

Managing your data

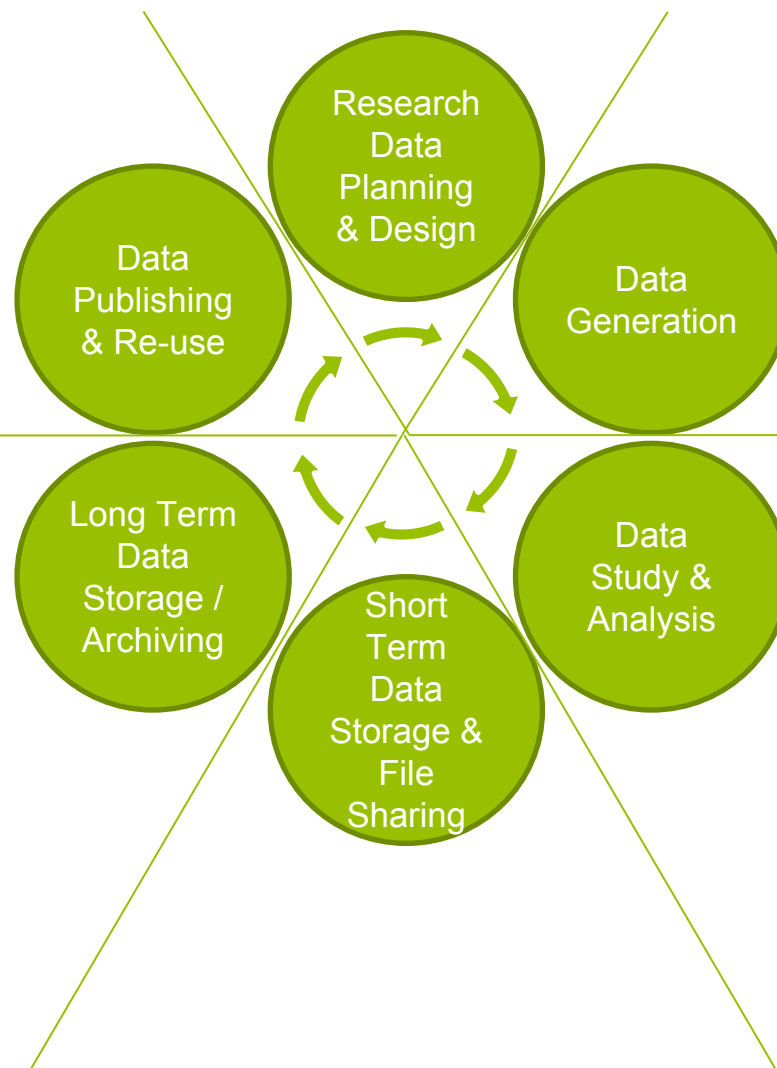
Niclas Jareborg, NBIS

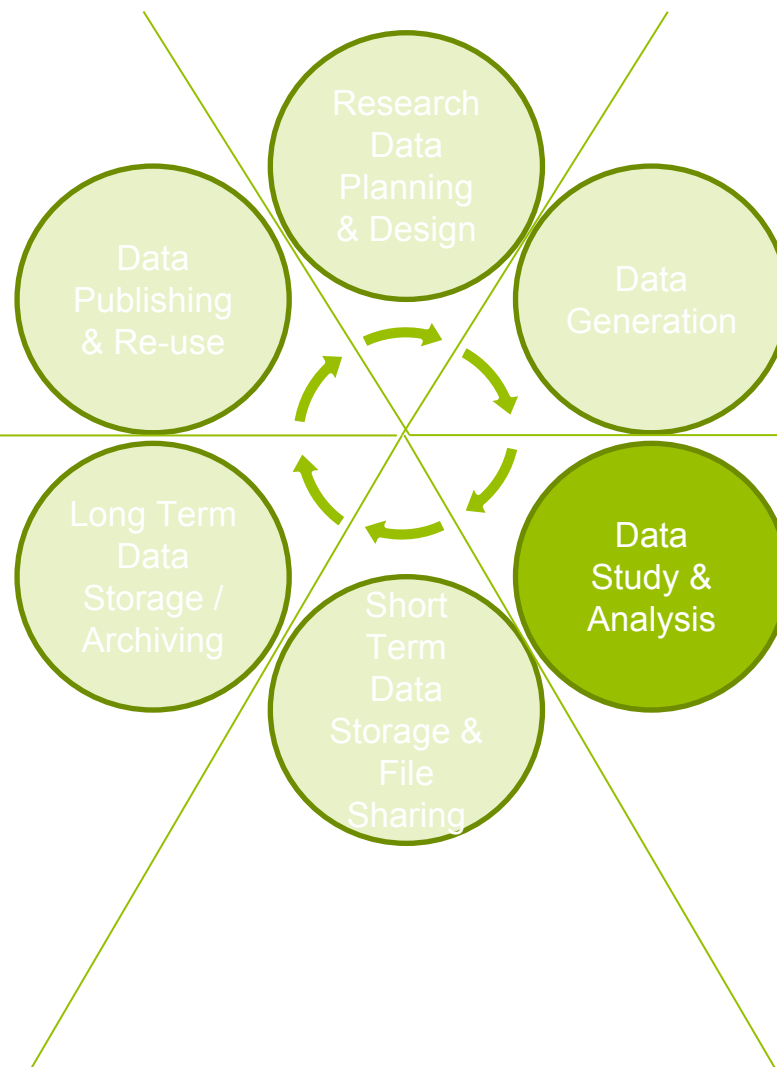
niclas.jareborg@nbis.se

Introduction to Bioinformatics using NGS data



**How do you know how an
old result was generated?**





Human derived data

- 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](#)



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.”

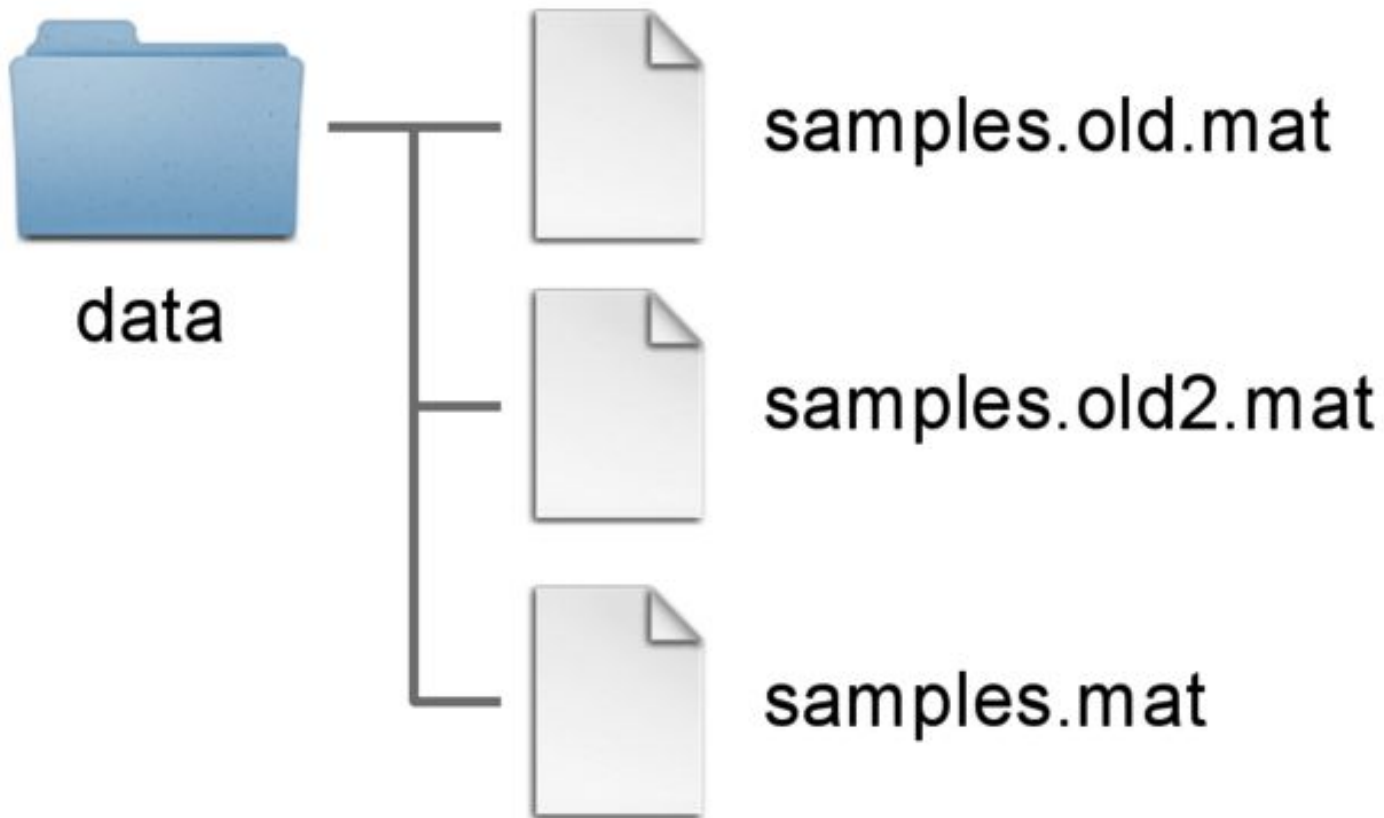


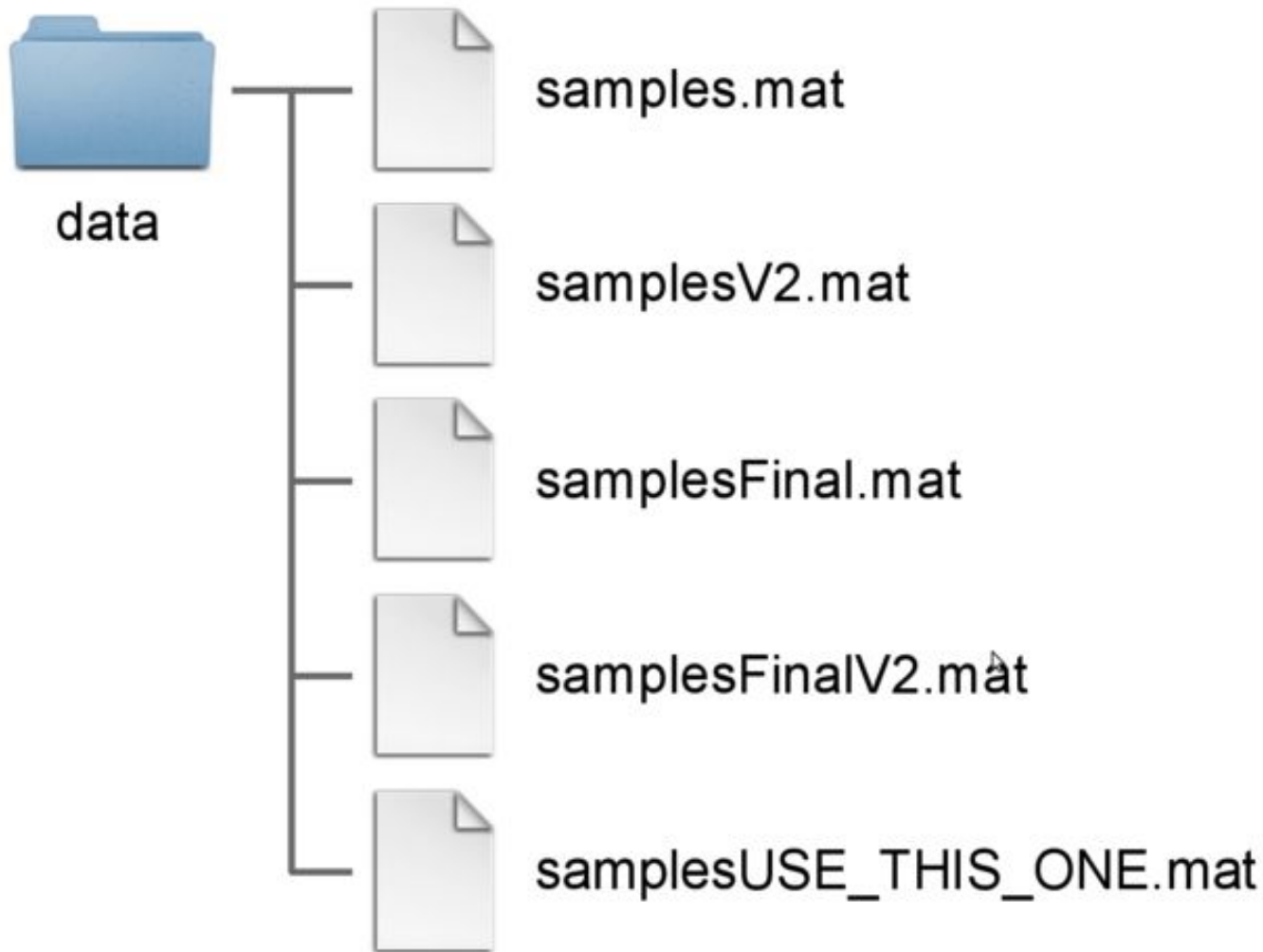
data

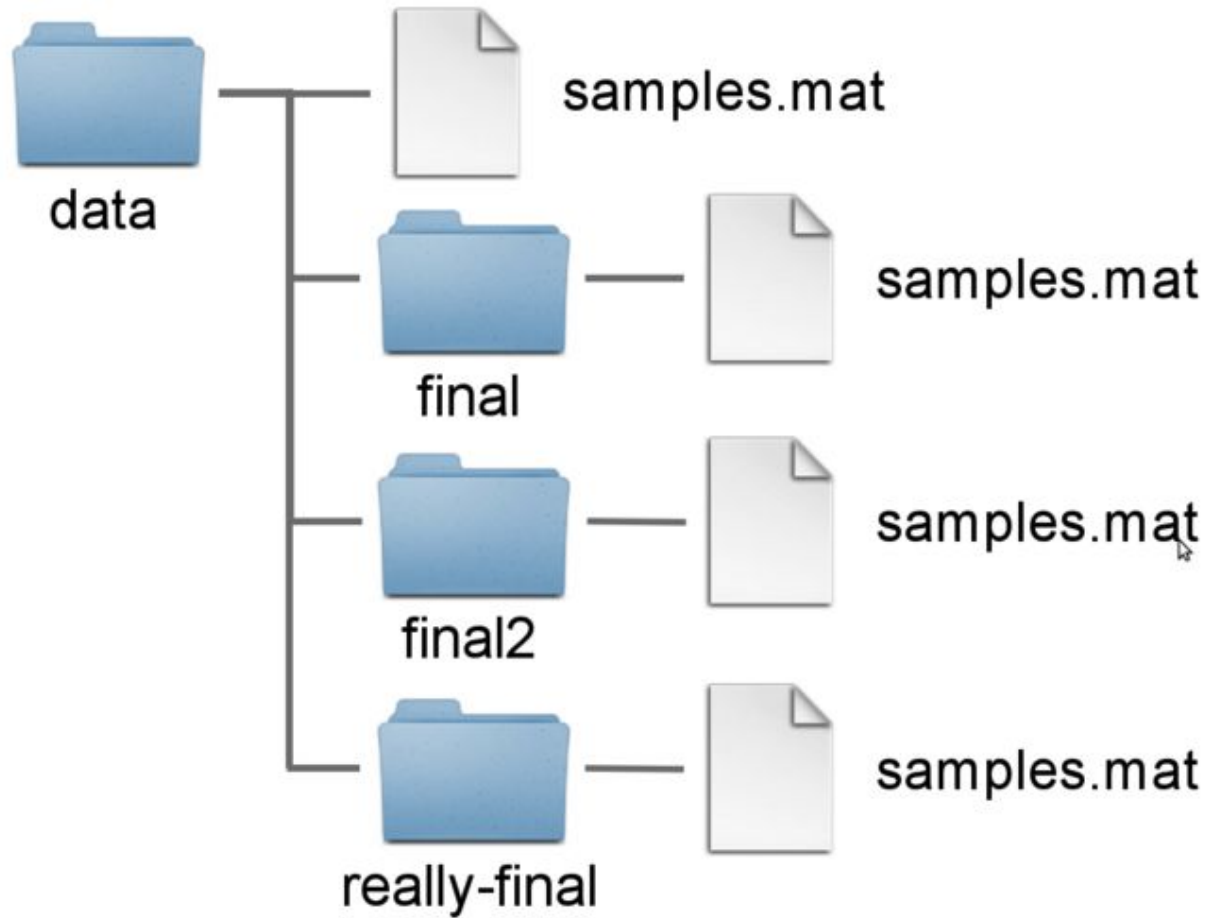


samples.mat





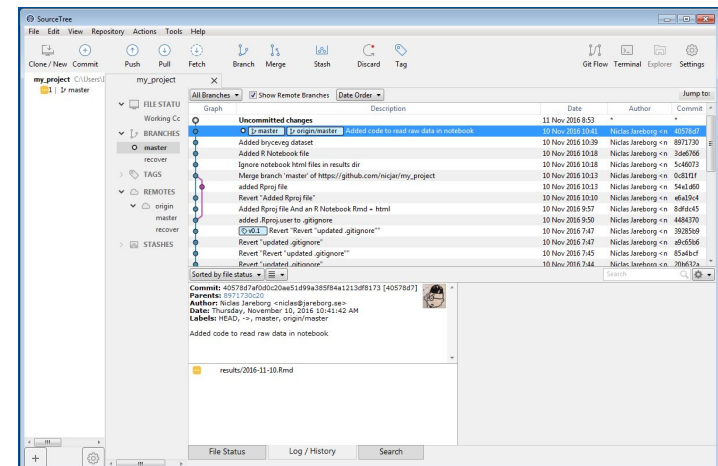






-
- There is a **folder for the raw data**, which do not get altered, or intermixed with data that is the result of manual or programmatic manipulation. I.e., derived data is kept separate from raw data, and **raw data are not duplicated**.
 - **Code is kept separate from data.**
 - Use a **version control 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...

- 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



-
- 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...

- 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

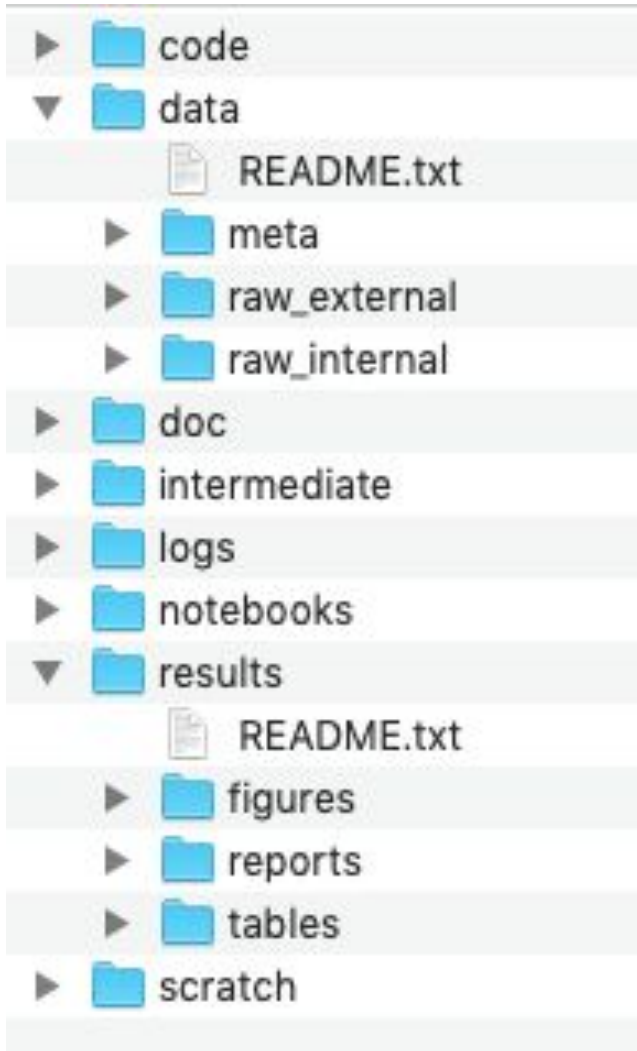
1986-01-28_raw-data-from-challenger-o-rings.txt

-
- 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...

- 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 formatting
 - Can be converted to lots of different outputs
 - HTML, pdf, MS Word, slides etc
 - Works nicely with **version control systems**

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





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

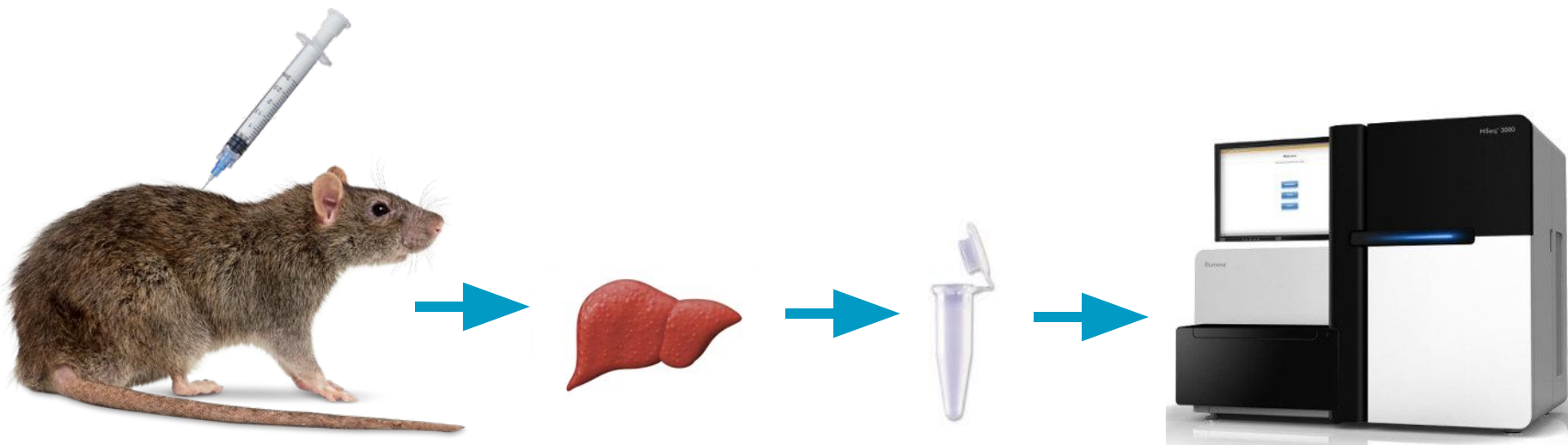
output files from different analysis steps, *can be deleted*

logs from the different analysis steps

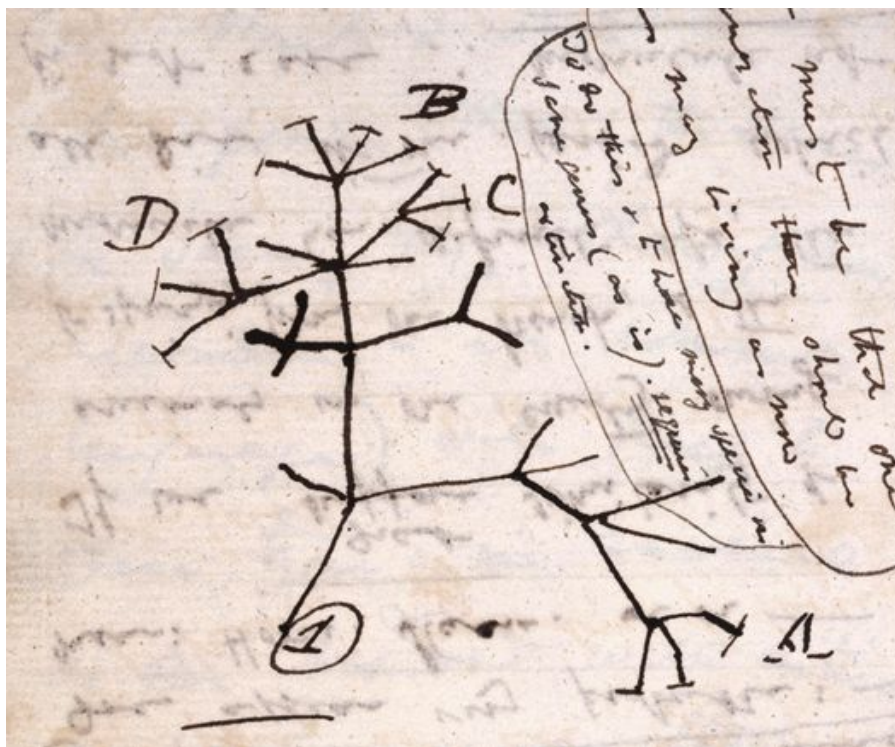
output from workflows and analyses

temporary files that can be safely *deleted or lost*

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

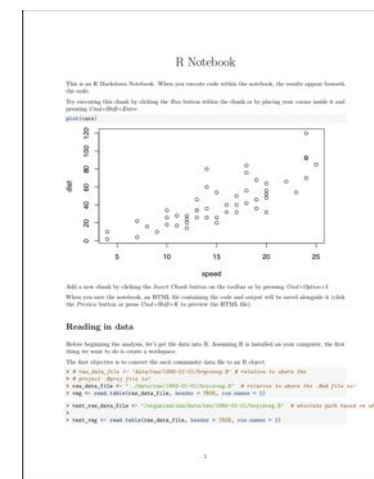
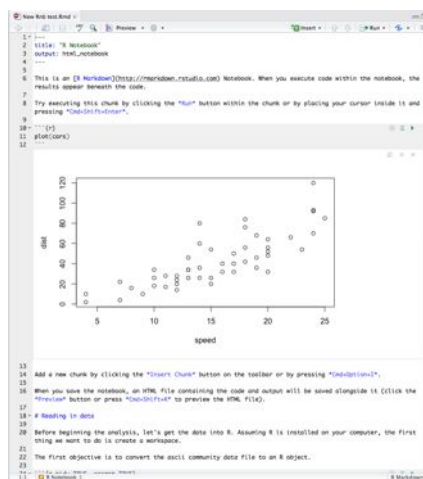
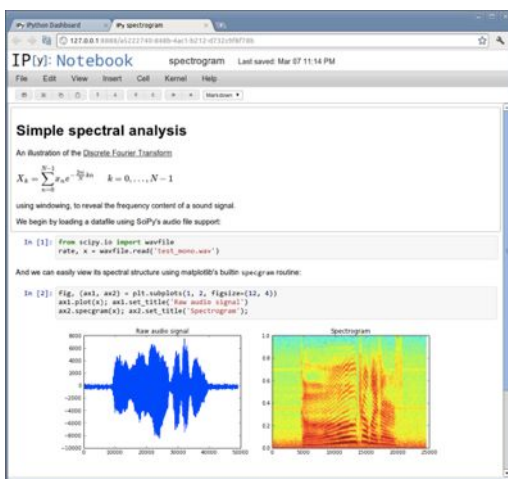


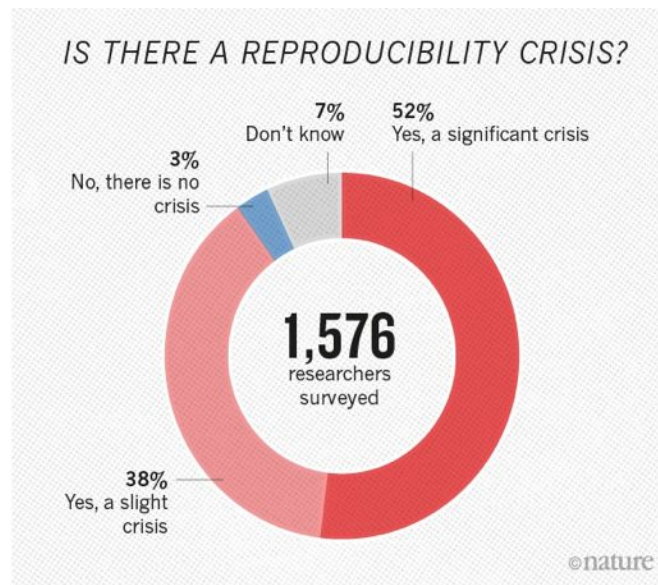
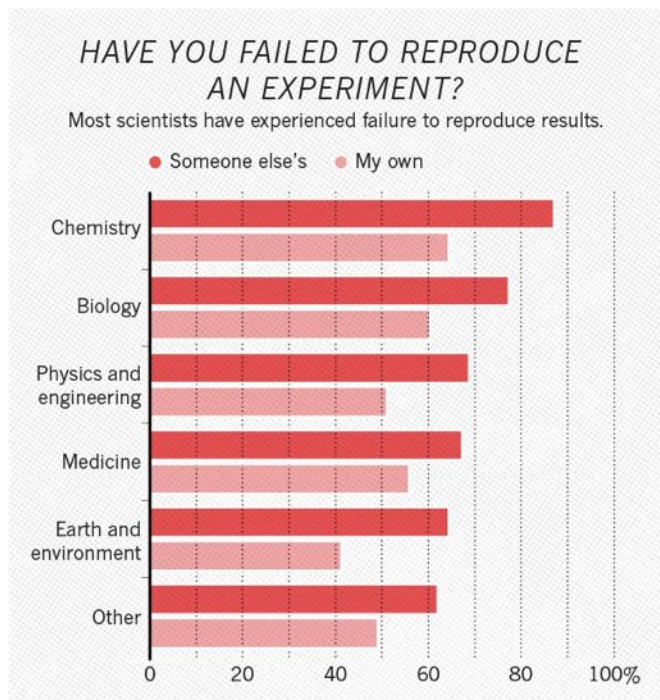
- Why?
 - You have to understand what you have done
 - **Others should be able to reproduce what you have done**



-
- 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

- 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



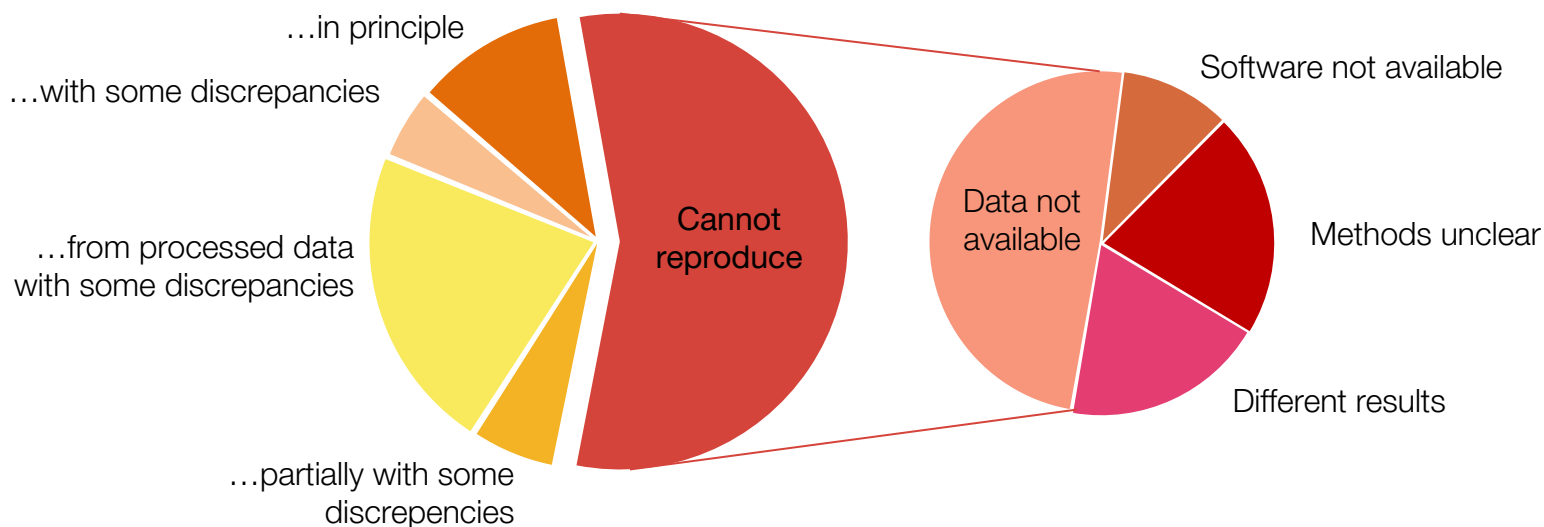


[1] "1,500 scientists lift the lid on reproducibility". Nature. 533: 452–454

[2] Begley, C. G.; Ellis, L. M. (2012). "Drug development: Raise standards for preclinical cancer research". Nature. 483 (7391): 531–533.

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

Can reproduce...



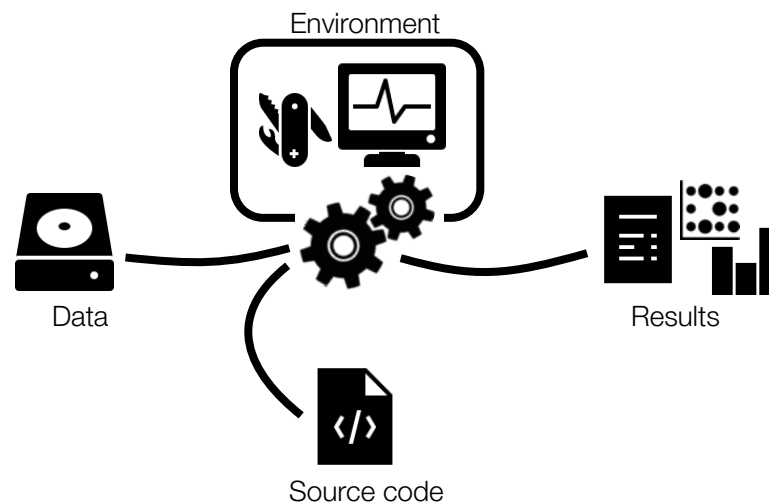
Summary of the efforts to replicate the published analyses.

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:



🏠 NBIS Reproducible research course

Search docs

- Welcome
- About
- The course
- Schedule
- Travel info
- Feedback
- Tutorials

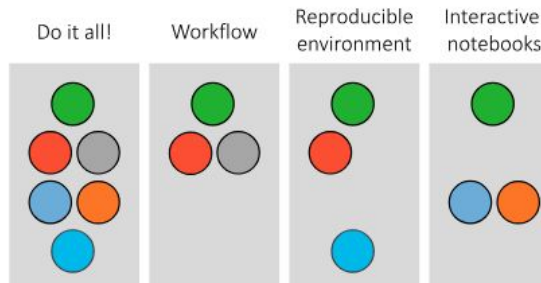
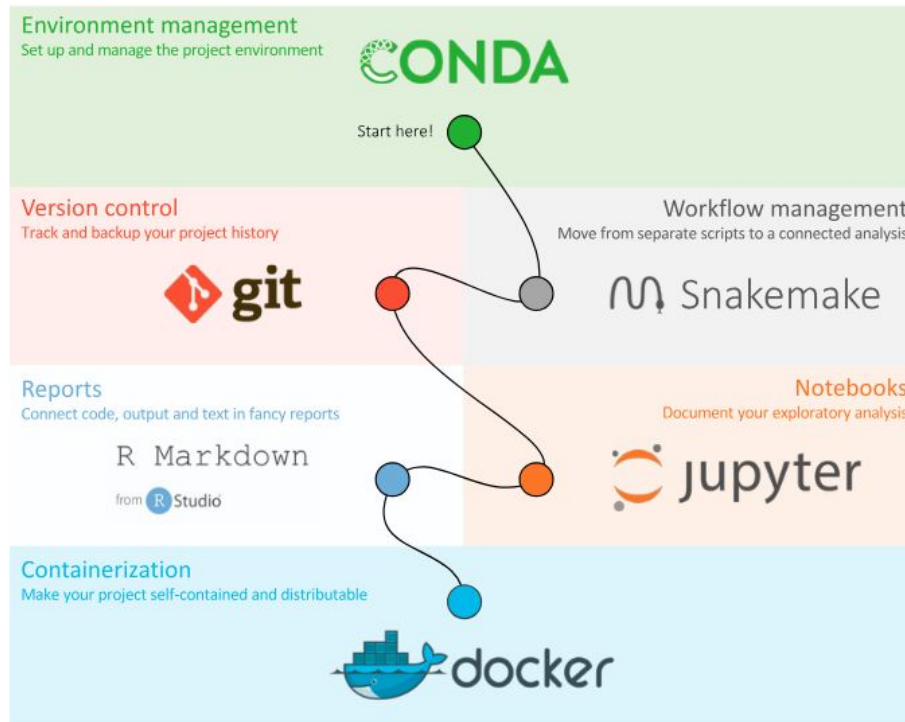
Introduction to the tutorials

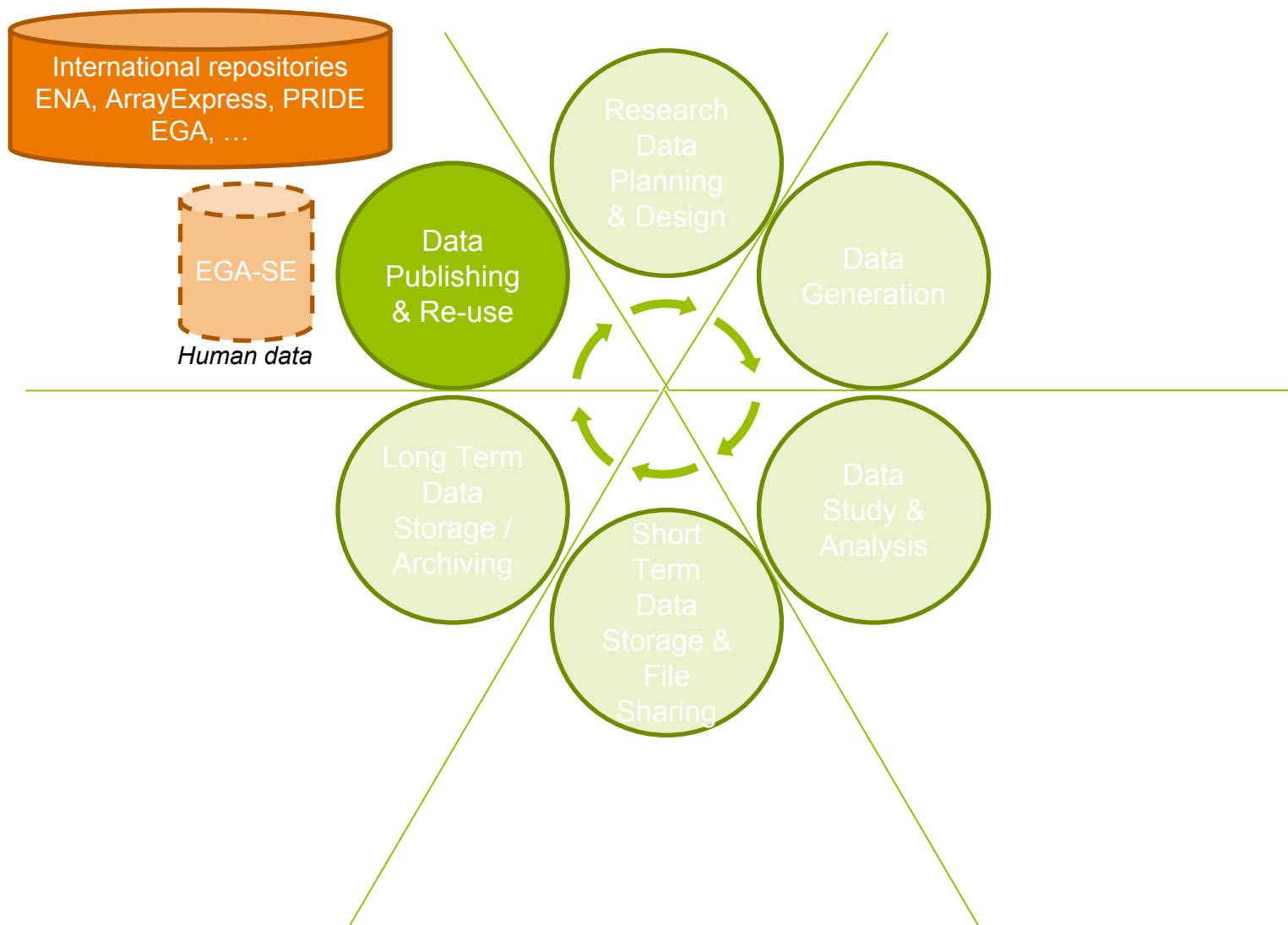
- Introduction to the tutorials
- The case study
- Setup
- For Mac / Linux users
- For Windows users
- The tutorials

- Conda
- Snakemake
- Git
- Jupyter
- R Markdown
- Docker
- Take down

Cloud services developers love

📖 Read the Docs





**Why should you make
research data available
for others?**

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



- 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



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



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

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

SUBJECT CATEGORIES

- » Research data
- » Publication characteristics

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 measurable 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

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

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

HANGZHOU, CHINA 4-5 SEPTEMBER

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



- *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 iD profile for Niclas Jareborg. The profile includes a navigation bar with options for researchers, organizations, about, help, and sign in. The user's name, ORCID ID, and a list of known names are displayed. The profile is divided into sections for Education (2) and Employment (7). The Education section lists two degrees from Uppsala University: a PhD in Microbiology (1989-05 to 1995-05) and a BSc in Microbiology (1985-01 to 1989-04). The Employment section lists two positions: Data Manager at Stockholms Universitet (2015-01 to present) and a position at Kungliga Tekniska Hogskolan (2013-01 to 2014-12).

ORCID
Connecting Research and Researchers

FOR RESEARCHERS | **FOR ORGANIZATIONS** | **ABOUT** | **HELP** | **SIGN IN**

SIGN IN | REGISTER FOR AN ORCID ID | LEARN MORE

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

Niclas Jareborg

ORCID ID
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) | Sort

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

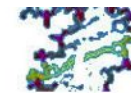
Uppsala Universitet: Uppsala, Sweden
1985-01 to 1989-04 (Microbiology)
BSc
Source: Niclas Jareborg | Created: 2015-04-09

Employment (7) | Sort

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)

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*

Study & Analysis

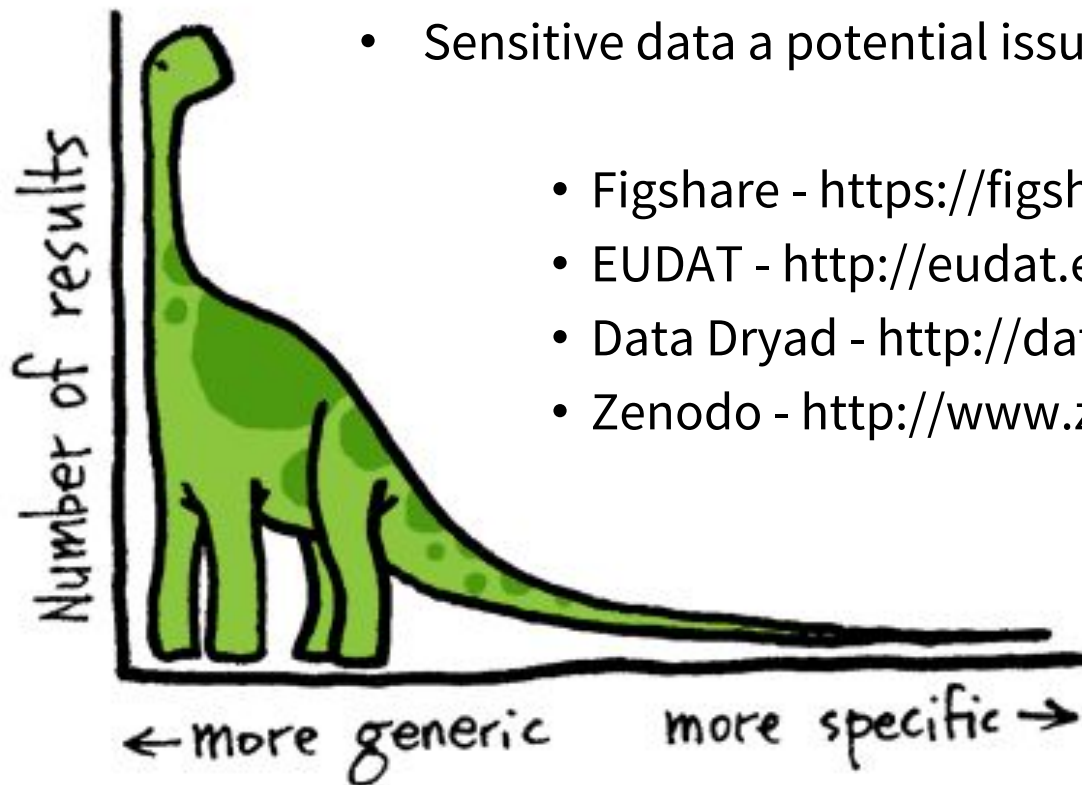


Deposition Database	Data type	International collaboration framework ¹
ArrayExpress	Functional genomics data. Stores data from high-throughput functional genomics experiments.	
BioModels	Computational models of biological processes.	
EGA	Personally identifiable genetic and phenotypic data resulting from biomedical research projects.	European Bioinformatics Institute and the Centre for Genomic Regulation
ENA	Nucleotide sequence information, covering raw sequencing data, contextual data, sequence assembly information and functional and taxonomic annotation.	International Nucleotide Sequence Database Collaboration
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.	

Deposition Database	Data type	International collaboration framework ¹
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
Support		
Pending incorporation into a Node Service Delivery Plan (see How countries join):		
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.	
EVA	The European Variation Archive covers genetic variation data from all species.	dbSNP and dbVAR
EMDB	The Electron Microscopy Data Bank is a public repository for electron microscopy density maps of macromolecular complexes and subcellular structures.	

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

- 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/>

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

Human Phenotype Ontology

Summary Classes Properties Notes Mappings Widgets

Jump To: All

- [-] Clinical modifier
- [-] Mode of inheritance
- [-] Monogenic/liging
- [-] 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 leukemia
 - [-] Chronic leukemia
 - [-] Lymphoid leukemia
 - [-] Myeloid leukemia
 - [-] Myeloproliferative disorder
 - [-] Lymphoma
 - [-] Lymphoproliferative disorder

Property	Value
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:D0015470 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)



Human Phenotype Ontology

Last updated: December 21, 2018



Summary **Classes** Properties Notes Mappings Widgets

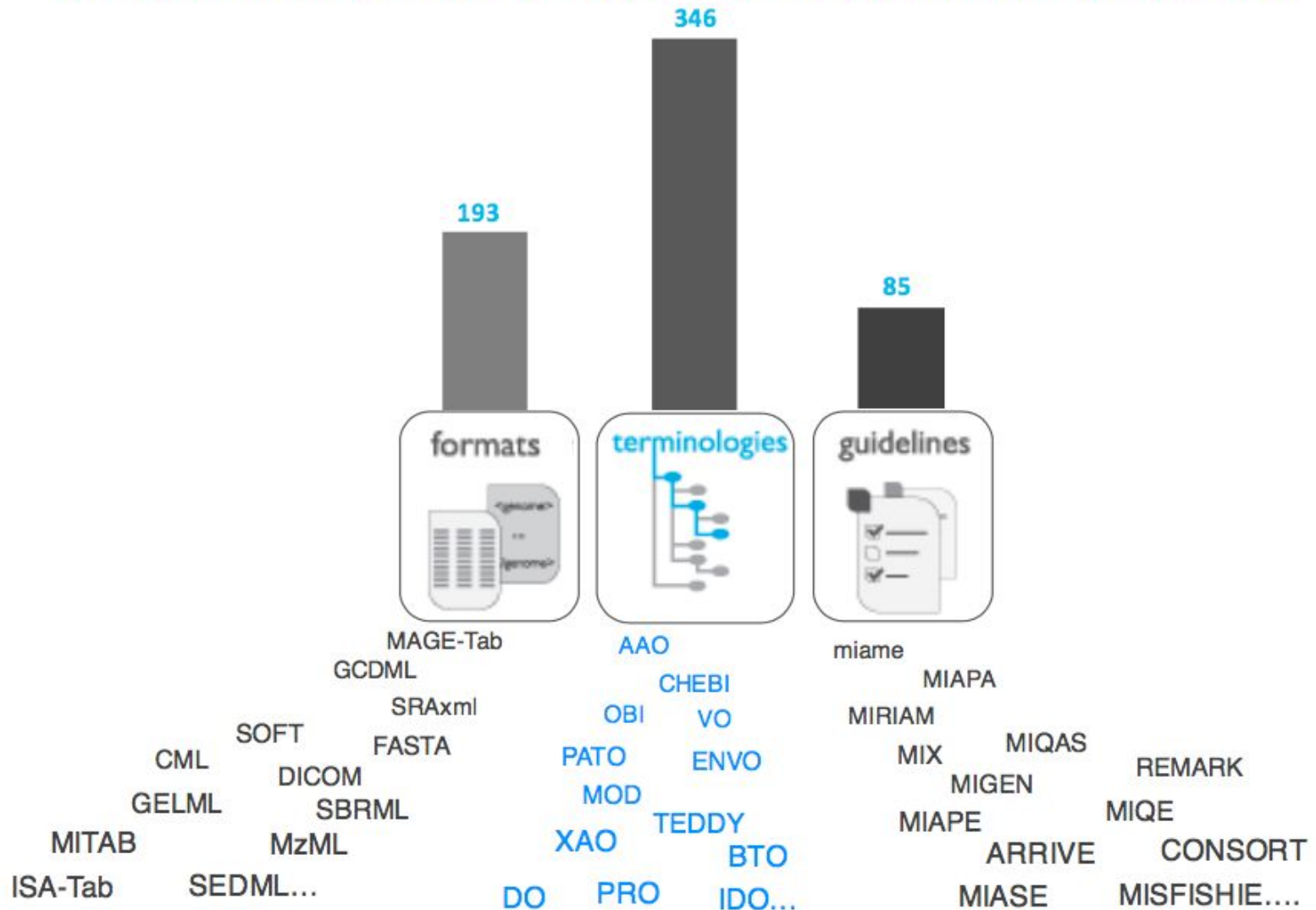
Jump to:

- [-] All
 - [-] Clinical course
 - [-] Clinical modifier
 - [-] Frequency
 - [-] Mode of inheritance
 - [-] Phenotypic abnormality
 - [-] Abnormal cellular phenotype
 - [-] Abnormality of blood and blood-forming tissues
 - [-] Abnormal bleeding
 - [-] Abnormal erythrocyte morphology
 - [-] Abnormal erythrocyte physiology
 - [-] Abnormal leukocyte morphology
 - [-] Abnormal myeloid cell morphology
 - [-] Abnormal thrombocyte morphology
 - [-] Abnormal thrombosis
 - [-] Arterial thrombosis
 - [-] Disseminated intravascular coagulation
 - [-] Peripheral thrombosis
 - [-] Thromboembolism
 - [-] Venous thrombosis
 - [-] Abnormality of bone marrow cell morphology
 - [-] Abnormality of coagulation
 - [-] Extramedullary hematopoiesis
 - [-] Hematological neoplasm
 - [-] Leukemia
 - [-] Acute leukemia
 - [-] Acute lymphoblastic leukemia
 - B Acute Lymphoblastic Leukemia**
 - [-] Ph-positive acute lymphoblastic leukemia
 - [-] T-cell acute lymphoblastic leukemias
 - [-] Acute megakaryocytic leukemia
 - [-] Acute monocytic leukemia
 - [-] Acute myeloid leukemia
 - [-] Acute myelomonocytic leukemia
 - [-] Acute promyelocytic leukemia
 - [-] Biphenotypic acute leukaemia

Details Visualization Notes (0) Class Mappings (7)

Preferred Name	B Acute Lymphoblastic Leukemia
Synonyms	Pre-B-cell acute lymphoblastic leukemia
Definitions	A type of ALL characterized by elevated levels of B-cell lymphoblasts in the bone marrow and the blood.
ID	http://purl.obolibrary.org/obo/HP_0004812
definition	A type of ALL characterized by elevated levels of B-cell lymphoblasts in the bone marrow and the blood.
has_exact_synonym	Pre-B-cell acute lymphoblastic leukemia
has_obo_namespace	human_phenotype
id	HP:0004812
label	B Acute Lymphoblastic Leukemia
notation	HP:0004812
prefLabel	B Acute Lymphoblastic Leukemia
treeView	Acute lymphoblastic leukemia
subClassOf	Acute lymphoblastic leukemia

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



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



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



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



- 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?*

- 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)

- **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/aktuell-projekt/dataskyddsfordning>

KI

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

KTH

<https://intra.kth.se/anstallning/anstallningsvillkor/att-va-ra-statligt-an/behandling-av-person/dataskyddsfordning-gdpr-1.800623>

LiU

<https://insidan.liu.se/dataskyddsfordningen/anmalan-av-personuppgiftsbehandling?l=sv>

LU

<https://personuppgifter.blogg.lu.se>

SU

<https://www.su.se/medarbetare/organisation-styrning/juridik/personuppgifter/dataskyddsf%C3%B6rordning>

UmU

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

UU

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

“As open as possible, as closed as necessary”

- **EGA** – European Genome-phenome Archive



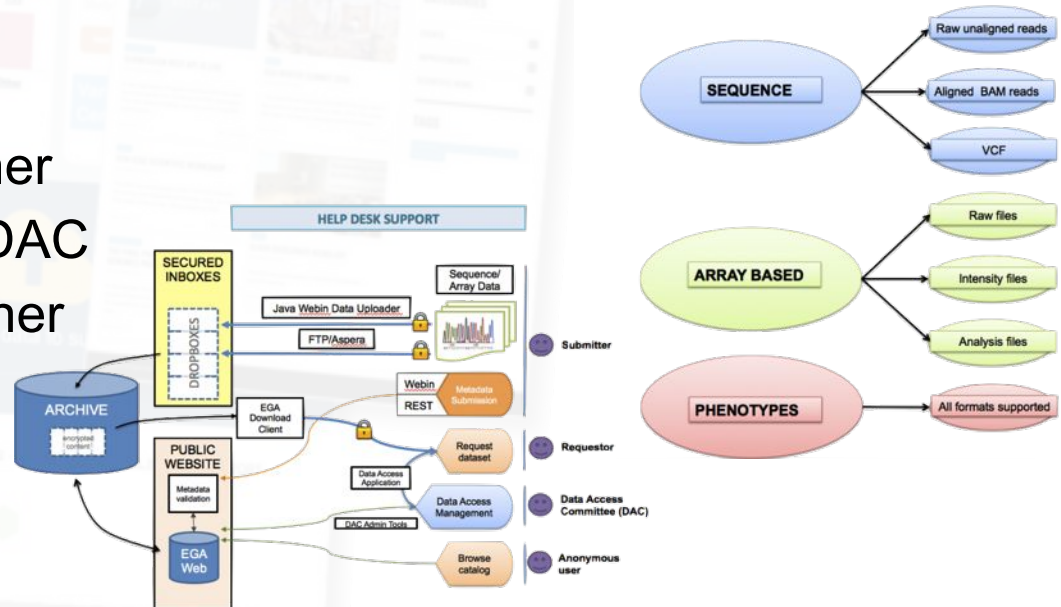
- Repository that promotes the distribution and sharing of **genetic and phenotypic data** consented for specific approved uses but **not fully open, public distribution.**
- All types of sequence and genotype experiments, including case-control, population, and family studies.

- Data Access Agreement

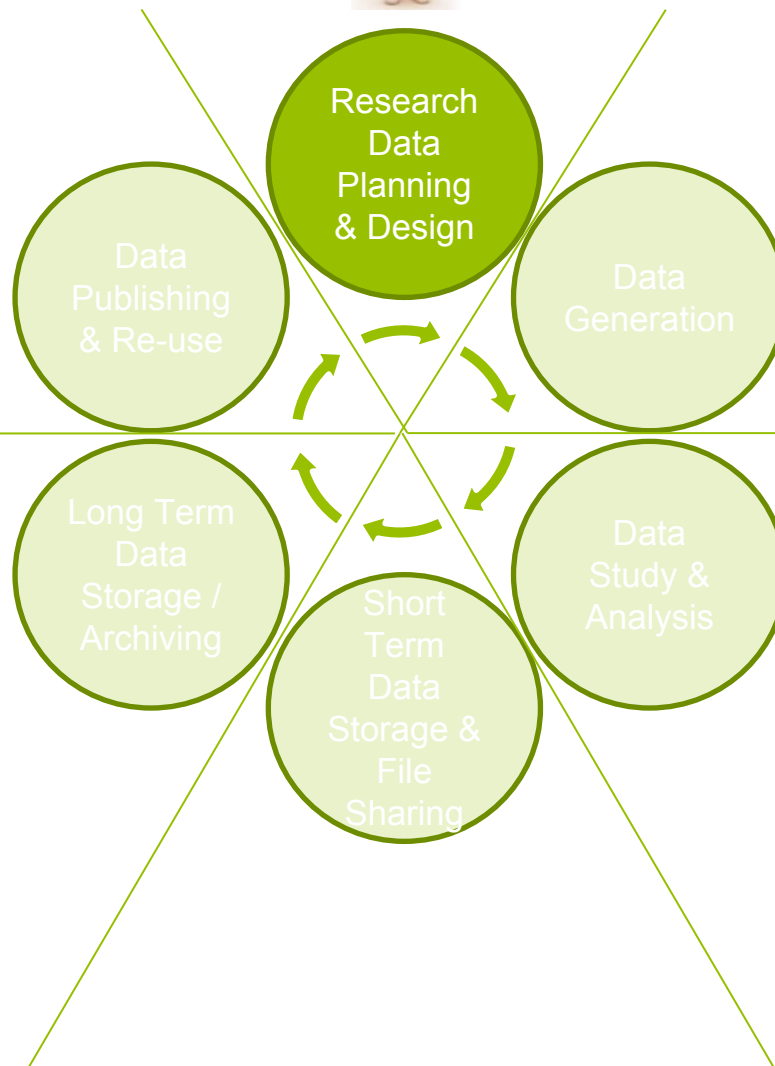
- Defined by the data owner

- Data Access Committee – DAC

- Decided by the data owner



**When should you start
thinking about how to
manage your data?**



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

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 administrating 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.

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

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

Project Details | Plan overview | Data Section Enabling Technologies Helix | Datamanagement ZonMw | Share | Download

expand all | collapse all

- 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 / 6)

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

ELIXIR Data Stewardship Wizard

DSW
DATA STEWARDSHIP WIZARD

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

Design of experiment

- Data design and planning
- Data Capture/Measurement
- Data processing and curation
- Data integration
- Data interpretation
- Information and insight

Is there any pre-existing data?

Are there any data sets available in the world that are relevant to your planned research?

No

Yes

Will you be using any pre-existing data (including other people's data)?

Will you be referring to any earlier measured data, reference data, or data that should be mined from existing literature? Your own data as well as data from others?

No

Yes

What reference data will you use?

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

- 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)

- 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
 - Define and **document** the *purpose* and *legal basis* for handling personal data in your project
- *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](https://doi.org/10.1371/journal.pcbi.1000424). 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 Wigge & 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/dataskyddsfordningen/>
 - ... and probably others I have forgotten