

Data Management for Bioinformatics

NBIS Data Management Team
data-management@scilifelab.se

NBIS Advisory Programme Grand Meeting
Uppsala, Sweden
25 May 2023

Managing Research Data

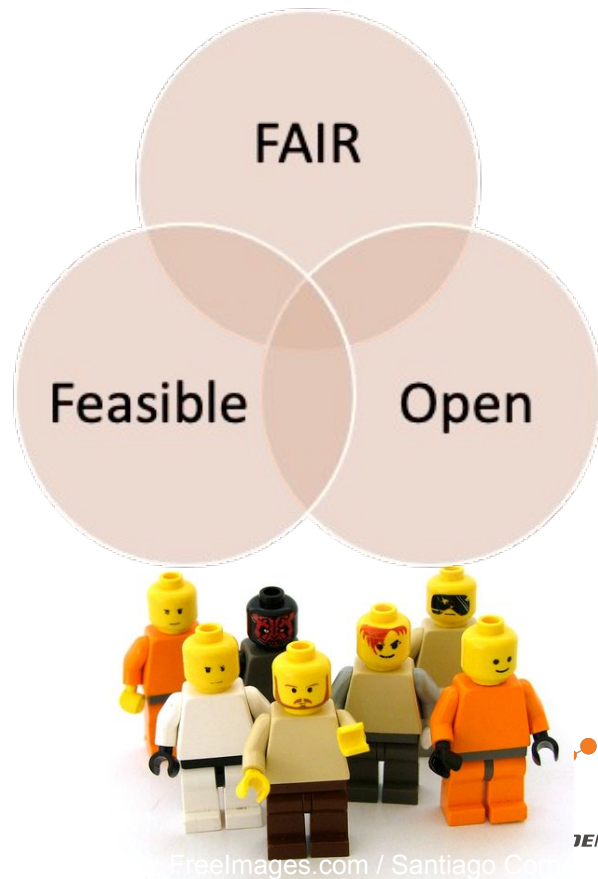


Illustration from Digitalbevaring.dk / Jørgen Stamp (CC BY 2.5 Denmark license).

- Research **data is a core component** of any research project or publication.
- Good data management practices are **important in all phases** of research
 - Ethics and legislation
 - Information security
 - Research documentation
 - Project organisation
- Research data needs to be secured **beyond the project's** time frame



- Funding agencies increasingly require data management plans to improve the impact of the distributed funds
- **Reduce risks of data loss** and scientific misconduct by securing adequate resources and skills for the project
- Improve impact in society and future research by **facilitating reuse and verification of results**
- Transparency & community of best practice



Optimise your practices for reuse



- Make project more efficient by implementing **good practices for handling research data**
- Establish procedures to **address all aspects of data management** throughout the data life cycle
- Adopt best-practice guidelines that encourage **Reproducible Research, Open Science & FAIR data principles**

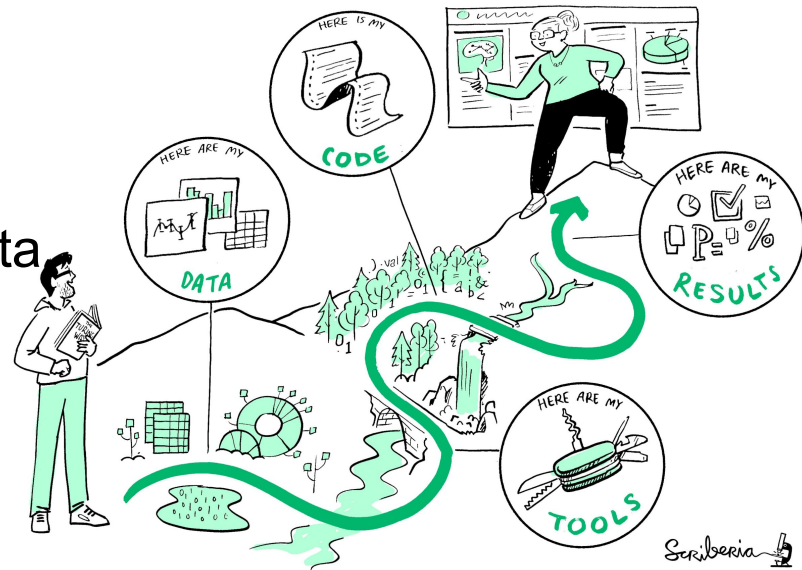


Best-practice Guidelines



Illustration: This image was created by Scriberia for The Turing Way community and is used under a CC-BY licence.

- **FAIR data principles** – make your data Findable, Accessible, Interoperable & Reusable – *as open as possible and as closed as necessary*
- **Reproducible Research** – ensure that results can be reproduced using the data code, and documentation provided
- **Open Science** – aim for unlimited, barrier free, open access to research outputs and transparency in the whole research cycle





- ❑ **Secure/organise data & analyses**, by managing back-ups, access restrictions, versioning, docs, scripts and transcripts
- ❑ **Deposit and share data** using restricted or public access data repositories that promote FAIR data principles
- ❑ **Adhere to community standards**, such as file formats, data dictionaries, controlled vocabularies and metadata
- ❑ **Maintain a Data Management Plan**, outlining the project's data management practices

Team vision

“Swedish Life Science researchers **apply good Data Management practices** so that the **research outputs** produced are **available** to the global research community, and to society at large, according to the principles of *Open Science, Reproducible Research, and FAIR*”

Team mission

Support – Training – Collaboration



Markus Englund
Data Steward



Erik Hedman
Data Steward



Yvonne Kallberg
Data Steward



Elin Kronander
Data Steward



Stephan Nylinder
Data Steward



Mattias Strömberg
Data Steward



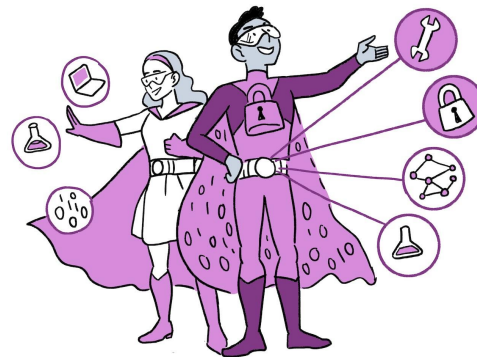
**Wolmar N.
Åkerström**
Data Steward



Niclas Jareborg
Data Manager



- Guide writing a data management plan
- Identify a suitable repository for publishing data
- Assist during the submission process when publishing data and code
- Advice on describing data with proper metadata for documentation and publishing
- Advice on what needs to be done when working with sensitive human data
- Data transfers, data organisation, backup, and security procedures



Scribbleria

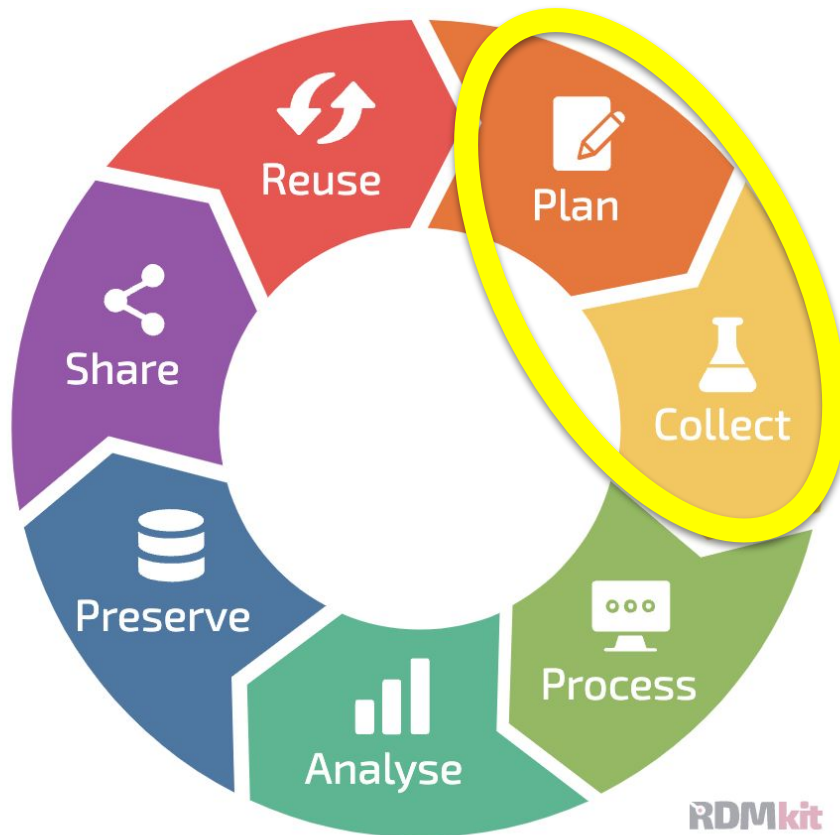
Contact us

- nbis.se/support/supportform
- data-management@scilifelab.se



<https://data-guidelines.scilifelab.se/> data-management@scilifelab.se

- **Get Support**—for anyone involved in life science research that is affiliated with a Swedish university or research institute.
- **Events & Training**—upcoming conferences, webinars, workshops, and training opportunities.
- **Topics**—covering the entire data life cycle with link to relevant resources.



Examples of other resources



Overview of good data management practices



The Research Data Management Kit (RDMkit) guides you through the whole data management life cycle and includes advice specific to your domain, your role and your country.

Step-by-step instructions



The FAIR Cookbook contains step-by-step recipes to accomplish specific data management tasks and to make your data FAIR (Findable, Accessible, Interoperable, Reusable).

Data management plan wizard



The Data Stewardship Wizard (DSW) is an online tool that guides researchers and data stewards through their data management planning.

The Research Data Management toolkit for Life Sciences

Best practices and guidelines to help you make your data FAIR (Findable, Accessible, Interoperable and Reusable)

What can we help you find?

Search RDMkit

Browse all topics by



Data life cycle

Start here to get an overview of research data management based on stages in the data life cycle.



Your role

Identify your role in research data management, find data management resources relevant for you, and information to help you progress in your career path.



Your domain

Learn about data management tasks that affect your domain or research community, and the solutions adopted to address them.



Your tasks

Find guidelines and solutions for tackling common data management tasks.



Tool assembly

Find concrete combinations of tools and resources assembled into an ecosystem for research data management.



National resources

Find pointers to country specific information resources and national research data management practices.



All tools and resources

Browse the RDMkit's catalogue of tools and resources for research data management.



All training resources

Browse all training resources mentioned in RDMkit pages.

The Research Data Management toolkit for Life Sciences

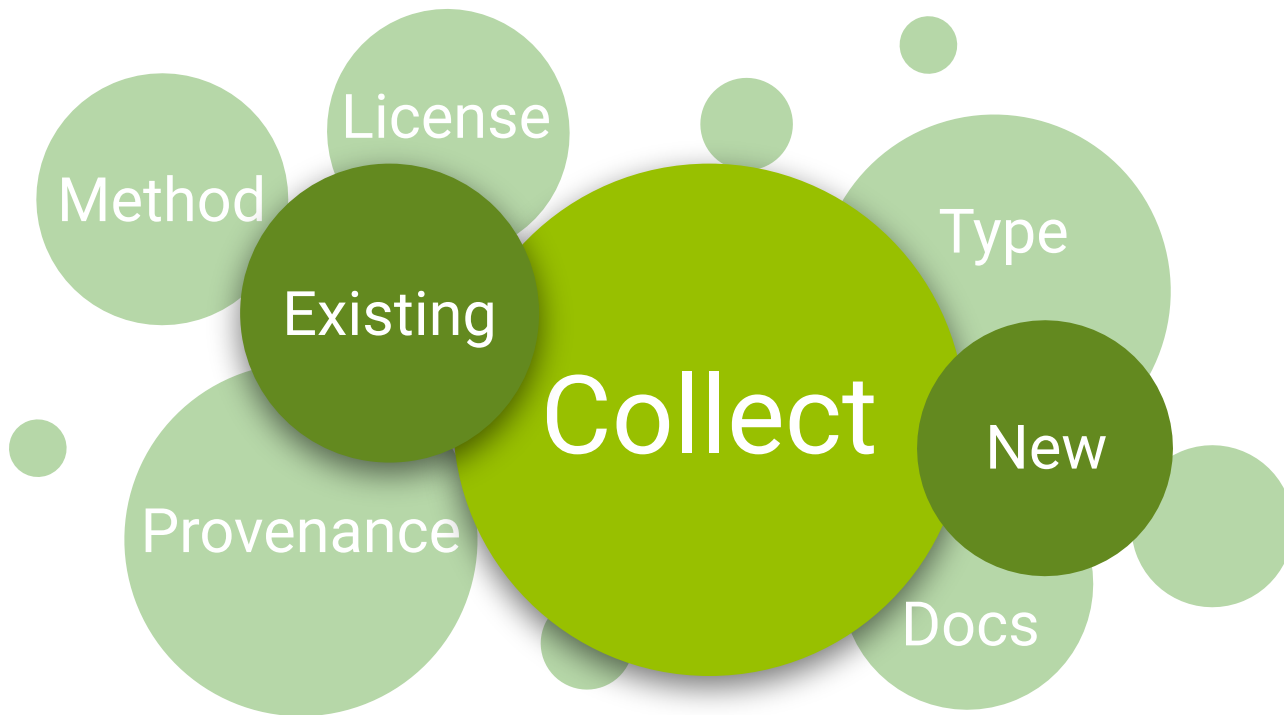
- RDM best practices and guidelines
- Links to tools/resources and training material given in specific DM context
- Examples of combination of tools for RDM

<https://rdmkit.elixir-europe.org/>

A Data Management Teaser



"Ready for BioData Management?" Training Data Stewards for Life Sciences: Intro Course, hosted by BioData.pt | ELIXIR PT, <https://doi.org/10.5281/zenodo.6599749>





- What is data collection?
- Why is data collection important?
- What should be considered for data collection?
- What are the related tasks?
- Where can training materials and events about data collection be found?

Contribute



Your domain

Learn about data management tasks that affect your domain or research community, and the solutions adopted to address them.



National resources

Find pointers to country specific information resources and national research data management practices.

Your domain ^

Bioimaging data

Biomolecular simulation data

Epitranscriptome data

Human data

Intrinsically disordered proteins

Marine metagenomics

Microbial biotechnology

Plant sciences

Proteomics

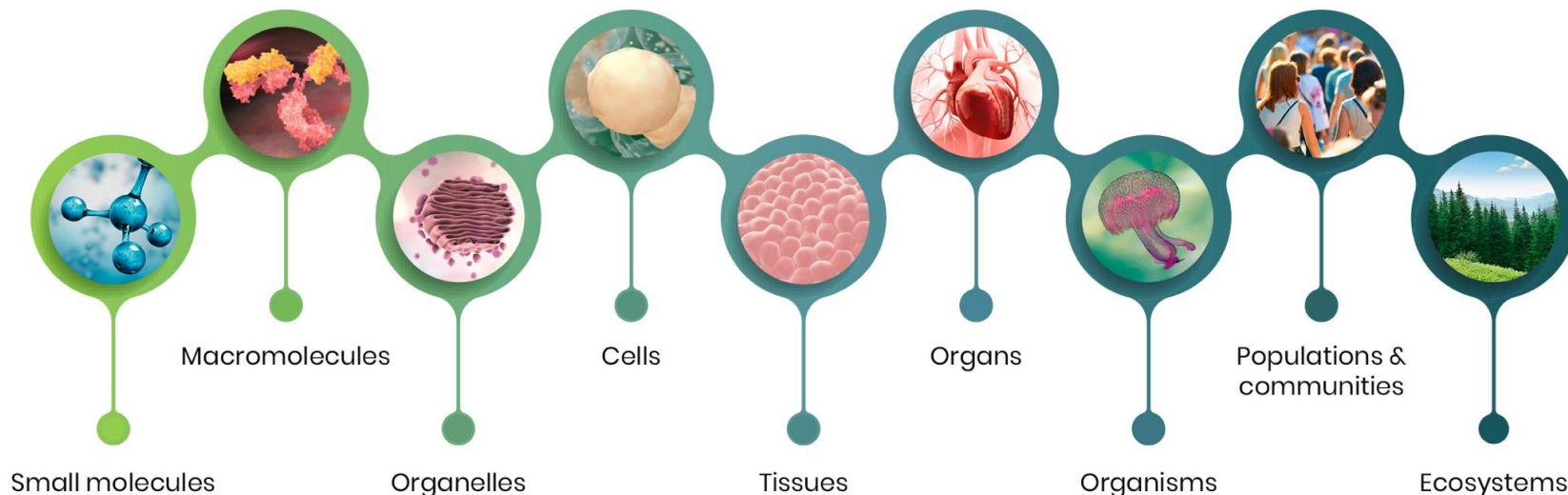
Rare disease data

Structural bioinformatics

Toxicology data



SciLifeLab generates a variety of data





National Genomics Infrastructure (NGI)

The National Genomics Infrastructure (NGI) provides services for next generation sequencing and SNP genotyping on all scales using a (...)

Ancient DNA

Use cleanroom labs and specialized molecular genetics techniques to extract, make libraries, sequence and analyze DNA in ancient and/or (...)

Clinical Genomics

Develops and provides clinical genetic tests using state-of-the-art genomic methods, such as next-generation sequencing, for (...)

Eukaryotic Single Cell Genomics

Provides service for high-throughput single cell genomics analysis

Microbial Single Cell Genomics

Provides streamlined single-cell sorting, lysis, whole-genome amplification and screening of individual microbial cells, as well as (...)

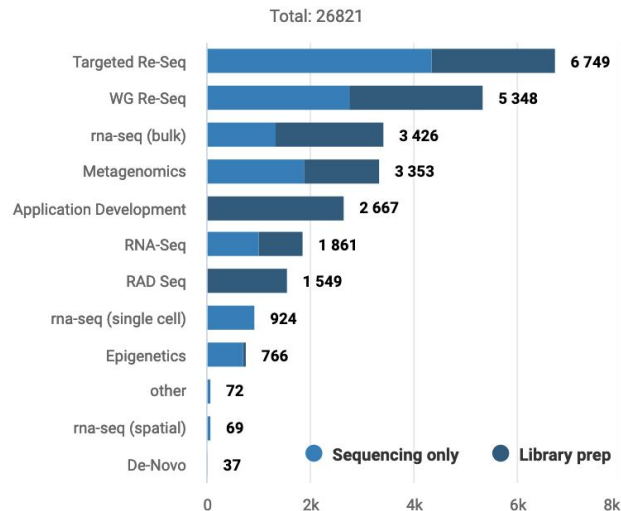
National Bioinformatics Infrastructure (NBIS)

Provides custom-tailored support with data analysis, computational tools, systems development and training.

- Throughputs 1 326 Gbp per day,
1 human genome equivalent per 3.51 min,
and 26 821 samples in 2022
- Instruments include *Ion Torrent*, *Illumina*,
Pacific Biosciences, and *Oxford Nanopore*
- Automated quality checks and analyses in a
number of applications
- Samples must be prepared to fulfill the
requirements for the method or kit



Samples in 2022

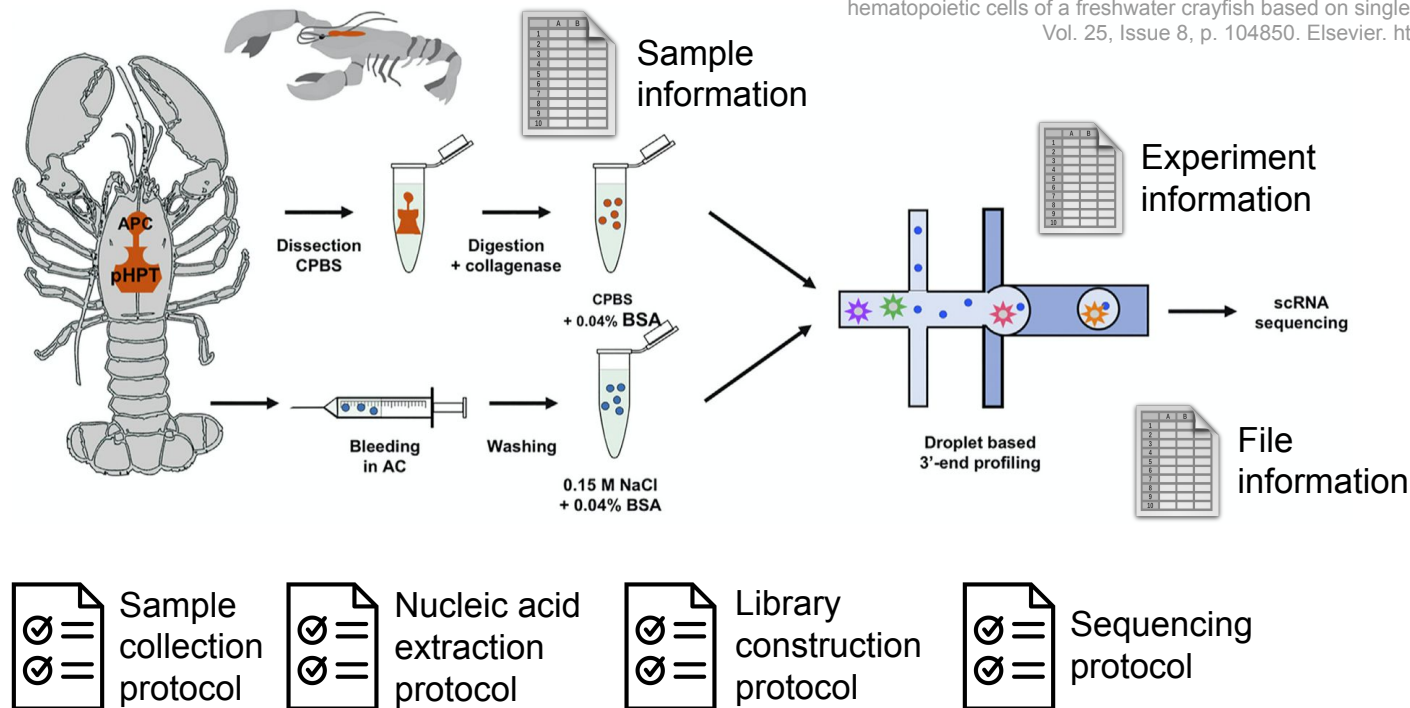


Single-cell RNA sequencing example



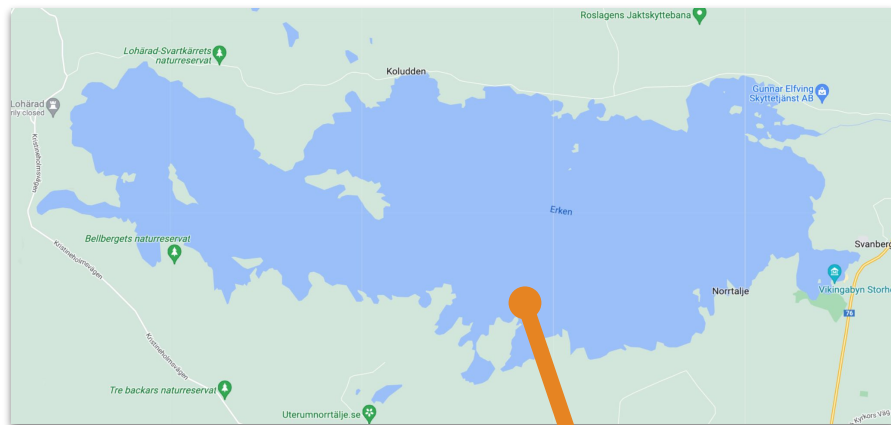
"Protocol" icon by Justin Blake from thenounproject.com

Söderhäll, I., Fasterius, E., Ekblom, C., & Söderhäll, K. (2022). Characterization of hemocytes and hematopoietic cells of a freshwater crayfish based on single-cell transcriptome analysis. In *iScience* Vol. 25, Issue 8, p. 104850. Elsevier. <https://doi.org/10.1016/j.isci.2022.104850>



- checksums.md5
- SampleSheet.csv
- ▼ SI-GA-F2_1
 - TJ-2700-1_S1_L001_R1_001.fastq.gz
 - TJ-2700-1_S1_L001_R2_001.fastq.gz
 - TJ-2700-1_S1_L002_R1_001.fastq.gz
 - TJ-2700-1_S1_L002_R2_001.fastq.gz
- ▼ SI-GA-F2_2
 - TJ-2700-1_S2_L001_R1_001.fastq.gz
 - TJ-2700-1_S2_L001_R2_001.fastq.gz
 - TJ-2700-1_S2_L002_R1_001.fastq.gz
 - TJ-2700-1_S2_L002_R2_001.fastq.gz
- ▼ SI-GA-F2_3
 - TJ-2700-1_S3_L001_R1_001.fastq.gz
 - TJ-2700-1_S3_L001_R2_001.fastq.gz
 - TJ-2700-1_S3_L002_R1_001.fastq.gz
 - TJ-2700-1_S3_L002_R2_001.fastq.gz
- ▼ SI-GA-F2_4
 - TJ-2700-1_S4_L001_R1_001.fastq.gz
 - TJ-2700-1_S4_L001_R2_001.fastq.gz
 - TJ-2700-1_S4_L002_R1_001.fastq.gz
 - TJ-2700-1_S4_L002_R2_001.fastq.gz

- Obtain freshwater crayfish (adult males) from lake Erken, Sweden (59.8 N 18.6 E)
- Maintain in the crayfish facility in running tap water, 10-12°C, 12:12 light:dark cycle...
- Feed once a week



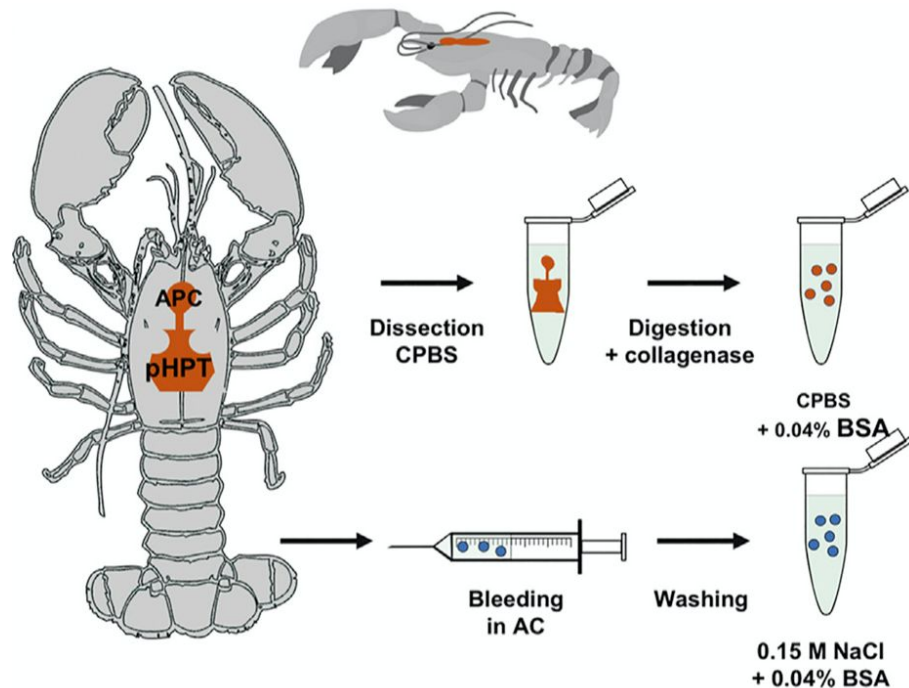
Map data © Google



Nucleic acid extraction protocol



Protocol illustration derived from Illustration in Söderhäll et al. (2022).



- Dissect and digest into single cells by incubation in 300 μ l of 0.1% collagenase ... at room temperature for 20 min on a rotating plate ... then filtered through a 40 mm cell strainer
- Pool isolated cells from four animals for scRNA-seq

Sample information as data



ENA Checklist: ERC00001 – ENA default sample checklist

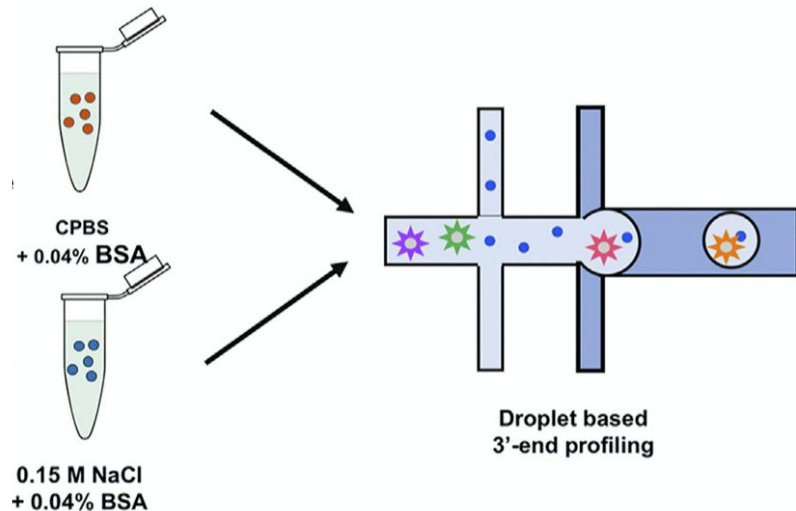
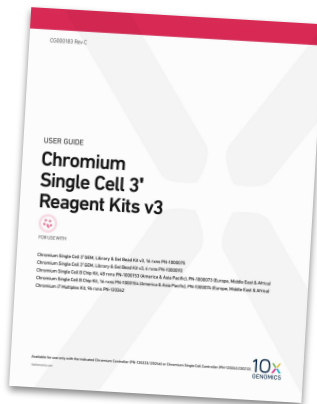
OLS / Experimental Factor Ontology

EFO

Population	Organism	<i>Pacifastacus leniusculus</i>
	Genotype	wild type genotype
	Geographical location	Sweden, Lake Erken
Individual	Growth condition	laboratory aquarium since Sep 2020
	Sampling date	2020-11-06
	Developmental stage	adult
	Body weight	35
	Sex	male
Specimen	Organism part	hematopoietic system
Sample	Cell type	hemocyte

- Prepare sequencing libraries using Chromium Single Cell 3' reagent kit v3 (cat# 1000075/1000073/120262, 10xGenomics)
- According to the manufacturer's protocol

CG000183
Single Cell 3' Reagent Kit
User Guide, v3 chemistry,
10xGenomics



Sequencing protocol



- 28+8+0+91 bp read length
- NovaSeq 6000 system
- SP flowcell
- v1 sequencing chemistry
- Include a sequencing library for the phage PhiX as 1% spike-in in the sequencing run

- ▼ SI-GA-F2_1
 - TJ-2700-1_S1_L001_R1_001.fastq.gz
 - TJ-2700-1_S1_L001_R2_001.fastq.gz
 - TJ-2700-1_S1_L002_R1_001.fastq.gz
 - TJ-2700-1_S1_L002_R2_001.fastq.gz
- ▼ SI-GA-F2_2
 - TJ-2700-1_S2_L001_R1_001.fastq.gz
 - TJ-2700-1_S2_L001_R2_001.fastq.gz
 - TJ-2700-1_S2_L002_R1_001.fastq.gz
 - TJ-2700-1_S2_L002_R2_001.fastq.gz
- ▼ SI-GA-F2_3
 - TJ-2700-1_S3_L001_R1_001.fastq.gz
 - TJ-2700-1_S3_L001_R2_001.fastq.gz
 - TJ-2700-1_S3_L002_R1_001.fastq.gz
 - TJ-2700-1_S3_L002_R2_001.fastq.gz
- ▼ SI-GA-F2_4
 - TJ-2700-1_S4_L001_R1_001.fastq.gz
 - TJ-2700-1_S4_L001_R2_001.fastq.gz
 - TJ-2700-1_S4_L002_R1_001.fastq.gz
 - TJ-2700-1_S4_L002_R2_001.fastq.gz

File Formats for Illumina Sequencing

Numerous options are available for converting data to compatible sequence file formats such as FASTQ files, and for downstream analysis of sequencing data. Illumina sequencers are designed so data can be easily streamed into Illumina Connected Analytics and BaseSpace Sequence Hub for cloud-based data management, analysis, and collaboration.

Raw data files are provided in sequence file formats that are compatible, or easily converted, to standardized data formats for streamlined aggregation and mining of large cohorts. With the DRAGEN BioT platform, the newest file format, FASTQ-ORA, is available. FASTQ-ORA is a lossless compression file reducing the size, time to transfer, and storage cost.

FASTQ Sequence File Format

FASTQ is a text-based sequencing data file format that stores both raw sequence data and quality scores. FASTQ files have become the standard format for storing NGS data from Illumina sequencing systems, and can be used as input for a wide variety of secondary data analysis solutions.

The MinSeq and MiSeq Sequencing Systems provide the option to automatically convert data from BCL to FASTQ format, so separate conversion software is not required.

[Learn More About FASTQ Files](#)

FASTQ-ORA Sequence File Format

FASTQ-ORA is a binary compressed file format of the text-based FASTQ sequencing data file format. fastq.ora files are up to 5x smaller than their corresponding fastq.gz files without compromising data integrity. All fastq.ora files can be read using the free decompression software available [here](#). Once installed, a simple command can be used to directly pipe the output of decompression on the fly into a wide range of popular mapping tools such as BWA, STAR, and Bowtie. DRAGEN-ORA compression is available with the DRAGEN server and on-board the NextSeq1000/2000.

Chromium Single Cell 3' Reagent Kits v3

FOR USE WITH

Chromium Single Cell 3' GEM Library & Gel Bead Kit v3, 16 rxns PN-1000075
Chromium Single Cell 2' GEM Library & Gel Bead Kit v3, 4 rxns PN-1000072
Chromium Single Cell 2' Chip Kit, 48 rxns PN-1000153 America & Asia Pacific, PN-1000073 Europe, ME
Chromium Single Cell 2' Chip Kit, 16 rxns PN-1000154 America & Asia Pacific, PN-1000074 Europe, ME
Chromium 2' Multiplex Kit, 16 rxns PN-120262

MultiQC v1.8

Report for project TJ-2700 on runfolder 201126_A00605_0172_BHVVTNDRXX

Report for project TJ-2700 on runfolder 201126_A00605_0172_BHVVTNDRXX

NGI Uppsala - SNP&SEQ Technology Platform

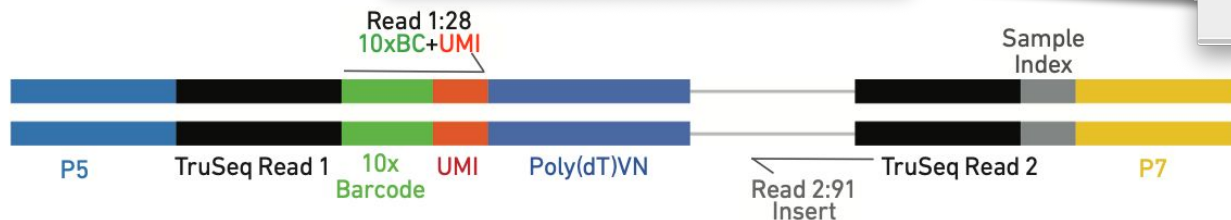
This is a report containing quality control information about your project run at the SNP&SEQ Technology Platform. If you have any questions, please do not hesitate to contact us at seq@medsci.uu.se

Report generated on 2020-11-27, 01:07 based on data in: /seqreports-data/nxf_work/2b/d14a7e223927c741f2a6713e8fa8

Welcome! Not sure where to start? [Watch a tutorial video](#) @0:00 don't show again

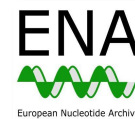
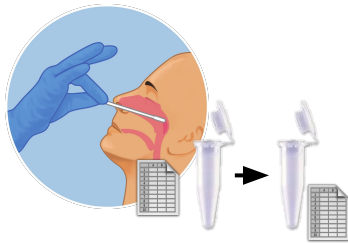
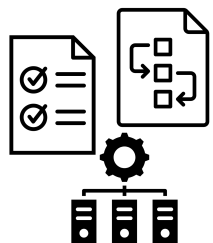
General Statistics

Sample Name	% GC	Length	M Seqs
TJ-2700-1_S1_L001_R1_001	46%	28 bp	53.5
TJ-2700-1_S1_L001_R2_001	48%	91 bp	53.5
TJ-2700-1_S1_L002_R1_001	46%	28 bp	53.5
TJ-2700-1_S1_L002_R2_001	48%	91 bp	53.5



- 201126_A00605_0172_BHVVTNDRXX-TJ-2700_multiqc_report.html
- checksums.md5
- SampleSheet.csv

“Protocol” & “project plan” icons by Justin Blake, and “infrastructure” icon by Eko Purnomo, from thenounproject.com



Study & data
design

Sampling
& specimen
collection

Sample
preparation

Sample analysis
& data generation

Data processing
to prepare inputs
for analysis

Data
analysis

Communicating
results

Procedures

data protection,
ethics permit,
infrastructure,
standards,
protocols,
data dictionaries,
data access, ...

Biosamples and instruments

populations (statistical) and inclusion criteria,
physical processing steps,
working storage conditions,
long-term storage location,
sample quality assessment,
sample annotations,
reagents, instruments, kits, ...

Data and computational workflows

digital processing steps,
working storage conditions,
long-term storage location,
data quality assessment,
sample/data annotations,
reference data,
analysis method...

Outputs

publications,
data,
tools,
workflows,
reports,
dashboards, ...



Sample collector's knowledge?

- Collection of specimen
- Pre-handover sample prep

Hurdles

- Shared terminology (ontology)
- File formats for data and documentation (metadata)
- Data delivery mechanisms

Data generator's knowledge?

- Technology specific preparation
- Data generation protocols
- Post generation processing
- Quality assessment before and after data generation



Ad Hoc

When it comes to my data, I have a "way of doing things" but no standard or documented plans.

One-Time

I create some formal plans about how I will manage my data at the start of a project, but I generally don't refer back to them.

Active and Informative

I develop detailed plans about how I will manage my data that I actively revisit and revise over the course of a project.

Optimized for Re-Use

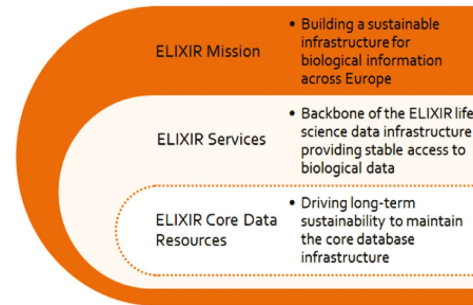
I have created plans for managing my data that are designed to streamline its future use by myself or others.

What data do you rely on?



<https://elixir-europe.org/platforms/data/core-data-resources>

- **Existing data to access**
Where and under what conditions are datasets available? How and when will you get access to them?
- **New data to be created**
What will be measured? Where? By whom? And using what instruments/methods?



National Genomics Infrastructure (NGI)

The National Genomics Infrastructure (NGI) provides services for next generation sequencing and SNP genotyping on all scales using a (...)

Ancient DNA

Use cleanroom labs and specialized molecular genetics techniques to extract, make libraries, sequence and analyze DNA in ancient and/or (...)

Clinical Genomics

Develops and provides clinical genetic tests using state-of-the-art genomic methods, such as next-generation sequencing, for (...)

Eukaryotic Single Cell Genomics

Provides service for high-throughput single cell genomics analysis

Microbial Single Cell Genomics

Provides streamlined single-cell sorting, lysis, whole-genome amplification and screening of individual microbial cells, as well as (...)

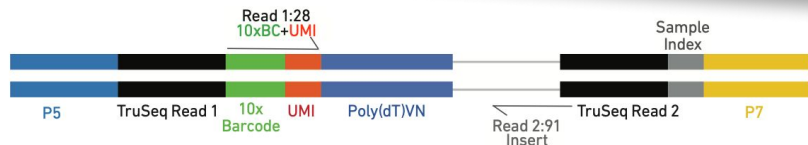
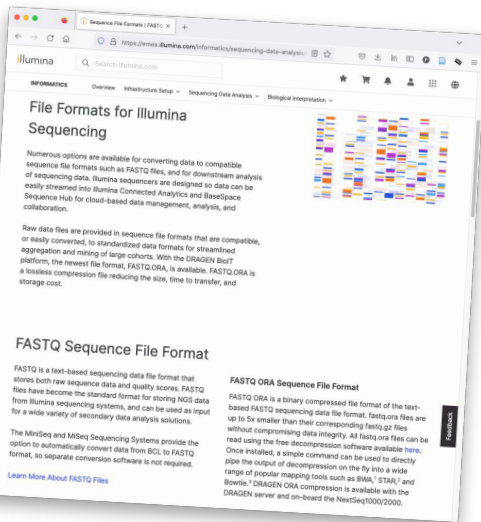
National Bioinformatics Infrastructure (NBIS)

Provides custom-tailored support with data analysis, computational tools, systems development and training.

What are their characteristics?



- ▼ SI-GA-F2_1
 - TJ-2700-1_S1_L001_R1_001.fastq.gz
 - TJ-2700-1_S1_L001_R2_001.fastq.gz
 - TJ-2700-1_S1_L002_R1_001.fastq.gz
 - TJ-2700-1_S1_L002_R2_001.fastq.gz
- ▼ SI-GA-F2_2
 - TJ-2700-1_S2_L001_R1_001.fastq.gz
 - TJ-2700-1_S2_L001_R2_001.fastq.gz
 - TJ-2700-1_S2_L002_R1_001.fastq.gz
 - TJ-2700-1_S2_L002_R2_001.fastq.gz
- ▼ SI-GA-F2_3
 - TJ-2700-1_S3_L001_R1_001.fastq.gz
 - TJ-2700-1_S3_L001_R2_001.fastq.gz
 - TJ-2700-1_S3_L002_R1_001.fastq.gz
 - TJ-2700-1_S3_L002_R2_001.fastq.gz
- ▼ SI-GA-F2_4
 - TJ-2700-1_S4_L001_R1_001.fastq.gz
 - TJ-2700-1_S4_L001_R2_001.fastq.gz
 - TJ-2700-1_S4_L002_R1_001.fastq.gz
 - TJ-2700-1_S4_L002_R2_001.fastq.gz



- **Kind of data**
Sequencing, numeric, textual, image, audio, video, etc...
- **Data and file formats**
What formats will you receive, get as output from software/equipment?
What docs do you need to use it?
- **Expected volumes**
What number of samples, rows, columns, files and/or their sizes?

What conventions do you adopt?

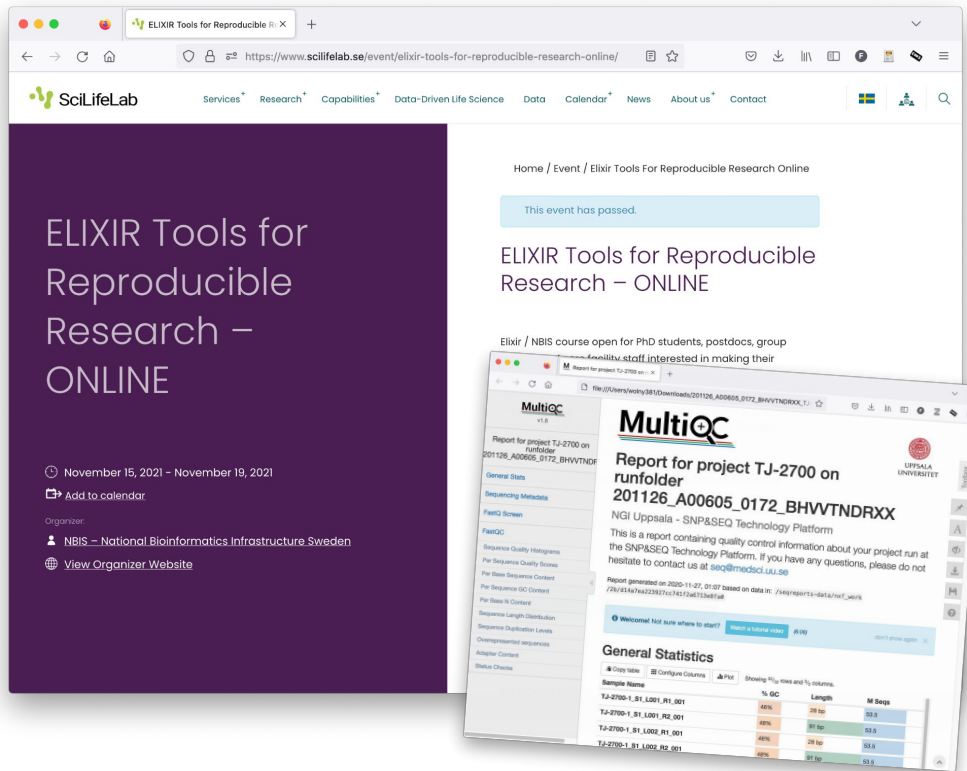


- **Community standards**
Data standards and terminologies used across related domains, e.g., data sharing platforms
- **Data organisation guidelines**
Project conventions for naming, locating and versioning files and documents
- **Documentation for reuse**
Data dictionaries, lab reports, pipelines, analysis transcripts...

The screenshot shows the RDMkit website interface. The top navigation bar includes 'Data management', 'About', 'Contribute', and 'GitHub'. The main content area is titled 'Your domain' and lists various data management solutions for different domains. A search bar is present. Below the main content, there is a table titled 'Sample characteristics as data' with columns for Population, Individual, Specimen, and Sample, and rows for various characteristics like Organism, Genotype, Geographical location, Growth condition, Sampling date, Developmental stage, Body weight, Sex, Organism part, and Cell type.

Population	Organism	Pacifastacus leniusculus
Individual	Genotype	wild type genotype
	Geographical location	Sweden, Lake Erken
	Growth condition	laboratory aquarium since Sep 2020
	Sampling date	2020-11-06
	Developmental stage	adult
	Body weight	35
	Sex	male
Specimen	Organism part	hematopoietic system
Sample	Cell type	hemocyte

What is done for quality assurance?

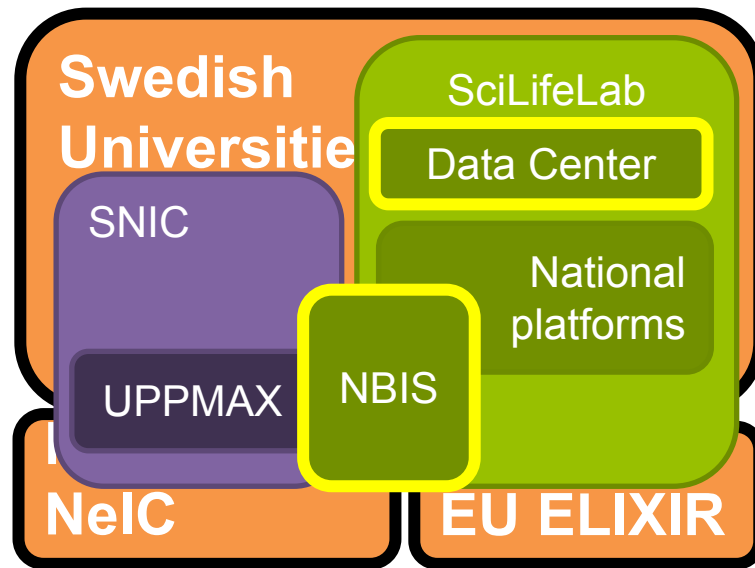


- **Traceability / provenance**
Instrument settings and calibration, standardised data capture, versioning pipelines, software and outputs
- **Data validation**
Repeated samples or measurements, peer review of data, or representation with controlled vocabularies

What platforms to you use?



- **Storage/processing locations**
For data collection, analysis, reporting, code, transfers etc.
- **Backup and data recovery**
Solutions to mitigate risks of data-loss and data corruption
(Beware of laptops and external storage)
- **Technical requirements**
Software and systems required to access / process the data?



What are the related policies?



Working with human data

2021-06-29 NBIS internal

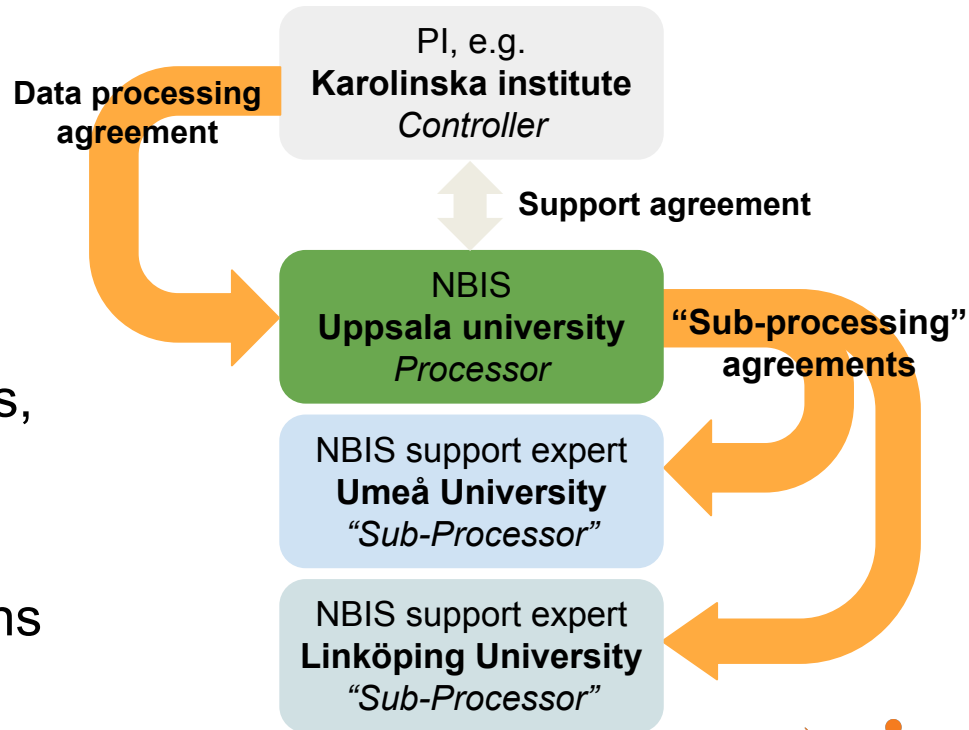


Data Stewardship Wizard Workshop



- **Information classification**
Suitable storage based on the characteristics of the data
- **Access control**
Who will have access to what data and how will it be enforced?
- **Data protection procedures**
Other strategies to mitigate risks of unwanted data disclosure or sabotage.

- **GDPR or sensitive data**
Regulation and agreements on personal, human, confidential or copyrighted information
- **Ownership & collaboration**
Sharing data between institutions, commercial entities, and nations
- **Contracts & licenses**
Legal, business & other limitations on processing or sharing data



What ethical considerations apply?



<https://fega.nbis.se/submission/ethical-approval.html>

<https://data-guidelines.scilifelab.se/>

Scilifelab RDM Guidelines
Knowledge hub for the management of life science research data in Sweden

Topics

Home / Topics / Research involving human data

On this page:

- What is human data?
- What is sensitive human data?
- Important regulations to follow
- Who is responsible for the data?
- Am I allowed to share data about humans?
- Repositories for publishing human data
- Considerations when working with human data
 - GDPR considerations
 - Ethical considerations
- Resources & Training

Research involving human data

Disclaimer! The PI, as well as everyone with access to sensitive personal data, is responsible for following current laws and regulations, and Scilifelab will not assume legal responsibility for advice provided in these guidelines.

What is human data?
Any data that directly or indirectly can be associated with a living person is considered personal data, e.g. name, address and personal identity number. See also legal reference regarding personal data.

What is sensitive human data?
Some personal data are regarded as sensitive and explicitly includes all genetic data and is likely to also apply to other data that might not be considered sensitive. Personal data should always be handled with care. See also legal reference regarding sensitive data.

Important regulations
Please find below an overview of the most important regulations regarding human data.

Ethical approval

Introduction

If you plan to do research in Sweden that involves processing of sensitive personal data, you must have an ethical approval from the Swedish Ethical Review Authority before you begin your research. An ethical approval may also be required under other circumstances that involve research on humans or on human biological material. You can enable others to reuse your data by specifying in the application how the data is going to be shared. The application that you send in to the Swedish Ethical Review Authority has to be in Swedish.

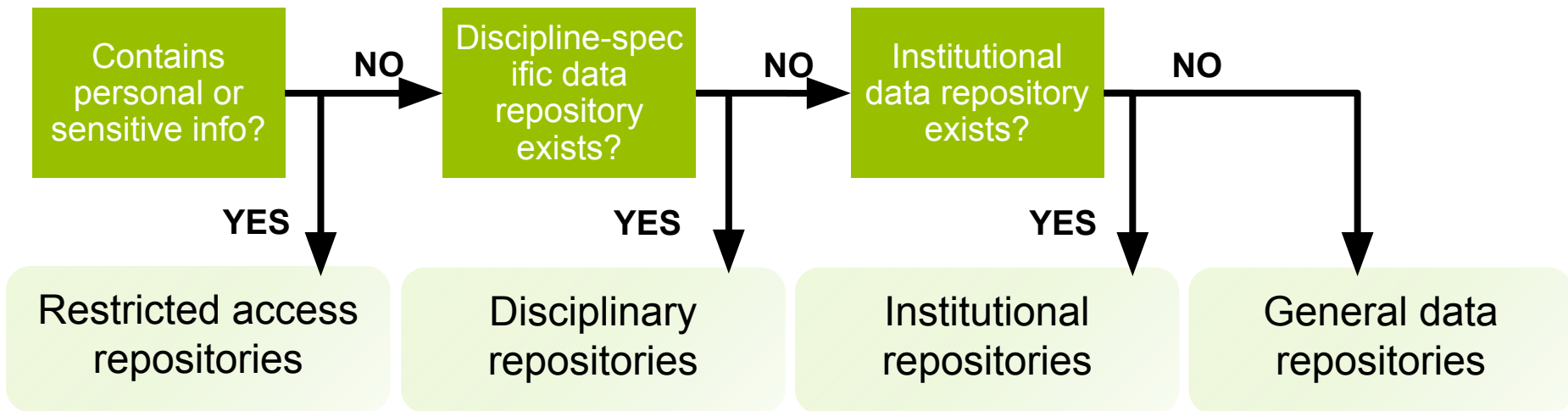
- **Ethical consent and review**
Limitations on processing data.
E.g., accessing, collecting, storing, analysing, etc...
- **Risk management**
Practices to minimise impact.
Granularity. Anonymisation.
- **Access and benefit sharing**
Balance communities / regions
using and providing data.

What will be available in the long-term?



- **Data sharing and restrictions**
Where to submit / deposit data
and under what conditions? E.g.,
embargo, committee, consent...
- **Data preservation and archiving**
What data to keep and for how
long? Where will it be archived?
- **Software and source code**
How to publish and maintain code,
research software and pipelines?

Selecting a data repository



What are the requirements for reuse?



- **Identifier and location**
Where and how to **Find**, **Access** and cite the data? DOI?
- **Open and common standards**
Data and file formats make your data **Interoperable**?
- **Permissive licensing**
How to enable and promote **Reuse** and future collaborations?



Ad Hoc

When it comes to my data, I have a "way of doing things" but no standard or documented plans.

One-Time

I create some formal plans about how I will manage my data at the start of a project, but I generally don't refer back to them.

Active and Informative

I develop detailed plans about how I will manage my data that I actively revisit and revise over the course of a project.

Optimized for Re-Use

I have created plans for managing my data that are designed to streamline its future use by myself or others.



Discuss in groups:

1. Where are you on the scale from ad-hoc to optimised practices?
2. What would be a possible next step towards optimised practices?

Group 1/E: The data

- What data do you rely on?
- What are their characteristics?

Group 2/D: Best practices

- What conventions do you adopt?
- What is done for quality assurance?

Group 3/C: Storage and access

- What platforms to you use?
- What are the related policies?

Group 4/B: Legal & ethics aspects

- What regulations and contracts apply?
- What ethical considerations apply?

Group 5/A: Sharing & long-term access

- What regulations and contracts apply?
- What ethical considerations apply?

What data do you rely on?



- **Existing data to access**

Where and under what conditions are datasets available? How and when will you get access to them?

- **New data to be created**

What will be measured?
Where? By whom? And using what instruments/methods?

Group 1

Ad-hoc or optimised?
Next steps?

What are their characteristics?



Group 1

Ad-hoc or optimised?

Next steps?

- **Kind of data**
Sequencing, numeric, textual, image, audio, video, etc...?
- **Data and file formats**
What formats will you receive, get as output from software/equipment?
What docs do you need to use it?
- **Expected volumes**
What number of samples, rows, columns, files and/or their sizes?

What conventions do you adopt?



- **Community standards**
Data standards and terminologies used across related domains, e.g., data sharing platforms
- **Data organisation guidelines**
Project conventions for naming, locating and versioning files and documents
- **Documentation for reuse**
Data dictionaries, lab reports, pipelines, analysis transcripts...

Group 2

Ad-hoc or optimised?
Next steps?



Group 2

Ad-hoc or optimised?

Next steps?

- **Traceability / provenance**
Instrument settings and calibration, standardised data capture, versioning pipelines, software and outputs?
- **Data validation**
Repeated samples or measurements, peer review of data, or representation with controlled vocabularies?

What platforms to you use?



- **Storage/processing locations**

For data collection, analysis, reporting, code, transfers etc.

- **Backup and data recovery**

Solutions to mitigate risks of data-loss and data corruption

(Beware of laptops and external storage)

- **Technical requirements**

Software and systems required to access / process the data?

Group 3

Ad-hoc or optimised?

Next steps?

What are the related policies?



Group 3

Ad-hoc or optimised?

Next steps?

- **Information classification**
Suitable storage based on the characteristics of the data
- **Access control**
Who will have access to what data and how will it be enforced?
- **Data protection procedures**
Other strategies to mitigate risks of unwanted data disclosure or sabotage.



- **GDPR or sensitive data**
Regulation and agreements on personal, human, confidential or copyrighted information?
- **Ownership & collaboration**
Sharing data between institutions, commercial entities, and nations?
- **Contracts & licenses**
Legal, business & other limitations on processing or sharing data?

Group 4

Ad-hoc or optimised?
Next steps?



Group 4

Ad-hoc or optimised?

Next steps?

- **Ethical consent and review**
Limitations on processing data?
E.g., accessing, collecting, storing, analysing, etc...
- **Risk management**
Practices to minimise impact?
Granularity? Anonymisation?
- **Access and benefit sharing**
Balance communities / regions
using and providing data?

What will be available in the long-term?



- **Data sharing and restrictions**
Where to submit / deposit data
and under what conditions? E.g.,
embargo, committee, consent...
- **Data preservation and archiving**
What data to keep and for how
long? Where will it be archived?
- **Software and source code**
How to publish and maintain code,
research software and pipelines?

Group 5

Ad-hoc or optimised?
Next steps?

What are the requirements for reuse?



Group 5

Ad-hoc or optimised?

Next steps?

- **Identifier and location**
Where and how to **Find**, **Access** and cite the data? DOI?
- **Open and common standards**
Data and file formats make your data **Interoperable**?
- **Permissive licensing**
How to enable and promote **Reuse** and future collaborations?