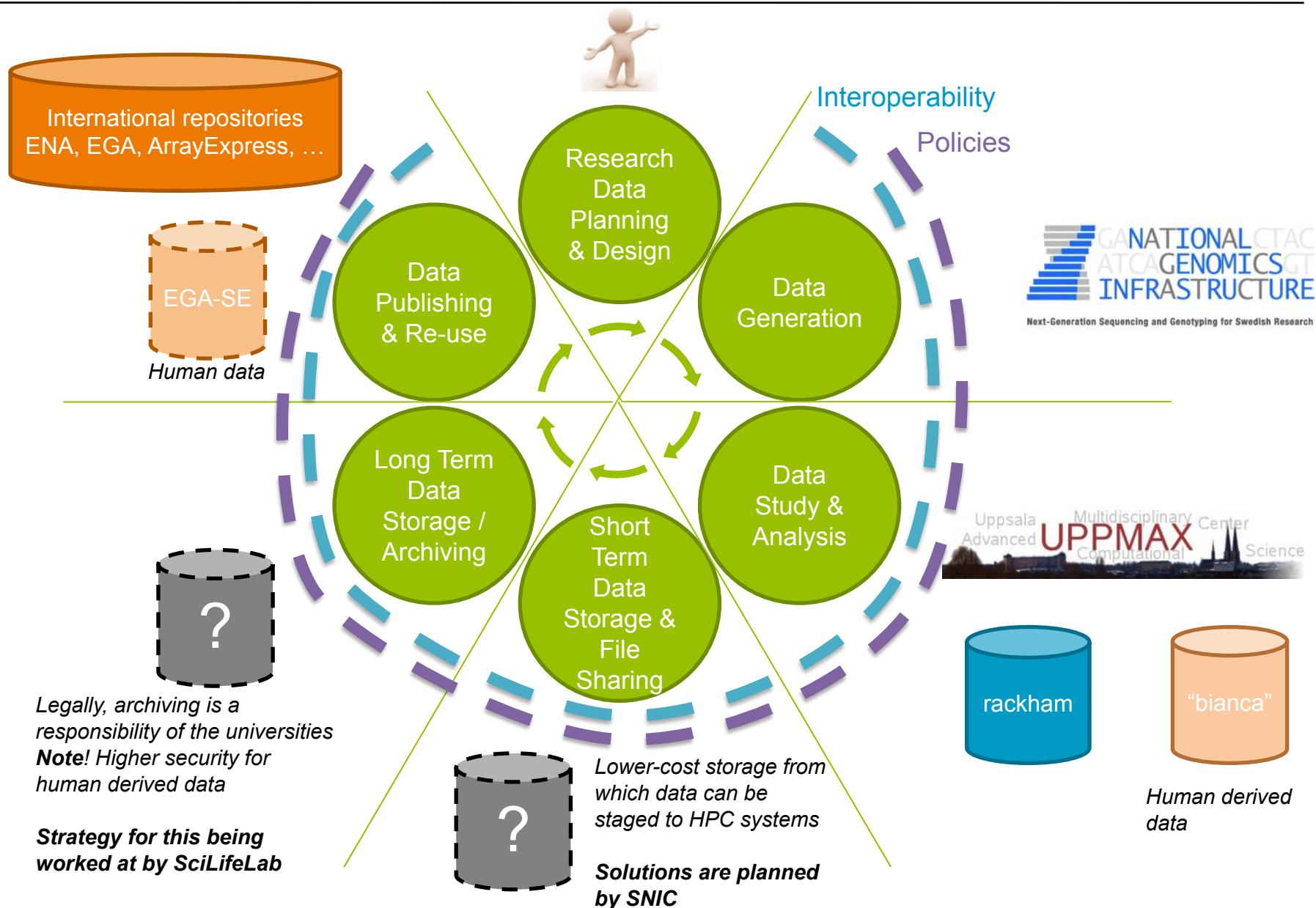


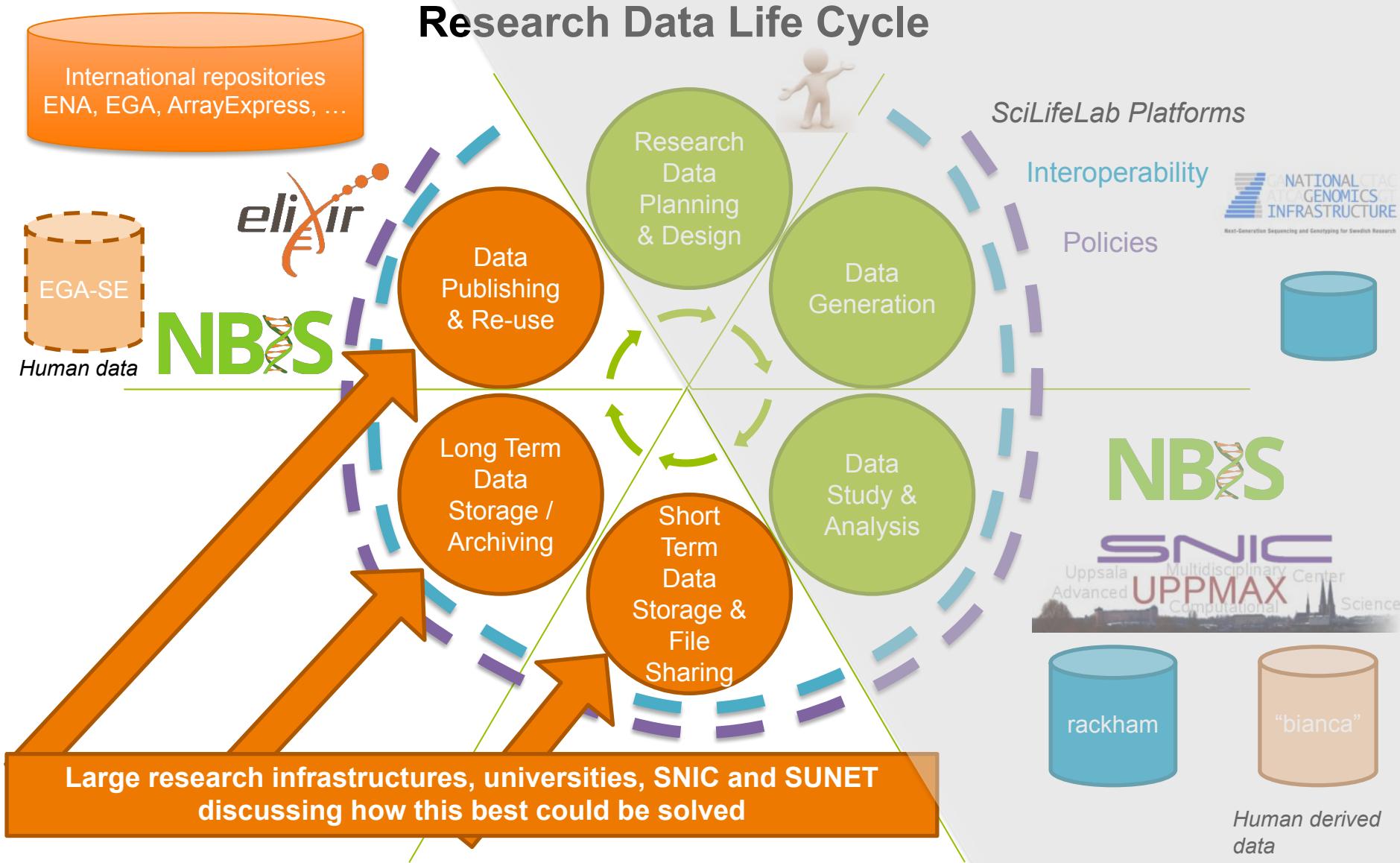
Future ways of accessing data



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







Data Centre / NBIS differences

SciLifeLab

Others



SciLifeLab

SciLifeLab



Researchers
–
Supported
projects

NBIS
Support for
research projects
Development
Data management
Training
ELIXIR hub

Data Centre
Data management
IT- and data services
and resources
IT coordination

Platforms
–
Data
production



- 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)

- Is it ethical to do bad/careless science?
 - Wasting resources
 - ... or even dangerous medical practices
 - Contribute to the current research credibility crisis
 - harming the profession
 - harming the public trust
- But!
 - Careless science -> longer CV

- You set the culture in your research group
- Promote best-practice data management
 - Pick a thought-through file and folder structure organization for your computational analyses
 - Strive for reproducibility
 - Data & Code
 - Organize project metadata from the start
 - In ways that makes it easy to submit to public repositories
 - Use available standards
 - Plan for submitting "raw data" to public repositories as early as possible
- Be aware that there are legal aspects to processing human data
- Make Data Management Plans for your projects
 - Ensure that your research output is FAIR
- *Ask for help if you need it!*

- 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
- FAIR
 - Luiz Bonino -
https://indico.neic.no/event/56/sessions/97/attachments/47/77/2.2._Day_2_PM_Late_-_Luiz_-_FAIR_Principles_explained_and_actionized.pdf
- 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/>
- Ethics
 - Rochelle Tractenberg “[Unexpected Ethical Challenges in Bioinformatics and Genomics](#).”
- ... and probably others I have forgotten

Will become a standard part of the research funding process



—
EDITORIAL • 13 MARCH 2018

Everyone needs a data-management plan

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

By 2019, all who receive grants from us must have a data management plan

As from spring 2019, if you are awarded a grant from the Swedish Research Council you must have a plan for how the research data generated within your project shall be managed.

You must not send in your data management plan to us when you apply for a grant, but your administering organisation will be responsible for ensuring that a data management plan is in place when you start your project or corresponding, and that the plan is maintained.

- VR & SUHF (Association of Swedish Higher Education Institutions)
 - **Central parts of a data management plan**
 - Based on Science Europe's "[Core Requirements for Data Management Plans](#)"
1. Description of data – reuse of existing data and/or production of new data
 2. Documentation and data quality
 3. Storage and backup
 4. Legal and ethical aspects
 5. Accessibility and long-term storage
 6. How will the use of unique and persistent identifiers, such as a Digital Object Identifier (DOI), be safeguarded?

1. Description of data – reuse of existing data and/or production of new data
 - How will data be collected, created or reused?
 - What types of data will be created and/or collected, in terms of data format and amount/volume of data?
2. Documentation and data quality
 - How will the material be documented and described, with associated metadata relating to structure, standards and format for descriptions of the content, collection method, etc.?
 - How will data quality be safeguarded and documented (for example repeated measurements, validation of data input, etc.)?
3. Storage and backup
 - How is storage and backup of data and metadata safeguarded during the research process?
 - How is data security and controlled access to data safeguarded, in relation to the handling of sensitive data and personal data, for example?
4. Legal and ethical aspects
 - How is data handling according to legal requirements safeguarded, e.g. in terms of handling of personal data, confidentiality and intellectual property rights?
 - How is correct data handling according to ethical aspects safeguarded?
5. Accessibility and long-term storage
 - How, when and where will research data or information about data (metadata) be made accessible? Are there any conditions, embargoes and limitations on the access to and reuse of data to be considered?
 - In what way is long-term storage safeguarded, and by whom? How will the selection of data for long-term storage be made?
 - Will specific systems, software, source code or other types of services be necessary in order to understand, partake of or use/analyse data in the long term?
6. How will the use of unique and persistent identifiers, such as a Digital Object Identifier (DOI), be safeguarded?
 - Responsibility and resources
 - Who is responsible for data management and (possibly) supports the work with this while the research project is in progress? Who is responsible for data management, ongoing management and long-term storage after the research project has ended?
 - What resources (costs, labour input or other) will be required for data management (including storage, back-up, provision of access and processing for long-term storage)? What resources will be needed to ensure that data fulfil the FAIR principles?

DMP tools

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?

My Dashboard Create plan Reference Help Language Rob Hooft

ELIXIR Data Stewardship Wizard

Go to App

DSW
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) (unsaved changes) Save

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?

Help

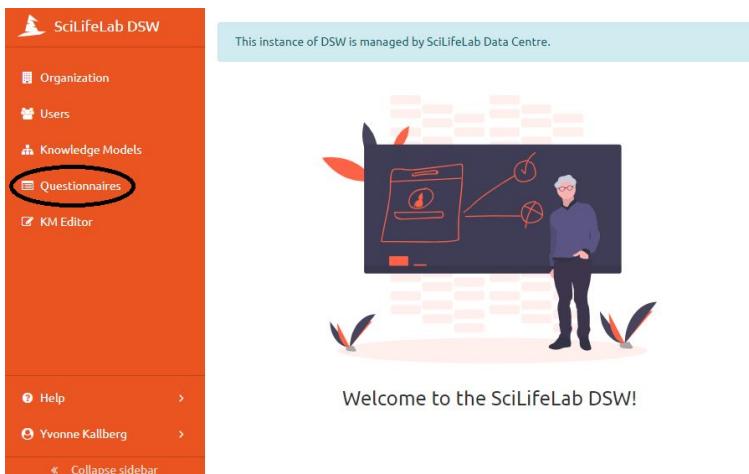
<https://dmponline.dcc.ac.uk/>

<https://ds-wizard.org/>

<https://dsw.scilifelab.se>

Hands-on DMP session

- Go to <https://dsw.scilifelab.se/>
- Create an account, login after confirmation email
- Click on Questionnaires in left-side menu
- Write a name based on your name and project
- Select ‘Common SciLifeLab DMP 1.1.0” as Knowledge Model
- Save and begin answering the questions



This instance of DSW is managed by SciLifeLab Data Centre.

Welcome to the SciLifeLab DSW!

Create Questionnaire

Name
YvonneKallberg_SDR

Knowledge Model
Common SciLifeLab DMP 1.1.0 (SciLifeLab:SciLifeLab-DMP:1.1.0)

Accessibility
 Private
 Questionnaire is visible only to you.

Public Read-Only
 Questionnaire can be viewed by other users, but they cannot change it.

Public
 Questionnaire can be accessed by all users.

Tags
There are no tags configured for the Knowledge Model

Cancel **Save**