NBIS
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

# Data Management Solutions

nf-core hackathon

July 15 2020

Elin Kronander

elin.kronander@nbis.se

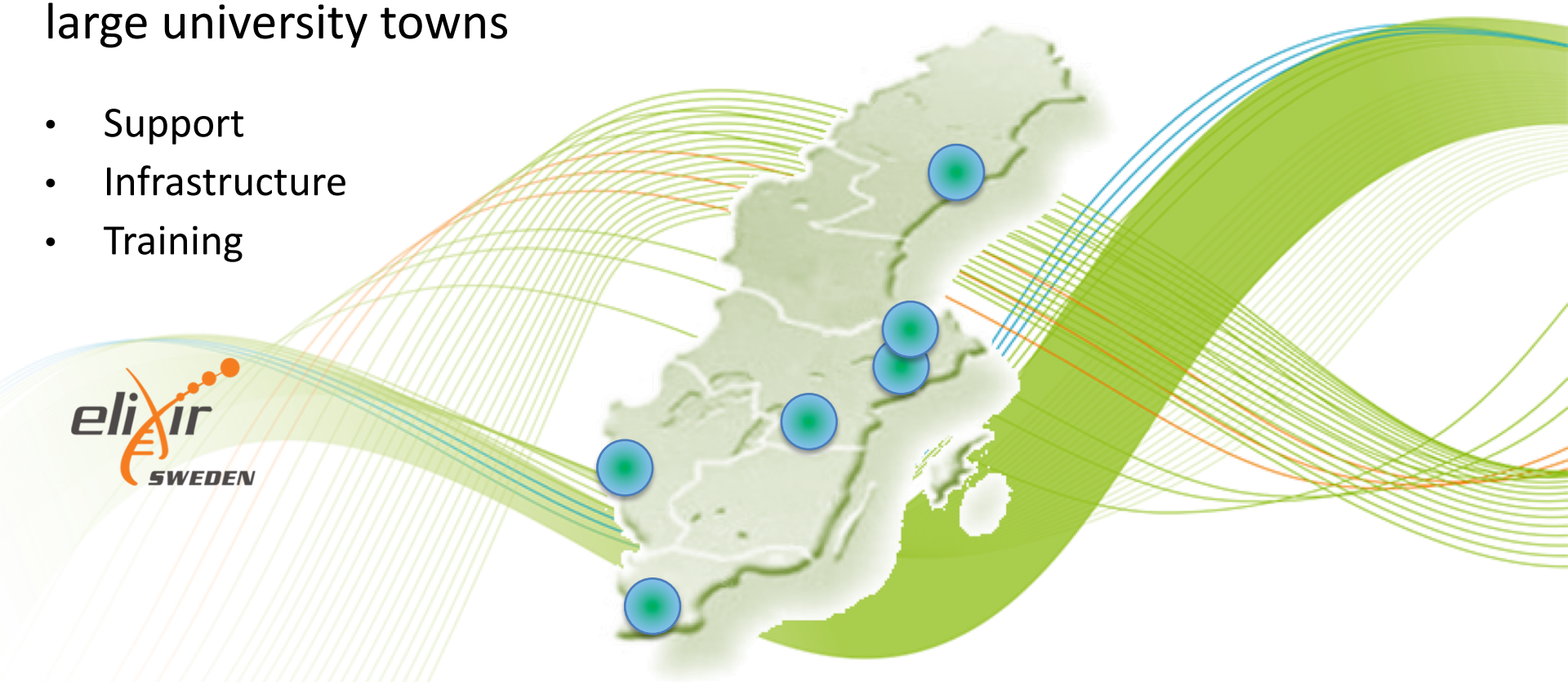
NBIS
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

# Bioinformatics Platform

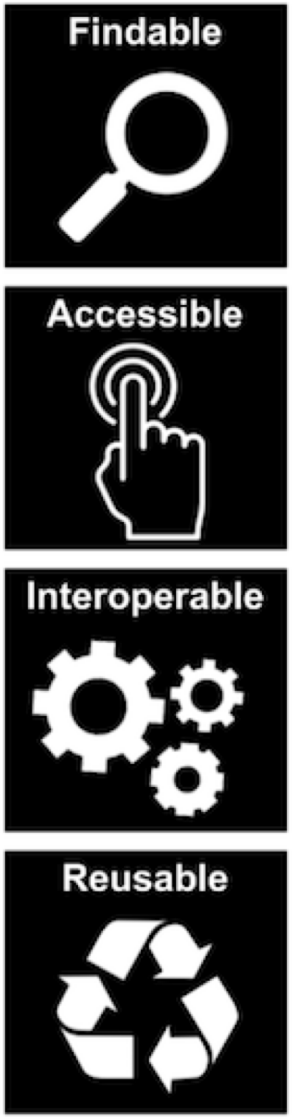
Distributed research infrastructure with nodes at each of the 6 large university towns

- Support
- Infrastructure
- Training



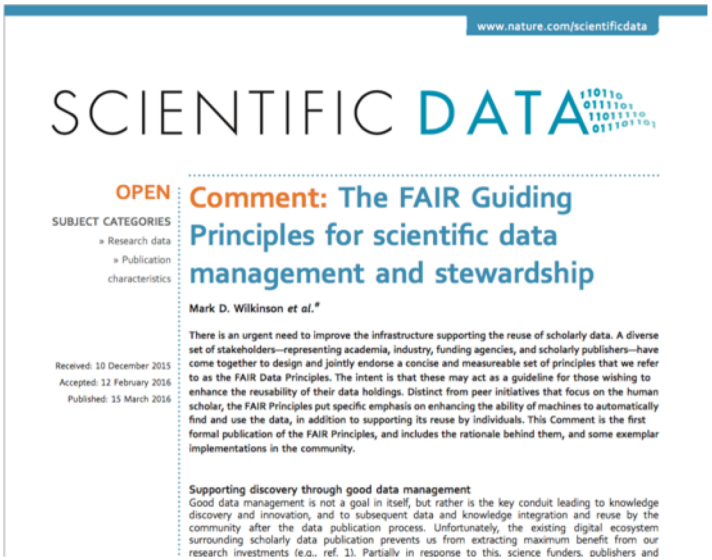
elixir SWEDEN

# FAIR Principles




To be useful for others data should be ***FAIR***

*… for both Machines and Humans*



www.nature.com/scientificdata

SCIENTIFIC DATA

OPEN

SUBJECT CATEGORIES
» Research data
» Publication characteristics

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

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

Mark D. Wilkinson *et al.*#

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

**Supporting discovery through good data management**

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

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

Image modified from www.ands.org.au

# FAIR Principles


Image Source: Australian National Data Service website (www.ands.org.au). Accessed 20 July 2020.

# International repositories











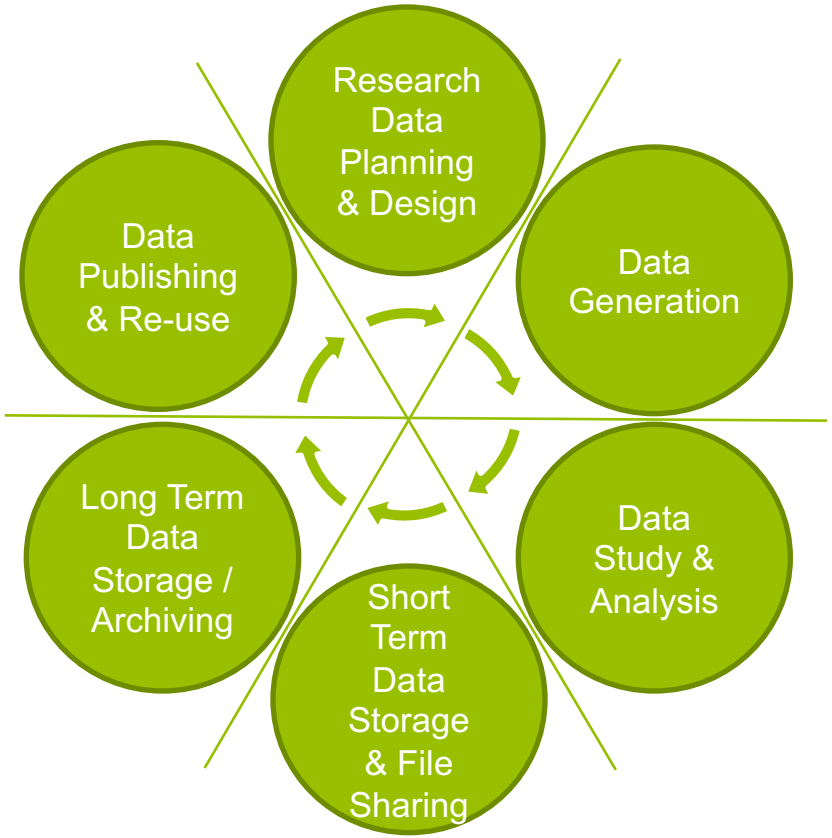

- Good way to make data **FAIR**
- Domain-specific metadata standards

Modified from slides created by Niclas Jareborg, Data Manager at NBIS

# What is research data?

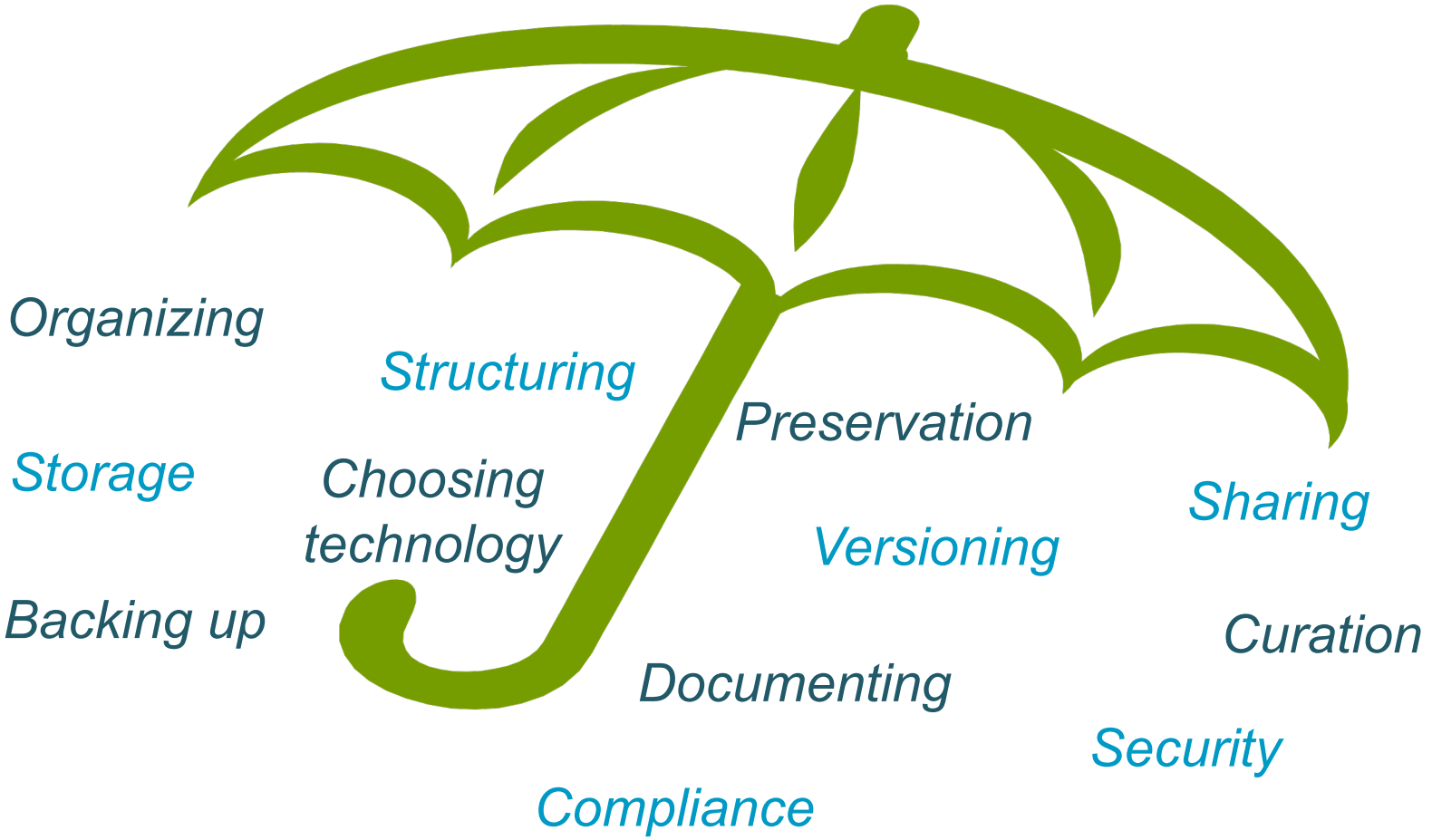
Any information you use in your research

The Research Data Life Cycle



# What is Data Management?

Purpose: Making the research process as efficient as possible




Modified from slides provided by the Research Support team, IT Service, University of Oxford

# Data Management Plan

- Revisable Document
- New & existing data
- Throughout data life cycle

### 1. Description of data

- What type of data will be generated
- From which type of samples will the data be generated
- What additional data will be collected

### 2. Documentation and data quality

- How will the analysis be documented
- Which metadata standards will be used
- What quality measures will be used
- Which file formats will be used
- What is your strategy concerning versioning

### 3. Storage and Backup

- How are the data stored?
- Are there back-up systems
- What is the estimated size of data
- Do you need to restrict access to data

### 4. Legal and ethical aspects

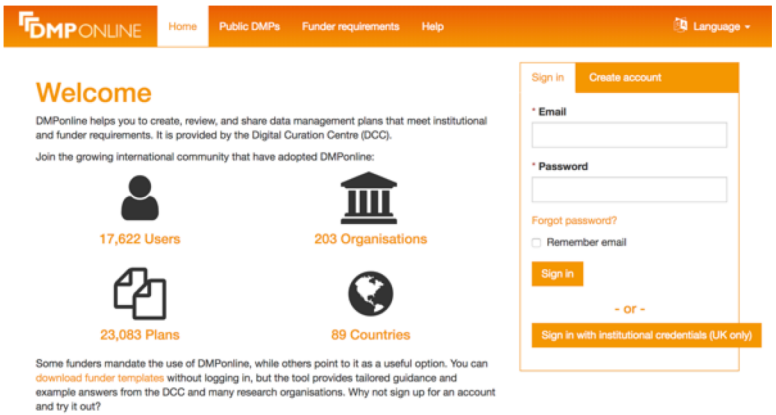
- Is sensitive human data part of your project
- Are there agreements in place with other stakeholders

### 5. Accessibility and long-term storage

- How and where will the data be shared
- How and where will the data be stored after the project's completion

# DMP tools

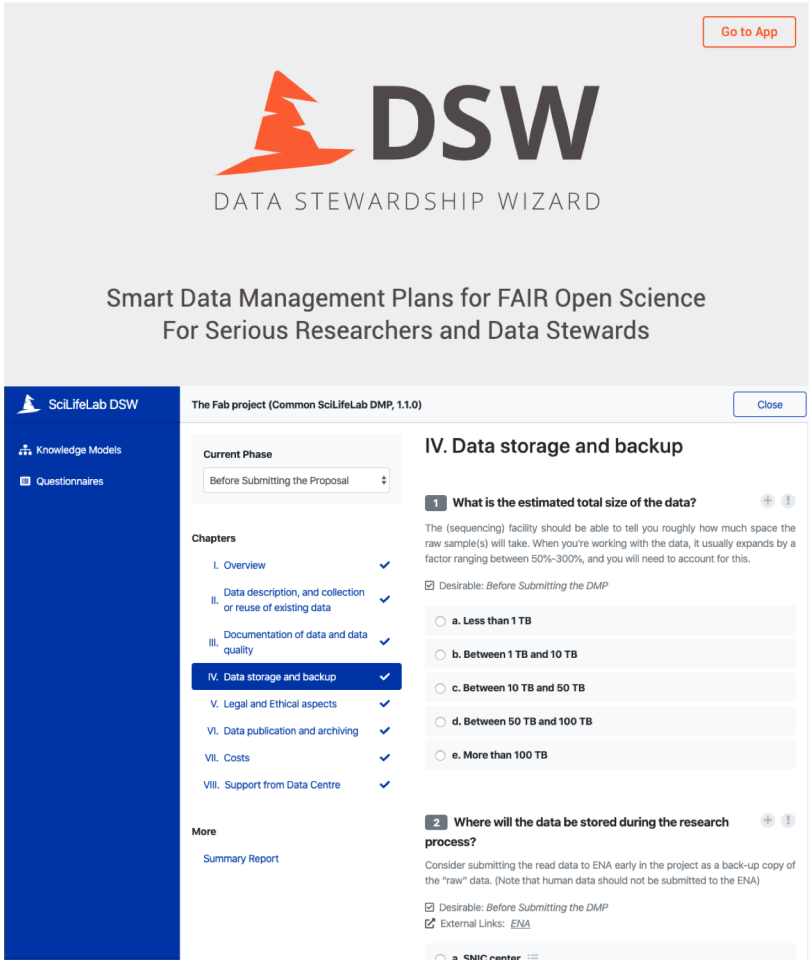
## DMPonline





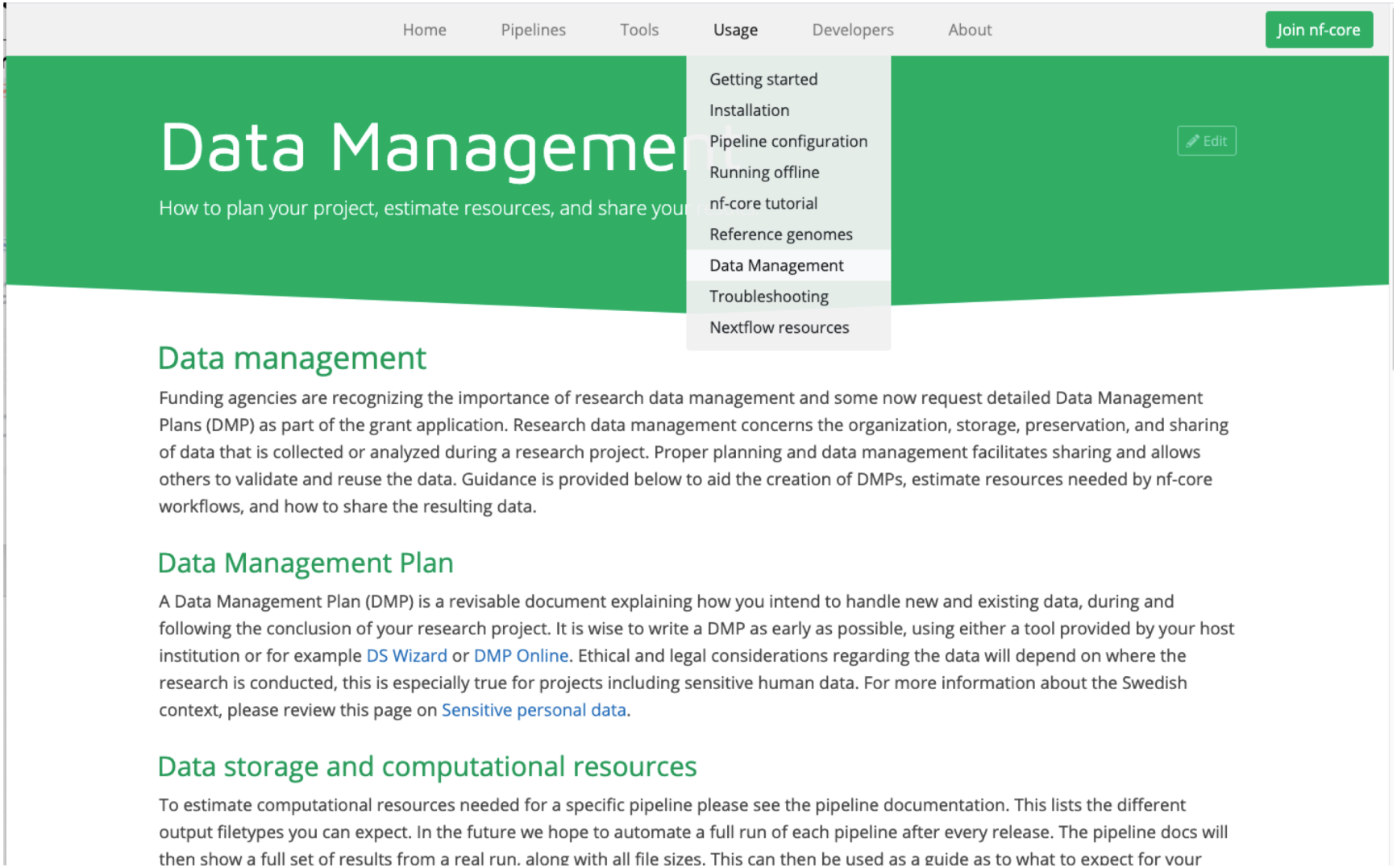
https://dmponline.dcc.ac.uk/

## ELIXIR Data Stewardship Wizard




https://ds-wizard.org/ https://dsw.scilifelab.se

Modified from slides created by Niclas Jareborg, Data Manager at NBIS


Home
Pipelines
Tools
Usage
Developers
About
Join nf-core
Getting started
Installation
Pipeline configuration
Running offline
nf-core tutorial
Reference genomes
Data Management
Troubleshooting
Nextflow resources
Edit
Data Managemen
How to plan your project, estimate resources, and share you
Data management
Funding agencies are recognizing the importance of research data management and some now request detailed Data Management Plans (DMP) as part of the grant application. Research data management concerns the organization, storage, preservation, and sharing of data that is collected or analyzed during a research project. Proper planning and data management facilitates sharing and allows others to validate and reuse the data. Guidance is provided below to aid the creation of DMPs, estimate resources needed by nf-core workflows, and how to share the resulting data.
Data Management Plan
A Data Management Plan (DMP) is a revisable document explaining how you intend to handle new and existing data, during and following the conclusion of your research project. It is wise to write a DMP as early as possible, using either a tool provided by your host institution or for example DS Wizard or DMP Online. Ethical and legal considerations regarding the data will depend on where the research is conducted, this is especially true for projects including sensitive human data. For more information about the Swedish context, please review this page on Sensitive personal data.
Data storage and computational resources
To estimate computational resources needed for a specific pipeline please see the pipeline documentation. This lists the different output filetypes you can expect. In the future we hope to automate a full run of each pipeline after every release. The pipeline docs will then show a full set of results from a real run, along with all file sizes. This can then be used as a guide as to what to expect for your

# Data Sharing exercise

1. Select a dataset or a project you are working on 
2. Identify a suitable repository 
3. Check the repository guidelines for: 
   - file formats
     - raw data
     - processed data
   - metadata
   - recommended vocabularies
   - quality measures
   - licensing
4. Prepare your data to conform to the repository guidelines 

# Data Sharing exercise

5. Submit your raw data 

   - easier to provide metadata
   - ready for publication
   - backup in different location
   - increase citations

   - processed data can be linked later
   - use embargo if necessary

# Choosing a repository

Does the data contain **personal** or **sensitive** information?

**YES** → 1. Controlled access repositories

**NO** → Is there a **discipline-specific** repository for your data?

**YES** → 2. Discipline-specific repositories

**NO** → Is there a **institutional** repository for your data?

**YES** → 3. Institutional repositories

**NO** → 4. General data repositories

ENA European Nucleotide Archive

BioSamples

EUROPEAN GENOME-PHENOME ARCHIVE

NCBI

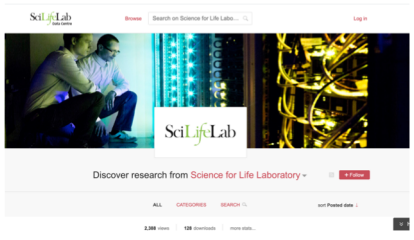
dbSNP Short Genetic Variations

PRIDE

Protein Data Bank in Europe
Bringing Structure to Biology

SND Svensk nationell datatjänst

figshare




zenodo

EUDAT

Etc...

Modified from slides created by Niclas Jareborg, Data Manager at NBIS

# Metadata example from ENA

The following metadata fields are supported in the manifest file:

STUDY: Study accession or unique name (alias)

SAMPLE: Sample accession or unique name (alias)

NAME: Unique experiment name

PLATFORM: See permitted values. Not needed if INSTRUMENT is provided.

INSTRUMENT: See permitted values

INSERT_SIZE: Insert size for paired reads

LIBRARY_NAME: Library name (optional)

LIBRARY_SOURCE: See permitted values

LIBRARY_SELECTION: See permitted values

LIBRARY_STRATEGY: See permitted values

DESCRIPTION: free text library description (optional)

# Metadata example from ENA

**Permitted values for library source**

GENOMIC: Genomic DNA (includes PCR products from genomic DNA).

GENOMIC SINGLE CELL:

TRANSCRIPTOMIC: Transcription products or non genomic DNA (EST, cDNA, RT-PCR, screened libraries).

TRANSCRIPTOMIC SINGLE CELL:

METAGENOMIC: Mixed material from metagenome.

METATRANSCRIPTOMIC: Transcription products from community targets

SYNTHETIC: Synthetic DNA.

VIRAL RNA: Viral RNA.

OTHER: Other, unspecified, or unknown library source material.

# Getting started

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

Life Science Data Management lecture by Niclas Jareborg
https://doi.org/10.17044/scilifelab.12057978.v1

*"Your primary collaborator is yourself six months from now, and your past self doesn't answer e-mails,"*

-Rachael Ainsworth