

# HeFDI Data Talk

Date	Topic	Presenter(s)
01 December 2023	NFDI4Microbiota - Infrastructure for scientific data	Dr. Michael Vockenhuber (SYNMIKRO - Philipps University Marburg)



## Abstract:

The talk will present the goals of the NFDI4Microbiota consortium and its services to the scientific community, including electronic lab notebooks, data management plans, workflows and training. There will also be an introduction to the local HPC and storage services (MaRC3 and MaSC).

## About the HeFDI Data Talks:

The HeFDI Data Talks are a bi-weekly open information and discussion event focused on data management in the context of science, in which relevant NFDI consortia as well as research data management services present themselves. The series discusses current topics and presents numerous – including local and regional – tools and services. The HeFDI Data Talks are an offer of the HeFDI Initiative (Landesinitiative HeFDI), which is funded by Hesse's Ministry for Science and Arts (HMWK).

DOI link: <https://doi.org/10.5281/zenodo.10301978>; License information: Creative Commons Attribution 4.0 International ([CC BY 4.0](#))



# NFDI4Microbiota

Infrastructure for scientific data

Michael Vockenhuber

Center for Synthetic Microbiology (SYNMIKRO)

Bioinformatics Core Facility

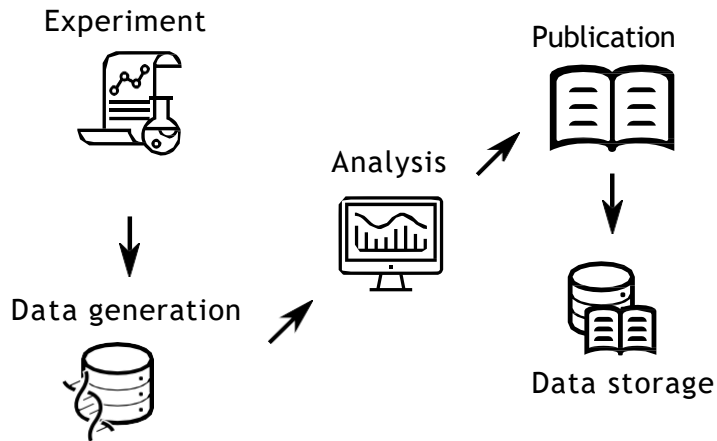
2023-12-01



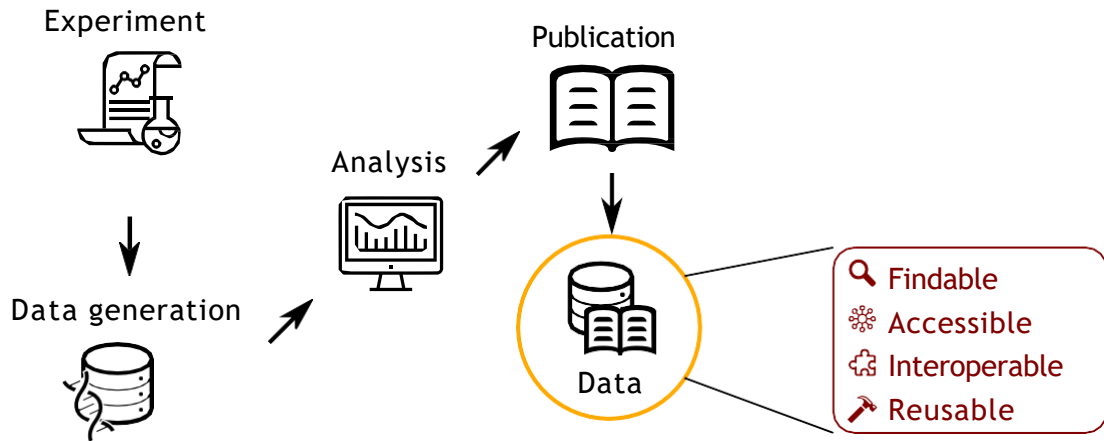
**Idea:** building a central hub to support researchers with

- data access
- data analysis
- (meta)data standards
- workflows / SOPs
- training

## Typical experimental workflow

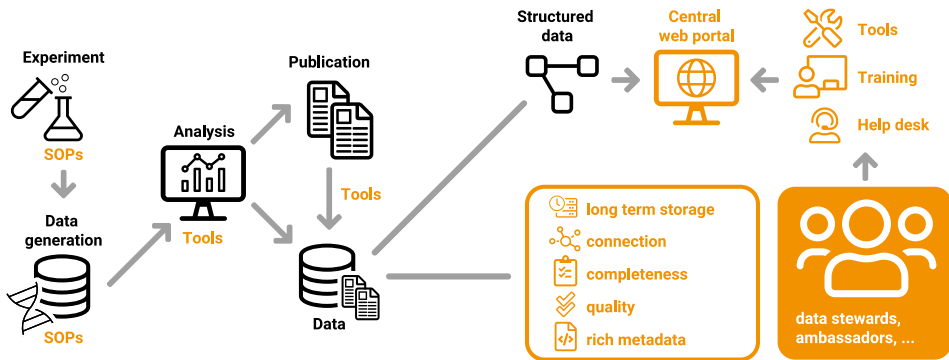


# Typical experimental workflow



Where does NFDI4Microbiota come in ?

# Where does NFDI4Microbiota come in ?



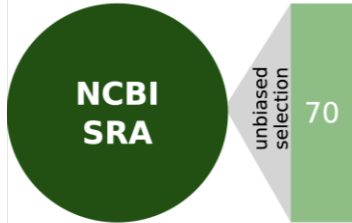
# Obtaining good quality data can be hard

Looking for plant RNAseq data in  
the NCBI Sequence Read Archive



**NCBI  
SRA**

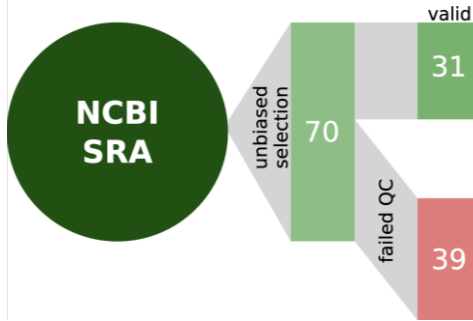
# Obtaining good quality data can be hard



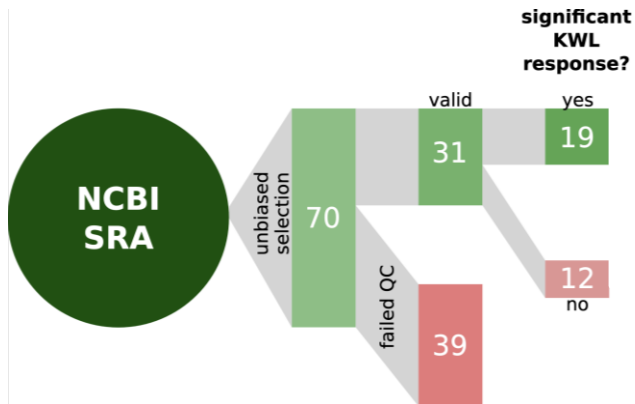
2.3 TB raw data



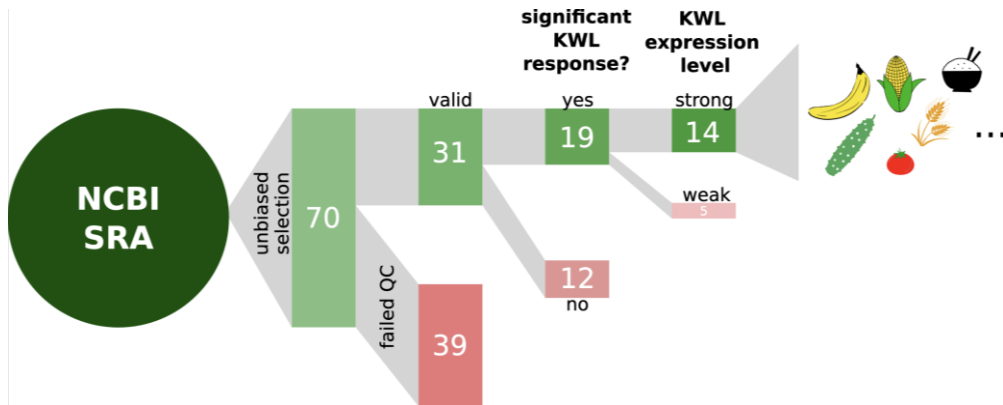
# Obtaining good quality data can be hard



# Obtaining good quality data can be hard



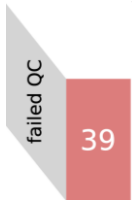
# Obtaining good quality data can be hard



55% loss of usable data due to (non-scientific) reasons  
i.e. low-quality metadata

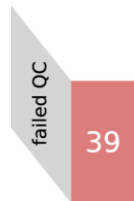
## Obtaining good quality data can be hard

What are common problems why data cannot be reused ?



# Obtaining good quality data can be hard

What are common problems why data cannot be reused ?



- obfuscated design - what sample represents what condition ?

☐ [5000070\\_5000439\\_5000488\\_36](#)

1. 1 ILLUMINA (Illumina Genome Analyzer II) run: 5.7M spots, 204.7M bases, 127.1Mb downloads  
Accession: SRX017414

☐ [5000070\\_5000438\\_5000487\\_36](#)

2. 1 ILLUMINA (Illumina Genome Analyzer II) run: 4.7M spots, 169.1M bases, 107.4Mb downloads  
Accession: SRX017413

☐ [5000070\\_5000440\\_5000489\\_36](#)

3. 1 ILLUMINA (Illumina Genome Analyzer II) run: 6.1M spots, 219M bases, 134.8Mb downloads  
Accession: SRX017412

☐ [5000070\\_5000441\\_5000490\\_36](#)

4. 1 ILLUMINA (Illumina Genome Analyzer II) run: 5M spots, 181.4M bases, 112.4Mb downloads  
Accession: SRX017411

☐ [5000070\\_5000444\\_5000548\\_36](#)

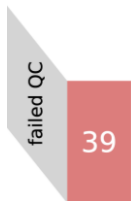
5. 1 ILLUMINA (Illumina Genome Analyzer II) run: 3.8M spots, 138.5M bases, 86.6Mb downloads  
Accession: SRX017410

☐ [5000070\\_5000443\\_5000547\\_36](#)

6. 1 ILLUMINA (Illumina Genome Analyzer II) run: 4.7M spots, 167.6M bases, 105.3Mb downloads  
Accession: SRX017409

## Obtaining good quality data can be hard

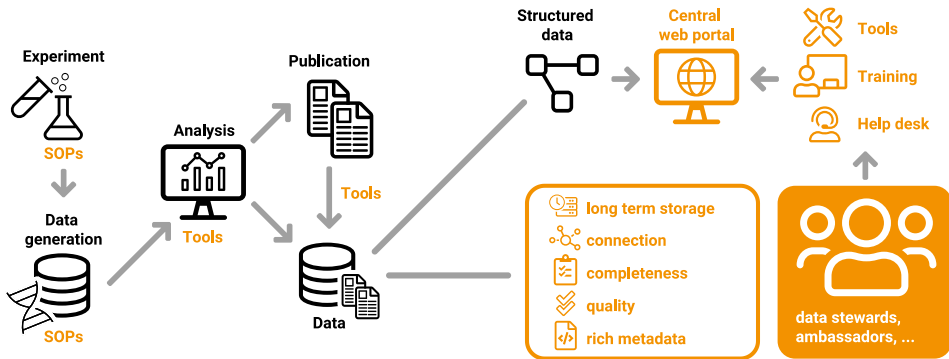
Why can these 39 datasets not be reused ?



- obfuscated design - what sample represents what condition ?
- incomplete description of study design
- no reproduction - n1
- wrong organism ? - description does not fit publication
- no publication connected to data - no scientific explanation

A lot of time and money was spent to create *dead* data

# How can we change this ?



# Help in detail

Experiment



Exp. Procedures



Training



Help desk

Data generation



Data Management



Meta Data



Quality assessment

Analysis



Workflows



Tools



IT Infrastructure

Data &  
Publication



Indexing



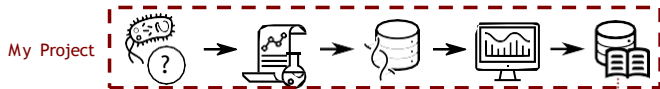
Reposition



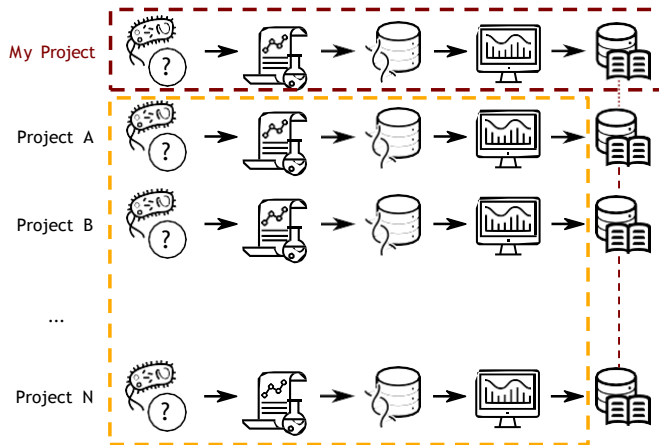
Transparency



# The Plan : Improved Reuseability

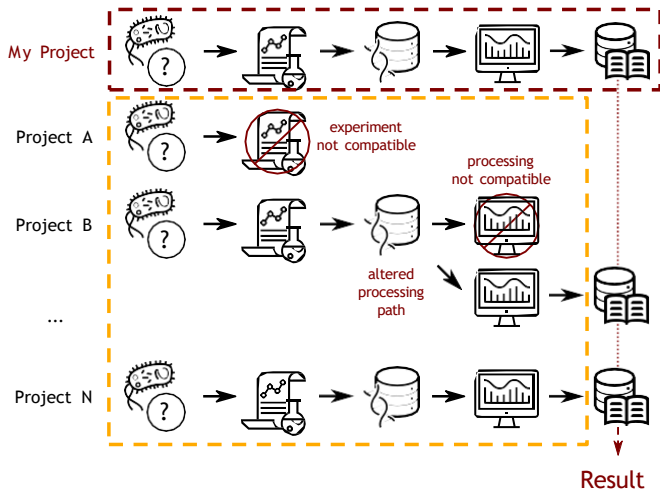


# The Plan : Improved Reuseability



Can I use those ?

# The Plan : Semi-Automated Decision Making



The System aides in the process of choosing Projects to include in my data set.

# (Co-)applicants



Microbiome research  
Phylogenomics  
Next-generation cultivation



de.NBI headquarter  
Microbial bioinformatics  
Cloud computing



Microbial resource center  
Microbial databases  
Phylogenomics



European research network  
Molecular biology  
Training



Software engineering  
Automated workflows  
Cloud computing



Computational metagenomics  
Pathogens  
Benchmarking



Virus bioinformatics  
RNA biology  
Host of EVBC



Imaging  
Metabolism  
Structural biology



Environmental research  
Research data management  
Natural & engineered ecosystems



Linked Open Data  
Discovery Services  
Data Stewardship

# What is already available / in prep ?

Available

- Workflows
- Knowledgebase
- Trainings
- Data Storage
- high performance computing (MaRC3)

## What is already available / in prep ?

### Available

- Workflows
- Knowledgebase
- Trainings
- Data Storage
- high performance computing (MaRC3)

### coming soon

- electronic Labnotebook
- semi-automatic DMP generator

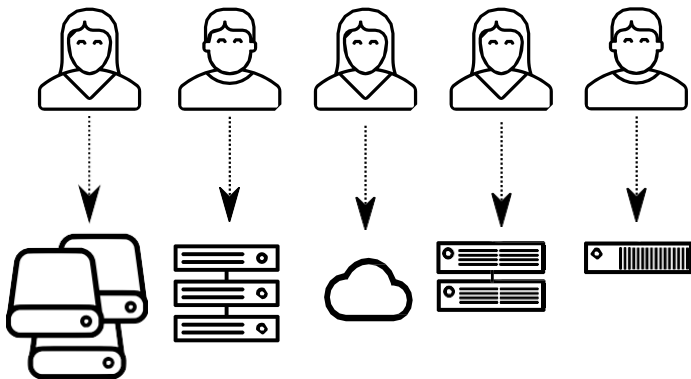


Be the central hub in Germany for supporting the microbiology community with data access, analysis services, data/metadata standards and training.

# MaSC - Marburger Storage Cluster

## Initial situation on big data

- no central big data storage available
- scattered landscape of individual solutions
- setup and administration drains scientific capacity
- know-how often lost during time (PhDs)







3 copies



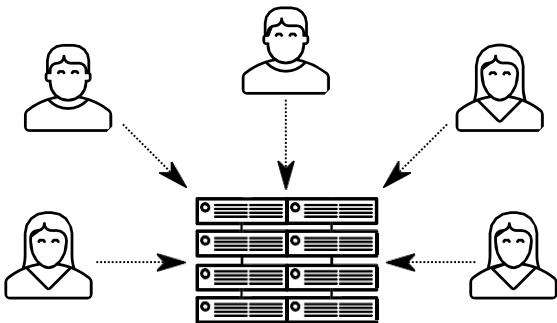
2 media types



1 off-site copy

## MaSC - Marburger Storage Cluster

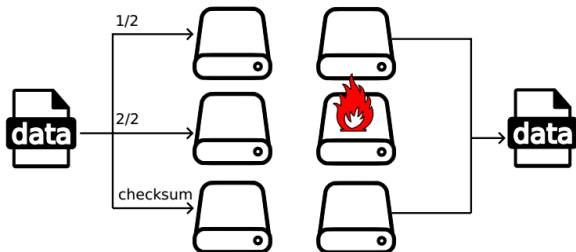
- crowd-funded storage system
- consolidate IT investments and maintenance (e.g. administration, power, cooling, network, ...)
- highly scalable, anyone can participate at any time



# IT basics - How to get reliable data capacity?

## Redundant array of independent disks (RAID)

- data is distributed over multiple disks in one server
- mirror or checksums used to recover from failures

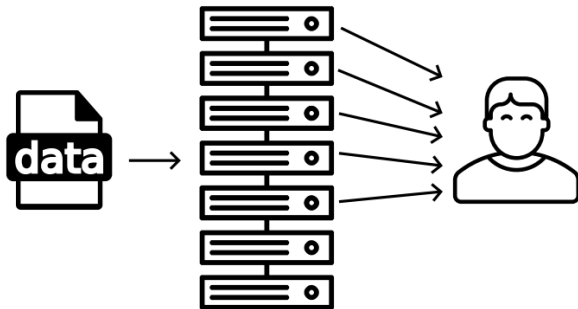


- hard to expand
- cannot extend beyond a single server

## IT basics - How to get reliable data capacity?

### Erasure coding (EC)

- data is distributed over multiple disks at any site
- fixed scheme (e.g. 5 fragments + 2 checksums)



- fast, highly scalable
- high availability

## MaSC - Marburger Storage Cluster

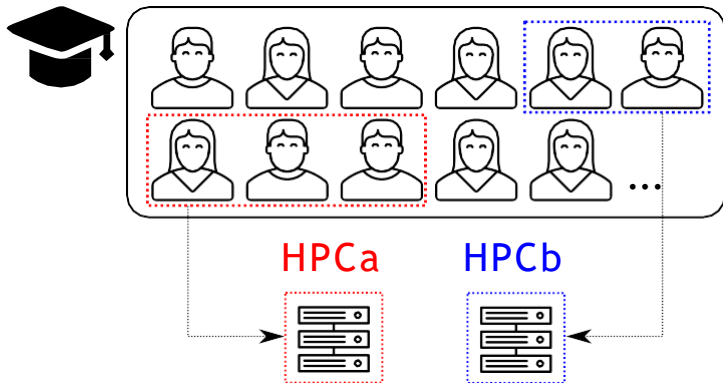
- current capacity: 4 PB (90% used, 2.4 Bil files) + 2 PB in January 2024
- Erasure coding (8+3), tolerates 3 server failures
- snapshots / 3-2-1 backup strategy / self healing
- fast access to UMRnet and HPC cluster MaRC3



# MaRC3 - Community Organization

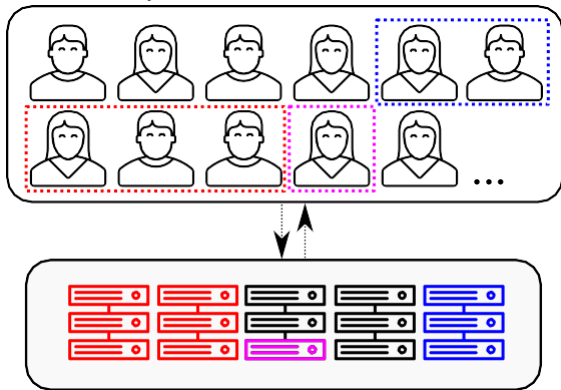
## Initial situation on HPC

- MaRC2 dying, no central successor HPC available
- funds only for distinct research fields or groups



## MaRC3 initiative

- provide compute power to **all** UMR researchers
- use HPC resources holistically while respecting funding schemes
  - centralized access to all systems
  - shared use of unused system-time



# MaRC3 - Marburger Rechen-Cluster 3



50 compute nodes



3200 CPU cores

AMD EPYC, 64 cores



256-1024 GB RAM



100 accelerators (GPUs, NVIDIA A100)



300 TB SSD-storage, MaSC access



25 Gbit Ethernet Network

100 Gbit to MaSC



# Bioinformatics Core Facility @ SYNMIKRO



Bioinformatics

Computational Support, Research Data Management

High Performance Computing (MaRC3)

Large-Scale Storage (MaSC)



Philipps



Universität  
Marburg



NFDI4  
MICROBIOTA



HRZ  
Uni Marburg

