GFBio: Publish your sequences with metadata!

Ivaylo Kostadinov, Ph.D.
GFBio e.V.

ivo@gfbio.org
@tigroumaniac
www.gfbio-ev.de

Today's Goals

1. Standardisation of the metadata according to the MIxS checklist (**M**inimal **I**nformation about Any(**x**) **S**equence)

2. The Data Submission and Brokerage System developed by GFBio to help archive and publish complex datasets through a single point of entry

3. Linking sequence data to PANGAEA records (reciprocally)

Agenda

- General Introduction (15 min)
  - Submitting data via GFBio | NFDI4Biodiversity
- MIxS (15 min)
  - Packages | Examples | ENVO
- Demo (30 min)
  - Metadata template | Walk-through
- Further Discussion (30 min)

Jimena Linares

Jimena Linares

DATA

Research Funding in DE

142 Billion €

http://www.dfg.de/sites/foerderatlas2021

DFG Funding Atlas

7.9 Billion Euro of third party funding was shared between Universities in Germany in 2019

https://www.dfg.de/sites/foerderatlas2021/

Time effort

.. for discovering and reusing multiple data sources

80%

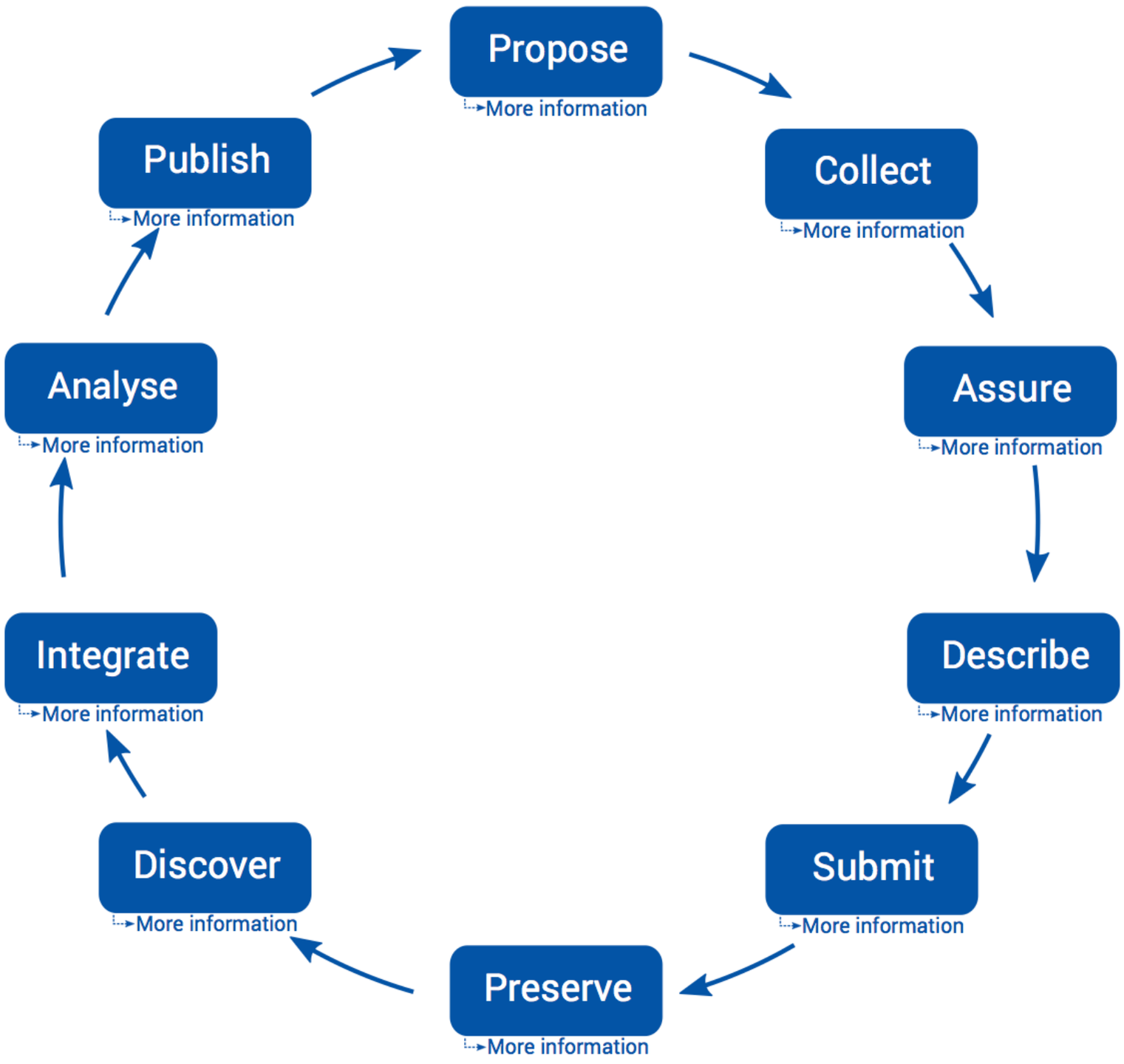
Mons, B. et al., doi:10.3233/ISU-1704824

dark data

[…] data that has never been published or otherwise made available to the rest of the scientific community.

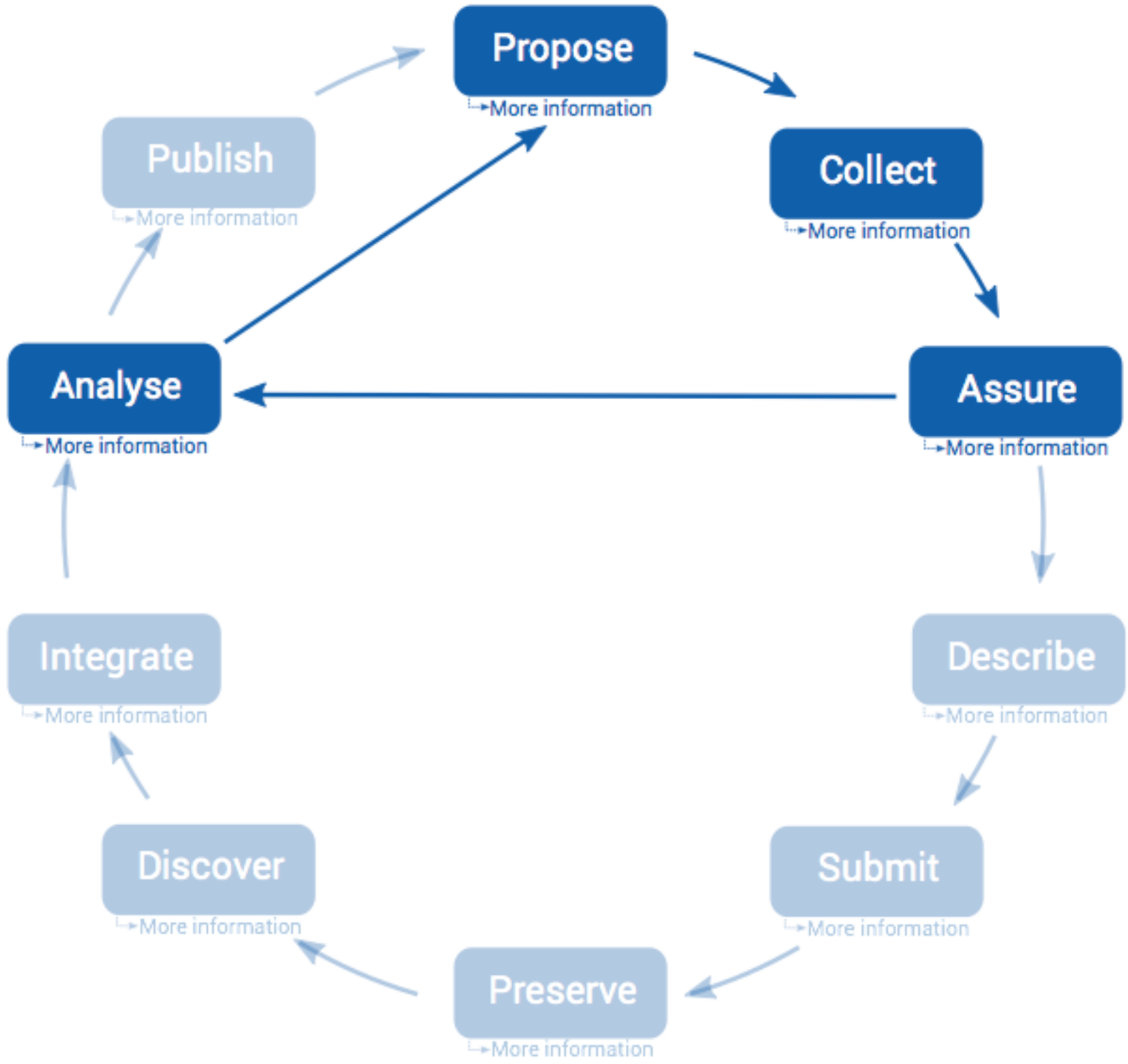
B. P. Heidorn Libr. Trends 57, 280–299; 2008

Data Lifecycle



http://www.gfbio.org/training/materials/data-lifecycle

Data Lifecycle



http://www.gfbio.org/training/materials/data-lifecycle

Data challenges

- Contextual data is not standardized

- <50% of contextual data ends up in a dedicated database

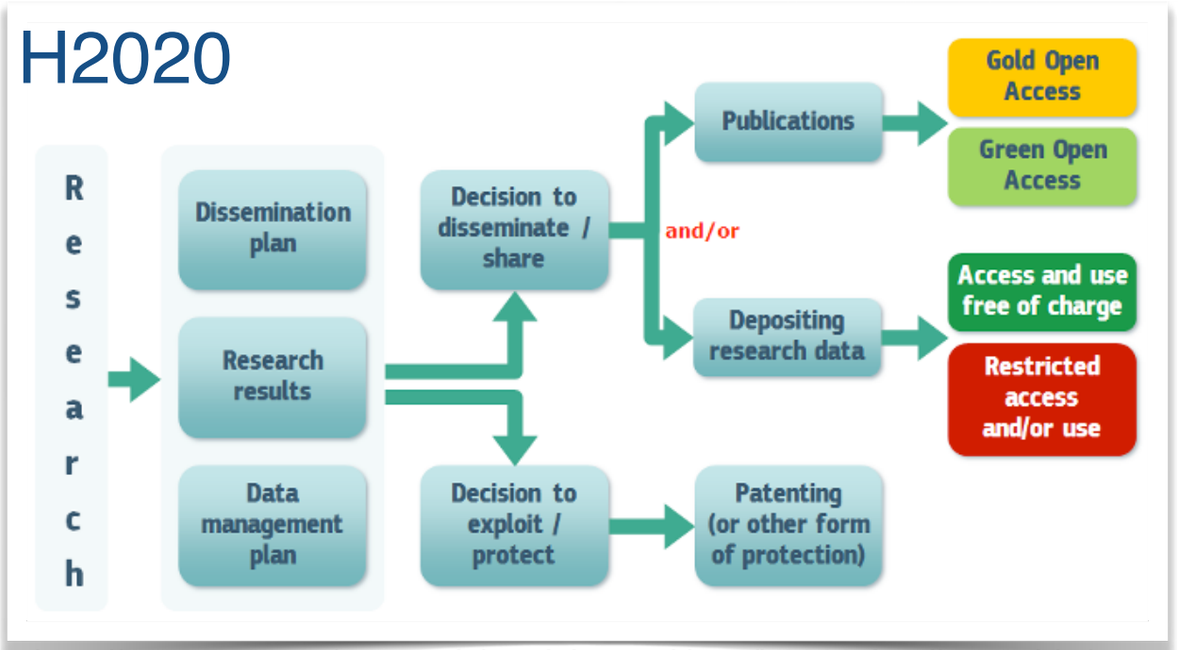
- Submitter = Manager = Analyst

Community response

OECD Principles and Guidelines for Access to Research Data from Public Funding

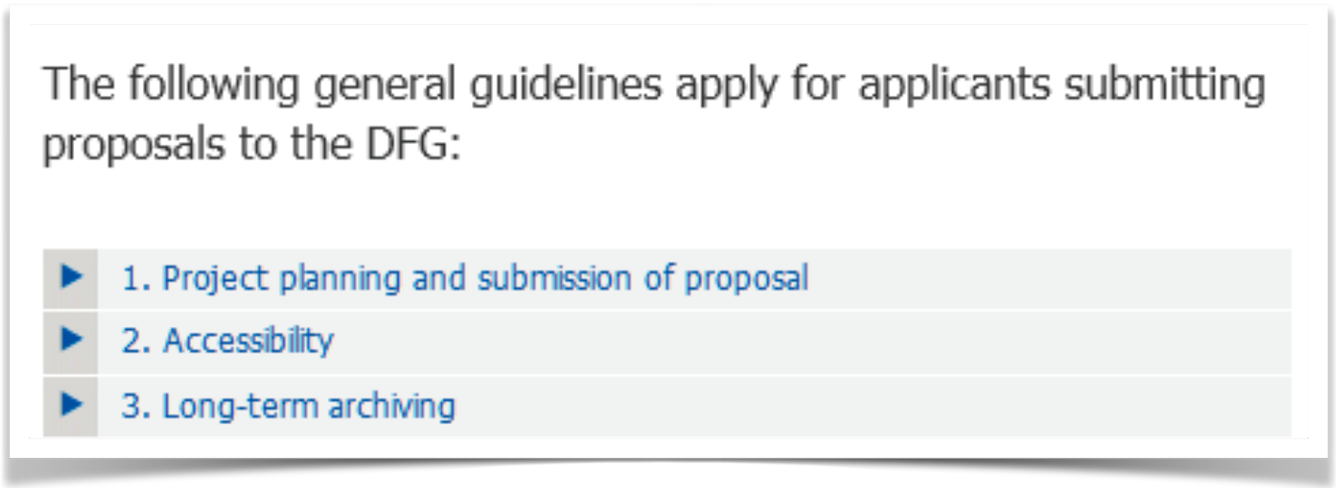
2007

http://www.oecd.org/sti/sci-tech/38500813.pdf



http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/open-access_en.htm

Rat für Informations Infrastrukturen



http://www.dfg.de/en/research_funding/proposal_review_decision/applicants/submitting_proposal/research_data/

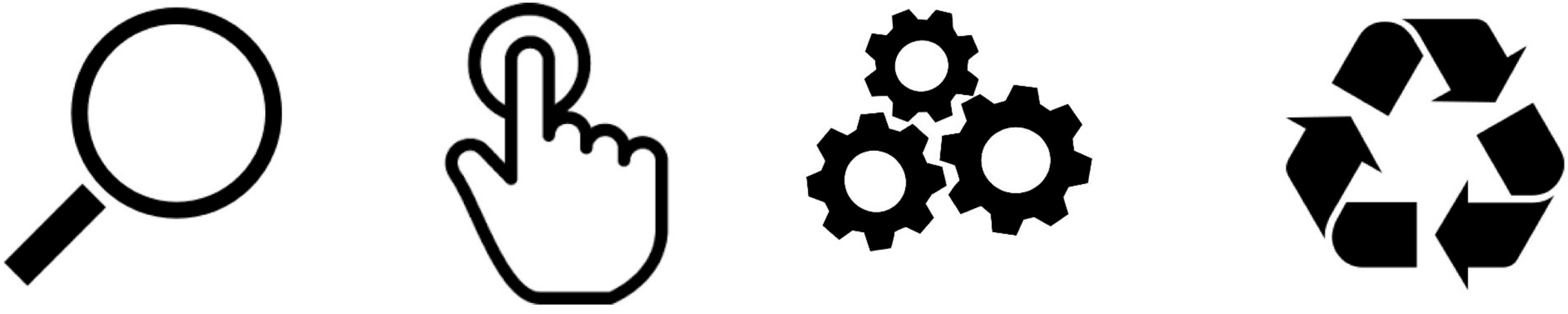
LEISTUNG AUS VIELFALT

Empfehlungen zu Strukturen, Prozessen und Finanzierung des Forschungsdatenmanagements in Deutschland

http://www.rfii.de

Community response

Findable Accessible Interoperable Reusable



By SangyaPundir [CC BY-SA 4.0], from Wikimedia Commons

www.nature.com/scientificdata

SCIENTIFIC DATA

OPEN

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

Mark D. Wilkinson *et al.*#

SUBJECT CATEGORIES
» Research data
» Publication characteristics

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

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.

Findable
Accessible
Interoperable
Reusable

Wilkinson, et al., Scientific Data, 2016
http://doi.org/10.1038/sdata.2016.18

FAIR Data

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

Wilkinson, et al., Scientific Data, 2016 http://doi.org/10.1038/sdata.2016.18

ARCHIVAL & PUBLICATION

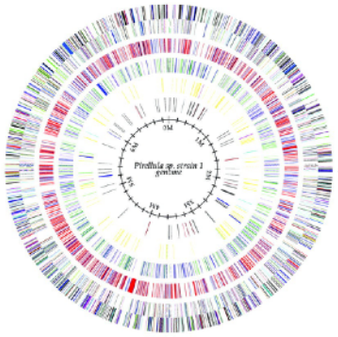
Archival & Publication

Findable

Accessable

Interoperable

Reproducible



land use nitrate salinity host relationship cell size motility calcium pertubation 16S sulfide bromide exoenzymes chemotaxis biofilm products antibiotics metabolism halophily magnesium substrate spectrum isolation oxygen pathogenicity light phosphate carbon classification genome organic matter pigmentation ammonium sulfate C/N ratio gram stain ph CO2 cultivation temperature

ENA European Nucleotide Archive

INSDC International Nucleotide Sequence Database Collaboration

PANGAEA

dedicated, long-term archives

NCBI

https://submit.ncbi.nlm.nih.gov/

gfbio e.v.

NCBI Site map | All databases | Search

Log in

Submission Portal

Submit data to NCBI Help

SEQUENCE DATA

GenBank

Genetic sequence database, an annotated collection of all publicly available DNA sequences. Submit Bacterial or Archaeal 16S ribosomal RNA sequences via the 16S rRNA Submission Tool (beta version) wizard. All other submission types should use one of the alternate submission tools (e.g. BankIt, Sequin, tbl2asn, etc.)

Genomes (WGS)

The Whole Genome Shotgun (WGS) database accepts prokaryotic and eukaryotic genomes that are draft or incomplete. This submission portal was updated on Feb. 3, 2014, and now accepts fasta sequences. To access submissions started before that date, go to the previous version of the portal.

Complete Genomes

Collection of genomic sequences that are used to represent the genome of an organism.

TSA

Computationally assembled sequences from primary data such as ESTs, traces and Next Generation Sequencing Technologies. TSA sequence records differ from EST and GenBank records because there are no physical counterparts to the assemblies.

SRA

The Sequence Read Archive (SRA) stores raw and aligned sequence data from NextGen sequencing platforms.

SRA Submissions (Beta)

The new submission interface for SRA. Data files can only be uploaded through this interface using the Aspera Connect plugin. Please use the other SRA submission tool if you need to upload your data using FTP or Aspera command line.

GEO

Next generation sequence submissions for functional genomic studies that

PROJECT DATA

BioProject

A collection of biological data related to a single initiative, originating from a single organization or from a consortium.

BIOLOGICAL MATERIALS

BioSample

Descriptions of biological source materials used in experimental assays.

MICROARRAY DATA

dbGaP

Microarray data from clinical studies that require controlled access.

GEO

Microarray submissions for functional genomic studies that examine gene expression, regulation or epigenomics.

MANUSCRIPT

NIHMS

An electronic version of your peer-reviewed final manuscript for inclusion in PubMed Central.

CLINICAL DATA

GTR

Genetic tests for inherited and somatic genetic variations, including newer types of tests such as arrays and multiplex panels.

VARIATION

Variation

Variation resources at NCBI: dbSNP, dbVar and ClinVar. dbSNP represents short variation in any organism including single nucleotide variants, insertions, deletions, and microsatellites. dbVar represents genomic structural variations from studies submitted on any organism or phenotype. ClinVar aggregates information about human sequence variation and its relationship to human health.

EBI-ENA https://www.ebi.ac.uk/ena/submit

EMBL-EBI Services Research Training About us

ENA European Nucleotide Archive

Search

Examples: BN000065, histone

Advanced Sequence

Home | Search & Browse | Submit & Update | Software | About ENA | Support

ENA > Submit and update

Submitting and updating data

We offer a number of services through which data (including updates) can be submitted to the European Nucleotide Archive (ENA). These technologies provide options appropriate for the scale and frequency of submission, the expertise and capacity of the submitter and the nature of the data to be transferred. The choices below lead users most directly to the appropriate submission route.

Submit to ENA

Email ENA Helpdesk

Programmatic submissions

Most types of submissions can also be made programmatically ... more information.

Submit & Update

- Data formats
- Taxon ID requests
- Uploading data files
- Reads
- Sequences
- Genome assembly submissions
- Taxonomy
- Sample checklists
- Environmental
- Epigenomic
- Species BARCODE
- Metadata model
- Register submission account
- Programmatic XML submissions

EBI-ENA https://www.ebi.ac.uk/ena/submit/sra/#home

EMBL-EBI Services Research Traini

Please specify the release date of your study:
This is when your study will be made public.
11-Aug-2015

Please provide a short name for the st

Please provide a short descriptive title

Please provide an abstract to describe

For genome assembly projects only: Ir
PLEASE ANSWER WITH YES IF YOU HAV
Yes
No

Please provide PubMed IDs of publicat
(numeric value)
PubMed IDs
Add

Please provide attributes to add a deep
Tag Value
Add

<< Previous

Please provide a short r

Please select the checklist attributes you would like to include with each sample. Recommended attributes can be unselected from within the corresponding attribute group on the left-hand side panel. You may also add custom attributes.

Filter attributes...

Add your own attribute + Add

- internal environment
- local environment conditions
- non-sample term
- sample collection
- organism characteristics
- concentration measurement
- host details
- other
- host description
- geography
- User Attributes

12 of 78 attributes selected

+ Expand − Collapse Download Template

Please complete any fields that you would like to apply to all samples. This will act as a template for the rest of the samples.

Template Basic Details

Unique Name Prefix:
* Title:
Description:

Organism Details

If your organism is not found please go here and email datasubs@ebi.ac.uk with the required details listed on the page in order for us to request a taxon Id for your organism.

Search:
* Tax Id:
* Scientific Name:
Common Name:

non-sample term

* investigation type:
* project name:
* sequencing method:

sample collection

<< Previous Next >>

PANGAEA

http://www.pangaea.de/submit/

gfbio e.V.

Logged in as **ikostadi** (log out, profile)

PANGAEA®
Data Publisher for Earth & Environmental Science

Submit Data to PANGAEA

Welcome to the PANGAEA data submission system. Any data from earth and life sciences are accepted. We highly appreciate you archiving and publishing your data with PANGAEA.

- **Benefits.** Published data are fully citable and can be cross-referenced with journal articles – read more...
- **Data preparation and quality control.** We will be in direct contact with you during preparation and archiving of your data – read more...
- **Costs.** PANGAEA can be used free of charge. Nevertheless, we appreciate any possible financial support – read more...

When you start the data submission process below, you will be redirected to the PANGAEA issue tracker that will assist you in providing metadata and uploading data files. Any communication with our editors will go through this issue tracker.

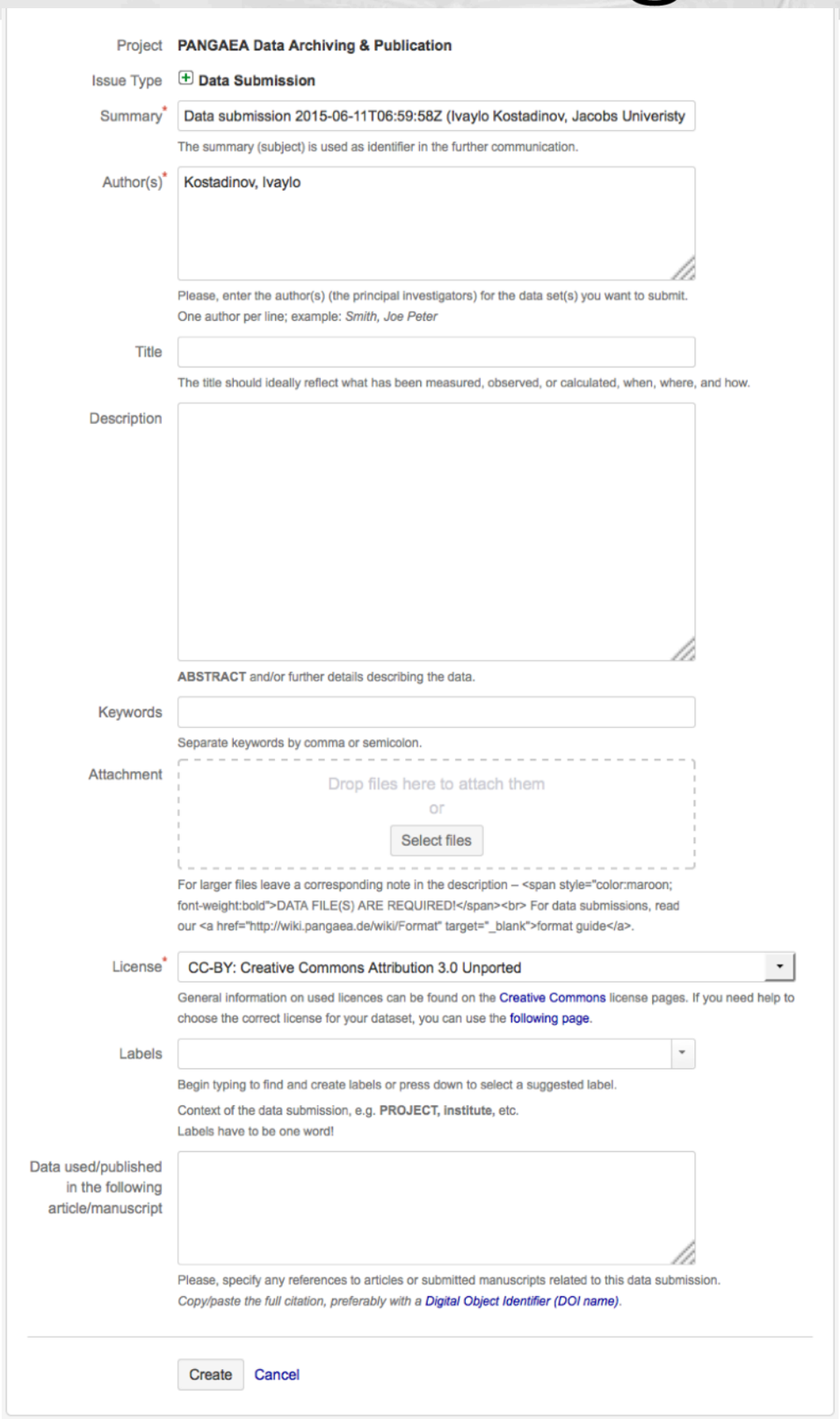
SUBMIT YOUR DATA

Further details about data submission to PANGAEA – if you have any comments on the data submission process, please contact us.

Contact

Easy for the user,

not so much for the curator.



Project **PANGAEA Data Archiving & Publication**

Issue Type **Data Submission**

Summary* Data submission 2015-06-11T06:59:58Z (Ivaylo Kostadinov, Jacobs Univeristy

The summary (subject) is used as identifier in the further communication.

Author(s)* Kostadinov, Ivaylo

Please, enter the author(s) (the principal investigators) for the data set(s) you want to submit.
One author per line; example: *Smith, Joe Peter*

Title

The title should ideally reflect what has been measured, observed, or calculated, when, where, and how.

Description

ABSTRACT and/or further details describing the data.

Keywords

Separate keywords by comma or semicolon.

Attachment

Drop files here to attach them
or
Select files

For larger files leave a corresponding note in the description – DATA FILE(S) ARE REQUIRED!
 For data submissions, read our format guide.

License* CC-BY: Creative Commons Attribution 3.0 Unported

General information on used licences can be found on the Creative Commons license pages. If you need help to choose the correct license for your dataset, you can use the following page.

Labels

Begin typing to find and create labels or press down to select a suggested label.
Context of the data submission, e.g. **PROJECT, institute,** etc.
Labels have to be one word!

Data used/published in the following article/manuscript

Please, specify any references to articles or submitted manuscripts related to this data submission.
Copy/paste the full citation, preferably with a Digital Object Identifier (DOI name).

Create Cancel

Archival & Pub. Challenges

- Different archives have different submission and curatorial workflows, requirements, response times.
- Your time is limited. Do you prefer to do analysis or submissions?
- Incentives (i.e. credits) for high-quality, FAIR data often unclear.

gfbio e.V.

gfbio
GERMAN FEDERATION
FOR BIOLOGICAL DATA

www.gfbio.org

A sustainable, service-oriented network of infrastructures and experts.

Funded by:
DFG Deutsche Forschungsgemeinschaft

www.gfbio.org

gfbio e.V.

https://www.gfbio.org

About Services Infothek Events GFBio e.V. Sign In

gfbio
GERMAN FEDERATION FOR BIOLOGICAL DATA

FAIR • Research • Data
Biodiversity, Ecology & Environmental Science

Enter a search term... FIND DATA

Plan Submit Visualize

GFBio Services

PLAN

Prepare a custom Data Management Plan (DMP).

SUBMIT

Submit your data to GFBio.

SEARCH

Search the GFBio data pool.

VISUALIZE & ANALYZE

Dynamically integrate, analyze and visualize GFBio datasets.

PUBLISH

Make your data citable.

TRAIN

Train your data management skills.

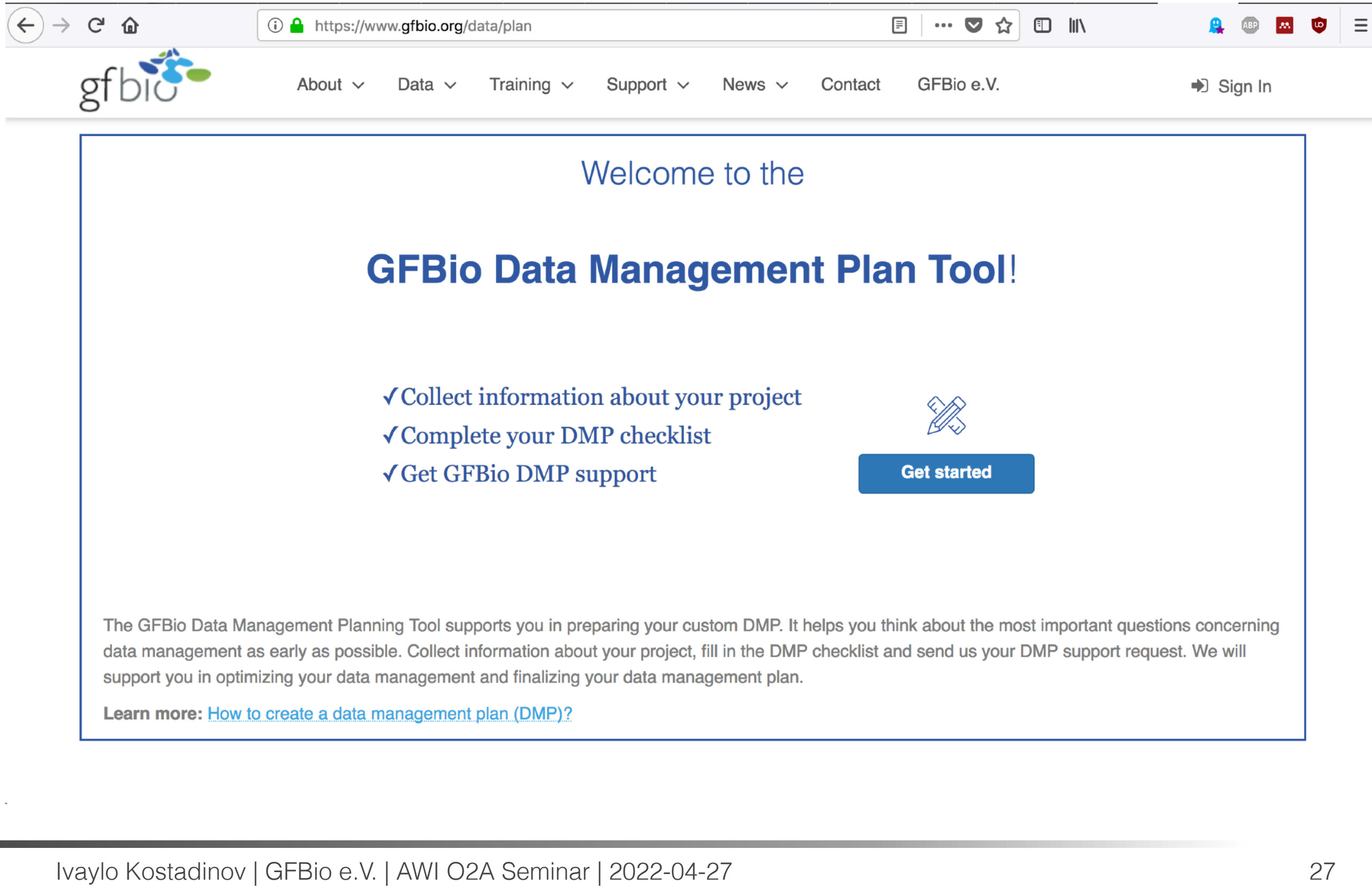
ARCHIVE

Deposit data and specimens in dedicated long-term archives.

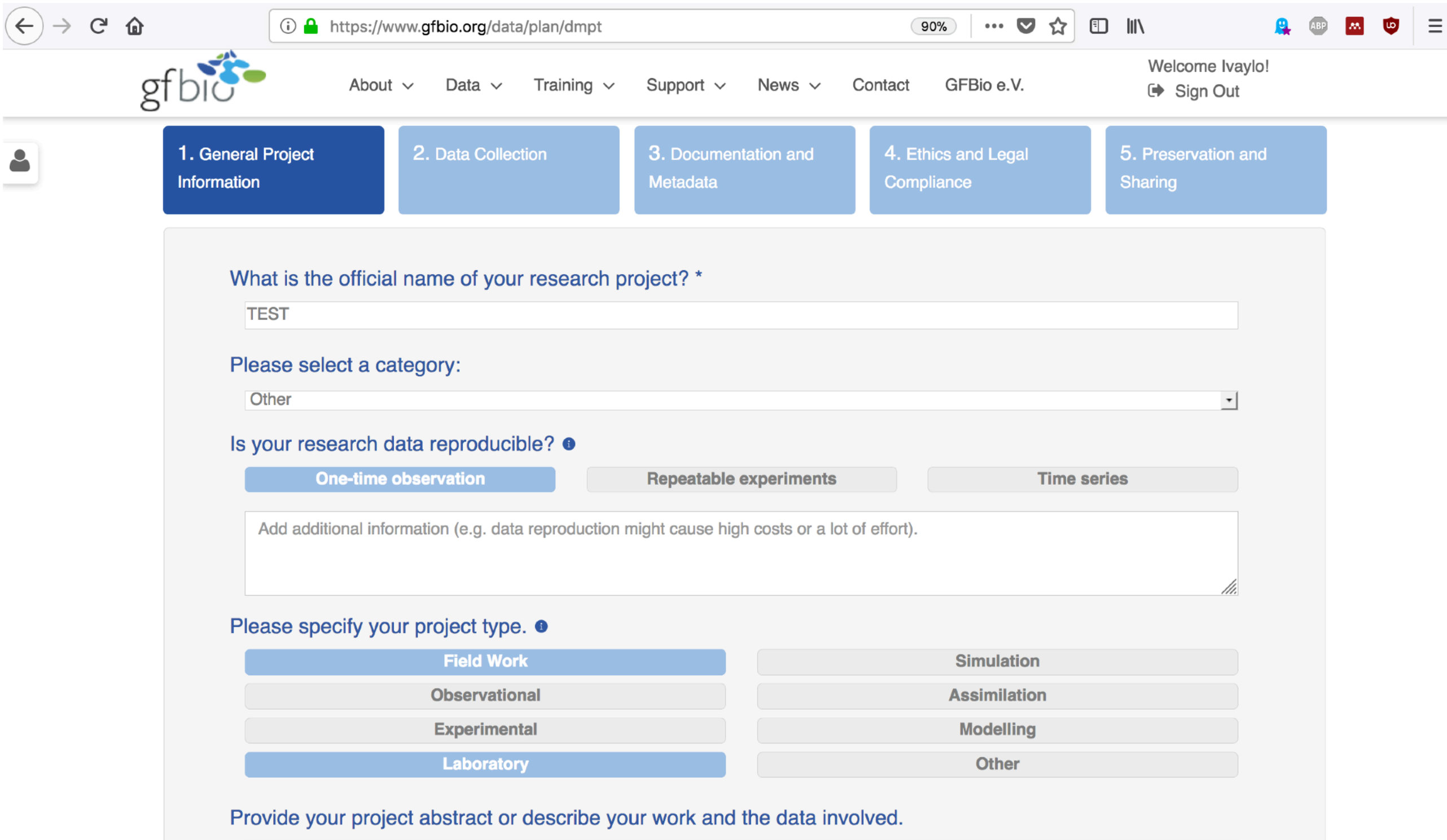
TERMINOLOGY SERVICE

Use the GFBio Terminology Service to describe your data and share terminologies with other researchers.

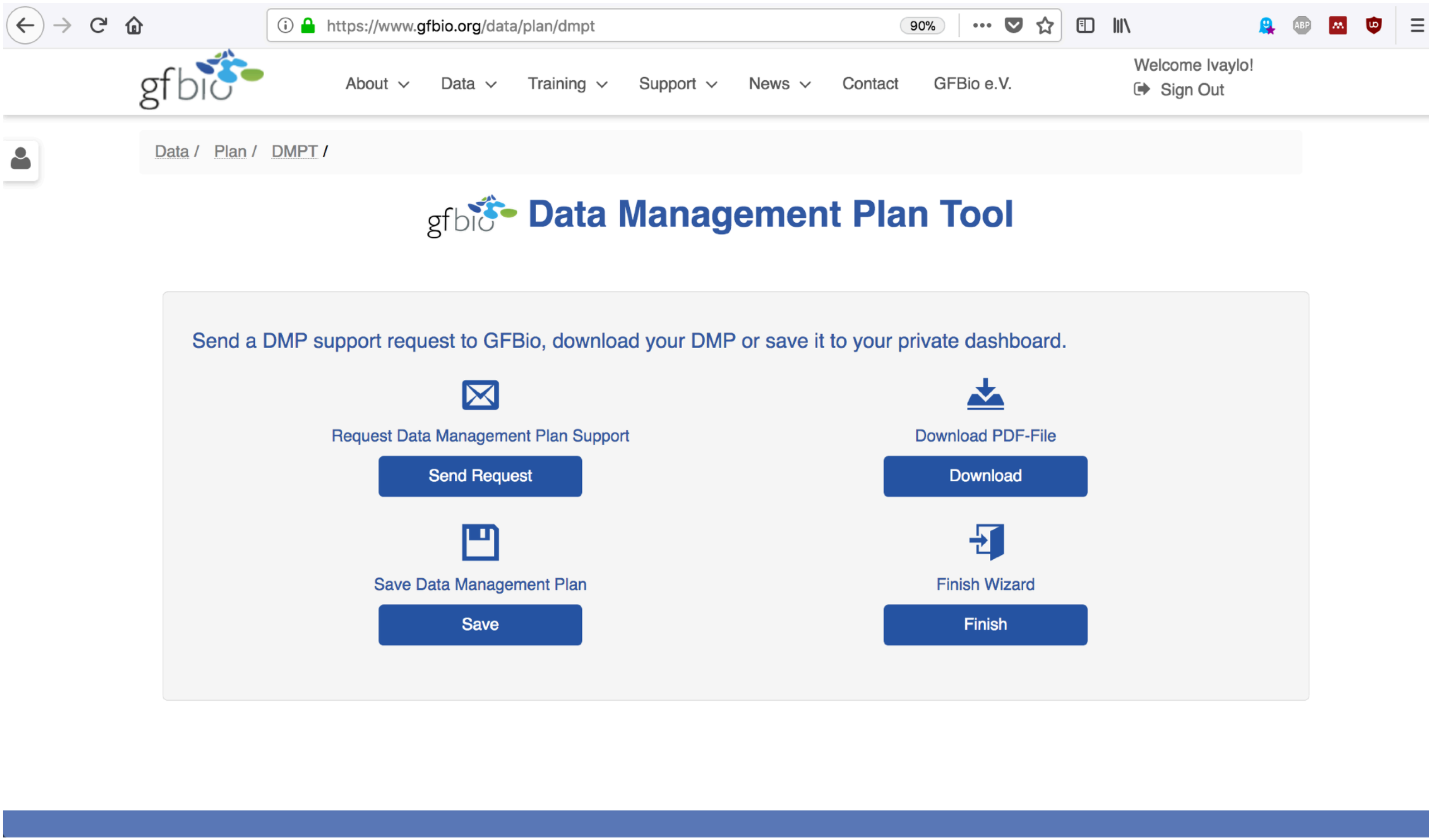
DMP Support


https://www.gfbio.org/data/plan
About
Data
Training
Support
News
Contact
GFBio e.V.
Sign In
Welcome to the
GFBio Data Management Plan Tool!
✓Collect information about your project
✓Complete your DMP checklist
✓Get GFBio DMP support
Get started
The GFBio Data Management Planning Tool supports you in preparing your custom DMP. It helps you think about the most important questions concerning data management as early as possible. Collect information about your project, fill in the DMP checklist and send us your DMP support request. We will support you in optimizing your data management and finalizing your data management plan.
Learn more: How to create a data management plan (DMP)?

DMP Support


https://www.gfbio.org/data/plan/dmpt
90%
gfbio
About
Data
Training
Support
News
Contact
GFBio e.V.
Welcome Ivaylo!
Sign Out
1. General Project Information
2. Data Collection
3. Documentation and Metadata
4. Ethics and Legal Compliance
5. Preservation and Sharing
What is the official name of your research project? *
TEST
Please select a category:
Other
Is your research data reproducible?
One-time observation
Repeatable experiments
Time series
Add additional information (e.g. data reproduction might cause high costs or a lot of effort).
Please specify your project type.
Field Work
Observational
Experimental
Laboratory
Simulation
Assimilation
Modelling
Other
Provide your project abstract or describe your work and the data involved.

DMP Support


https://www.gfbio.org/data/plan/dmpt
90%
About
Data
Training
Support
News
Contact
GFBio e.V.
Welcome Ivaylo!
Sign Out
Data / Plan / DMPT /
Data Management Plan Tool
Send a DMP support request to GFBio, download your DMP or save it to your private dashboard.
Request Data Management Plan Support
Send Request
Download PDF-File
Download
Save Data Management Plan
Save
Finish Wizard
Finish

DMP training opportunity

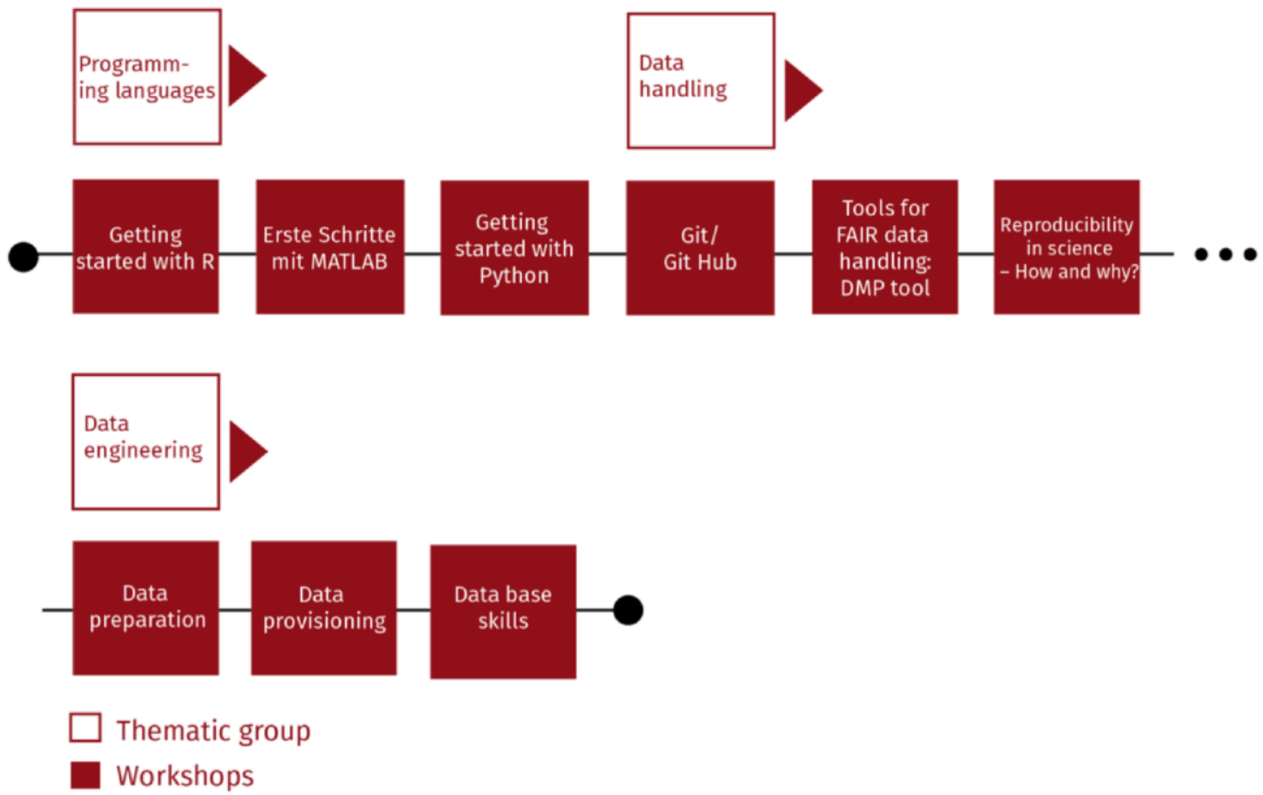
Universität Bremen

UBremen Research Alliance

The Alliance | Welcome Center | Research | Research data | Artificial Intelligence

Operator Track Data Steward

Start in June 2022



Workshops will start in June 2022

Hands-on workshops focussing on data steward skills for scientists.

Registration will start on 28 April 2022

To the courses in 2021! >

DMP training opportunity

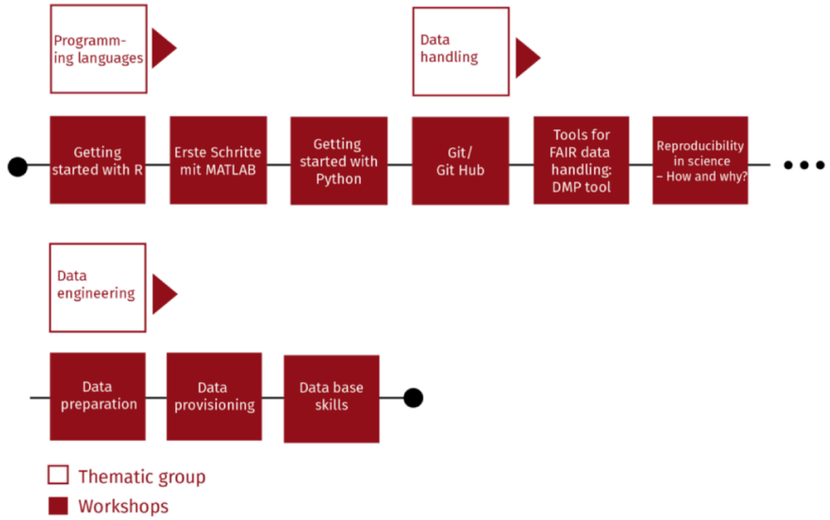
Universität Bremen

UBremen Research Alliance

The Alliance | Welcome Center | Research | Research data | Artificial Intelligence

Operator Track Data Steward

Start in June 2022



Workshops will start in June 2022

Hands-on workshops focussing on data steward skills for scientists.

Registration will start on 28 April 2022

To the courses in 2021!

UBRA Data Train
Data Steward Operator Track
Tools for FAIR data handling: DMP Tool

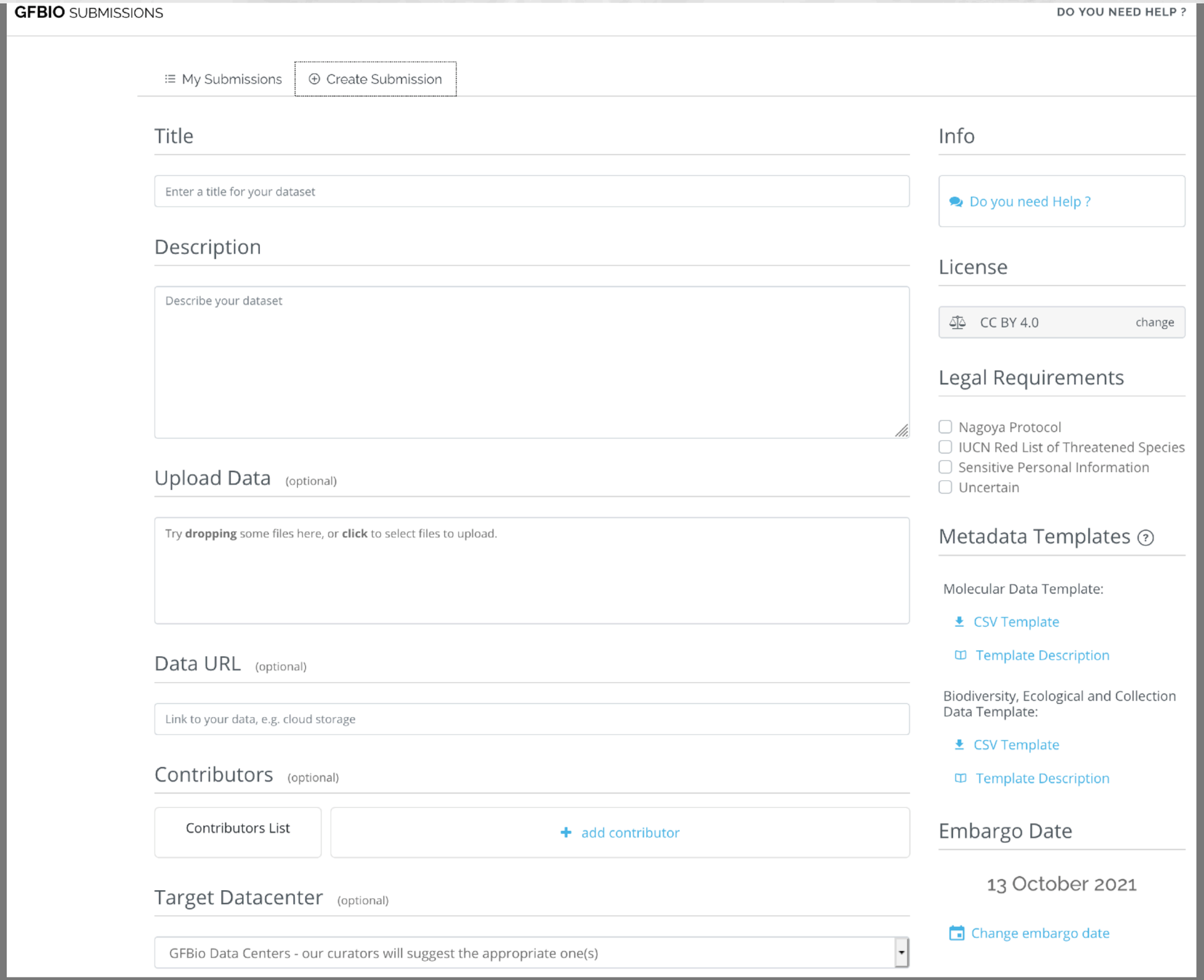
Jimena Linares

https://www.uni-bremen.de/en/research-alliance/research-data/data-train/data-train-curriculum

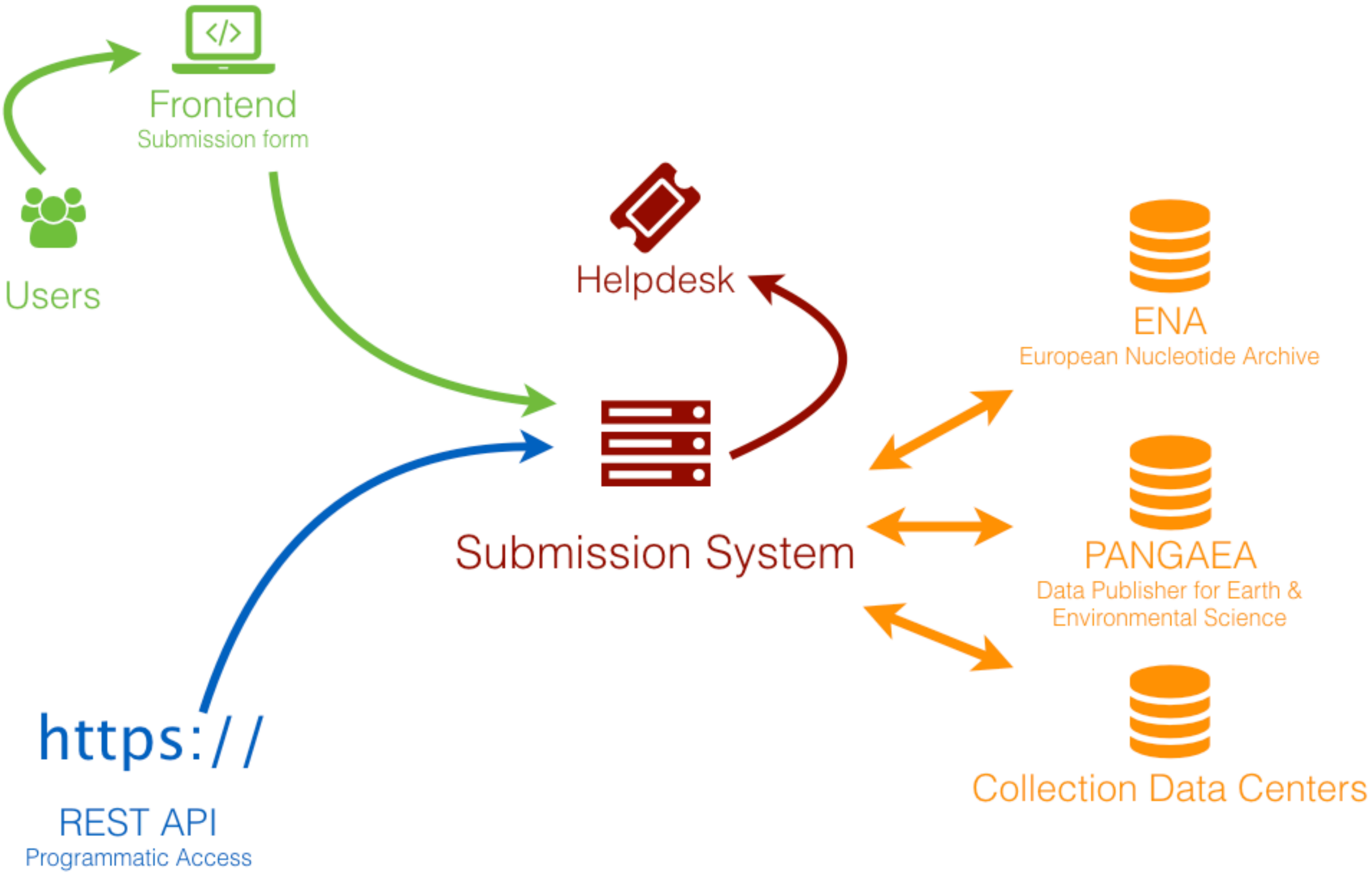
Data Submission

https://submissions.gfbio.org

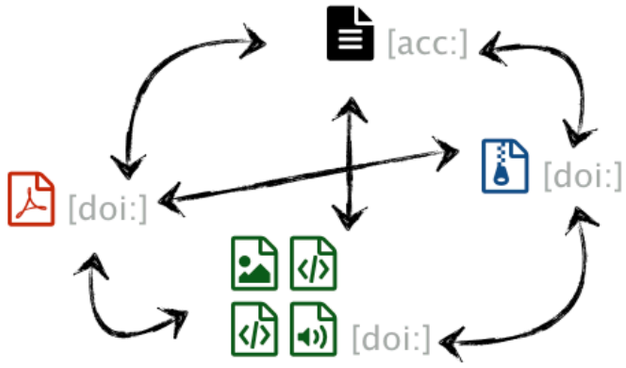
Data Submission



Data Submission



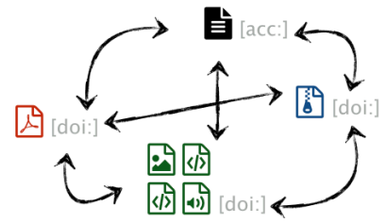
Data Submission



Added value at a glance:

- Single-point of contact - data is distributed to data centers and interlinked
- Expert support for metadata standardization - ABCD, DwC, **MIxS, ENVO**
- Manual and programmatic (API) operation
- Integration with local RDM systems

Data publication & citation



europepmc.org/abstract/PMC/PMC6265507

Sign in or create an account

Europe PMC | About | Tools | Developers | Help | Explore the beta version | Europe PMC plus

Search worldwide, life-sciences literature

Search | Advanced Search

E.g. "breast cancer" HER2 Smith J

☐ Environmental Drivers of Free-Living vs. Particle-Attached Bacterial Community Composition in the Mauritania Upwelling System.
(PMCID:PMC6265507)

Recent Activity | Export

Formats

Abstract | Full Text | PDF

Abstract | Citations | Related Articles | Data | BioEntities | External Links

Bachmann J[1], Heimbach T[1], Hassenrück C[1], Kopprio GA[1], Iversen MH[2], Grossart HP[3], Gärdes A[1]

Affiliations

Frontiers in Microbiology [23 Nov 2018, 9:2836]

Type: research-article, Journal Article
DOI: 10.3389/fmicb.2018.02836

Abstract

Saharan dust input and seasonal upwelling along North-West Africa provide a model system for studying microbial processes related to the export and recycling of nutrients. This study offers the first molecular characterization of prokaryotic particle-attached (PA; >3.0 µm) and free-living (FL; 0.2-3.0 µm) players in this important ecosystem during August 2016. Environmental drivers for alpha-diversity, bacterial community composition, and differences between FL and PA fractions were identified. The ultra-oligotrophic waters off Senegal were dominated by Cyanobacteria while higher relative abundances of Alphaproteobacteria, Bacteroidetes, Verrucomicrobia, and Planctomycetes (known particle-degraders) occurred in the upwelling area. Temperature, proxy for different water masses, was the best predictor for changes in FL communities. PA community variation was best explained by temperature and ammonium. Bray Curtis dissimilarities between FL and PA were generally very high and correlated with temperature and salinity in surface waters. Greatest similarities between FL and PA occurred at the deep chlorophyll maximum, where bacterial substrate

Show annotations in this abstract

☐ Chemicals
☐ Organisms

Data Citations

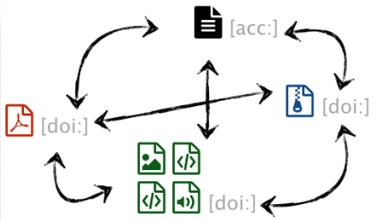
Identified 2 unique Data Citations in the full text

doi 10.1594/PANGAEA.889977 (2)

bioproject PRJEB26997 (1)

Show all items

Data publication & citation



European Bioinformatics Inst... (GB) | https://www.ebi.ac.uk/ena/data/vi 90%

ENA
European Nucleotide Archive

Search
Examples: BN000065, histone
Advanced
Sequence

Home | Search & Browse | Submit & Update | Software | About ENA | Support

Contact Helpdesk

Study: PRJEB26997

Environmental drivers of free-living vs. particle-attached bacterial community composition in the Mauritania upwelling system

View: Project XML Study XML **Download:** Project XML Study XML

Name
8733:dd9bf50b-64cf-4be3-9a80-18d39b289d50

Submitting Centre
Leibniz Centre for Tropical Marine Research

Secondary accession(s)
ERP109032

Broker Name
German Federation for Biological Data (GFBio)

Description
This study provides the first molecular characterisation of the bacterial community composition off the coast of Mauritania and Senegal. During the Meteor M129 cruise coastal, shelf and open water samples were taken for 16S Illumina MiSeq sequencing of both particle-attached (PA; >3.0 µm) and free-living (FL; 0.2-3.0 µm) fraction. Additionally, environmental drivers for α-diversity, bacterial community composition and differences between PA and FL bacteria were identified.

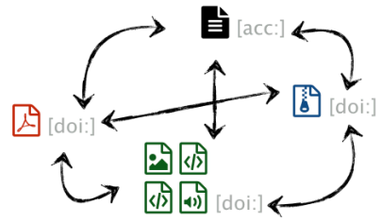
Navigation | Read Files | Portal | Attributes | Publications

Publications citing this record

[1] Bachmann, Jennifer; Heimbach, Tabea; Hassenrück, Christiane; Kopprio, Germa; Iversen, Morten Hvitfeldt; Grossart, Hans-Peter; Gärdes, Astrid (2018): Bacterial community composition off the coast of Mauritania and Senegal during METEOR cruise M129. PANGAEA, https://doi.org/10.1594/PANGAEA.889977, Supplement to: Bachmann, J et al. (2018): Environmental Drivers of Free-Living vs. Particle-Attached Bacterial Community Composition in the Mauritania Upwelling System. Frontiers in Microbiology, 9, https://doi.org/10.3389/fmicb.2018.02836
DOI

[2] **Environmental Drivers of Free-Living vs. Particle-Attached Bacterial Community Composition in the Mauritania Upwelling System.**
Bachmann J., Heimbach T., Hassenrück C., Kopprio GA., Iversen MH., Grossart HP., Gärdes A.
Leibniz Centre for Tropical Marine Research (ZMT), Bremen, Germany.
Front Microbiol 92836 (2018)
Show abstract
Europe PMC
PDF
doi

Data publication & citation



https://doi.pangaea.de/10.1594/PANGAEA.889977 90%

Not logged in

PANGAEA.

Data Publisher for Earth & Environmental Science

SEARCH SUBMIT ABOUT CONTACT

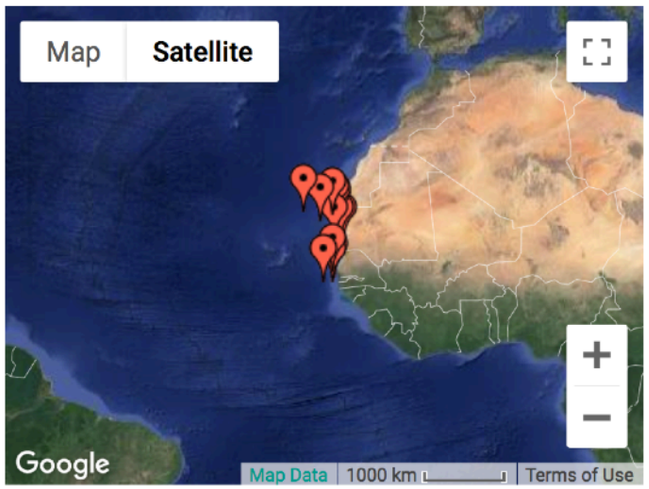
Citation:

Bachmann, Jennifer; Heimbach, Tabea; Hassenrück, Christiane; Kopprio, Germa; Iversen, Morten Hvitfeldt; Grossart, Hans-Peter; Gärdes, Astrid (2018): Bacterial community composition off the coast of Mauritania and Senegal during METEOR cruise M129. *PANGAEA*, https://doi.org/10.1594/PANGAEA.889977,

Supplement to: Bachmann, J et al. (2018): Environmental Drivers of Free-Living vs. Particle-Attached Bacterial Community Composition in the Mauritania Upwelling System. *Frontiers in Microbiology*, **9**, https://doi.org/10.3389/fmicb.2018.02836

Always quote above citation when using data! You can download the citation in several formats below.

RIS Citation | BibTeX Citation | Copy Citation | Facebook | Twitter | Show Map | Google Earth



Abstract:

Saharan dust input and seasonal upwelling along North–West Africa provide a model system for studying microbial processes related to the export and recycling of nutrients. This study offers the first molecular characterization of prokaryotic particle-attached (PA; >3.0 µm) and free-living (FL; 0.2–3.0 µm) players in this important ecosystem during August 2016. Environmental drivers for alpha-diversity, bacterial community composition, and differences between FL and PA fractions were identified. The ultra-oligotrophic waters off Senegal were dominated by Cyanobacteria while higher relative abundances of Alphaproteobacteria, Bacteroidetes, Verrucomicrobia, and Planctomycetes (known particle-degraders) occurred in the upwelling area. Temperature, proxy for different water masses, was the best predictor for changes in FL communities. PA community variation was best explained by temperature and ammonium. Bray Curtis dissimilarities between FL and PA were generally very high and correlated with temperature and salinity in surface waters. Greatest similarities between FL and PA occurred at the deep chlorophyll maximum, where bacterial substrate availability was likely highest. This indicates that environmental drivers do not only influence changes among FL and PA communities but also differences between them. This could provide an explanation for contradicting results obtained by different studies regarding the dissimilarity/similarity between FL and PA communities and their biogeochemical functions.

Related to:

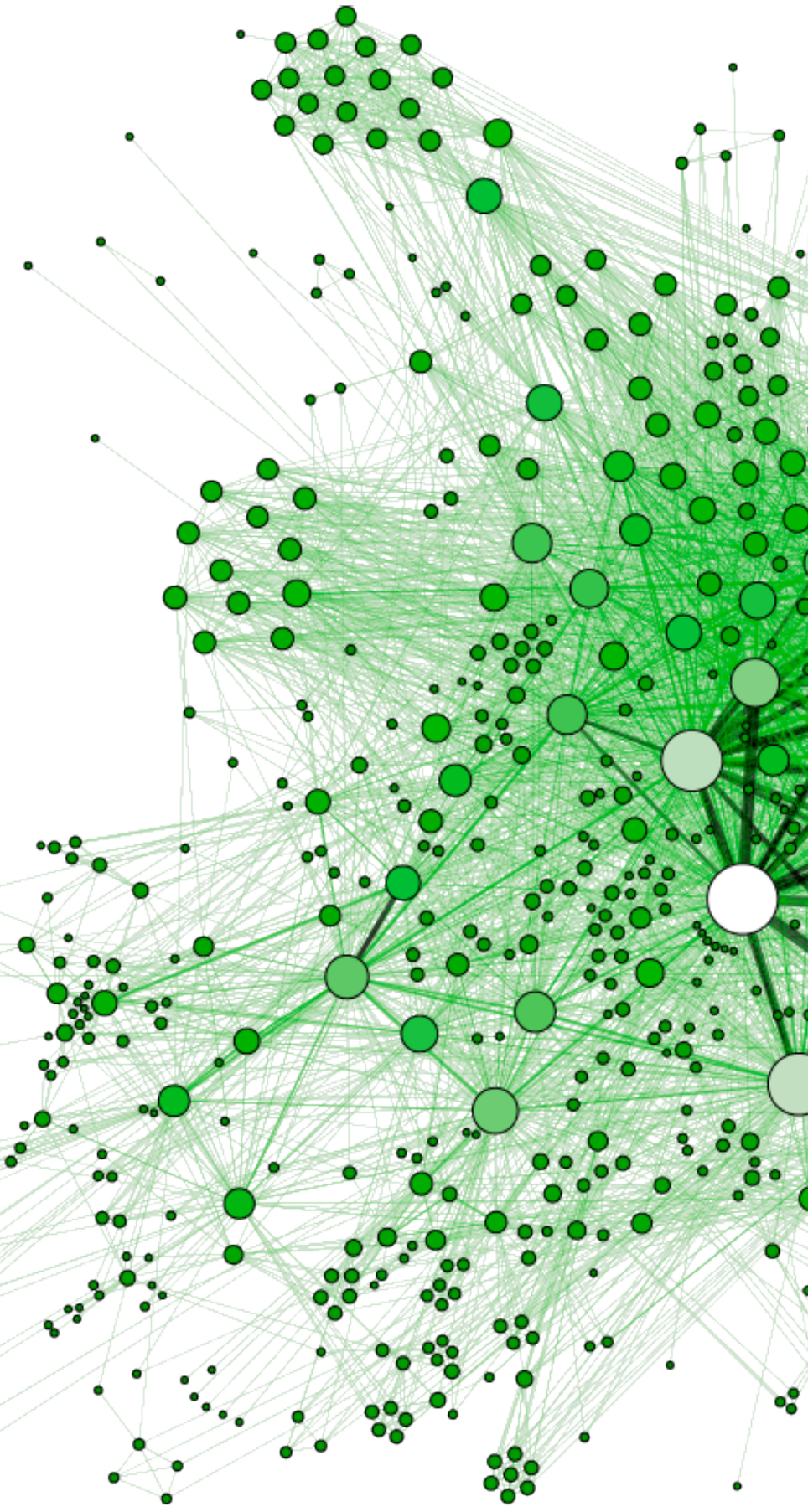
Bachmann, Jennifer; Heimbach, Tabea; Hassenrück, Christiane; Kopprio, Germa; Iversen, Morten Hvitfeldt; Grossart, Hans-Peter; Gärdes, Astrid (2018): Environmental drivers of free-living vs. particle-attached bacterial community composition in the Mauritania upwelling system. *European Nucleotide Archive (ENA)*, insdc:PRJEB26997

Project(s):

Leibniz Centre for Tropical Marine Research (ZMT)

NFDI 4
BIODIVERSITY

NFDI4Biodiveristy

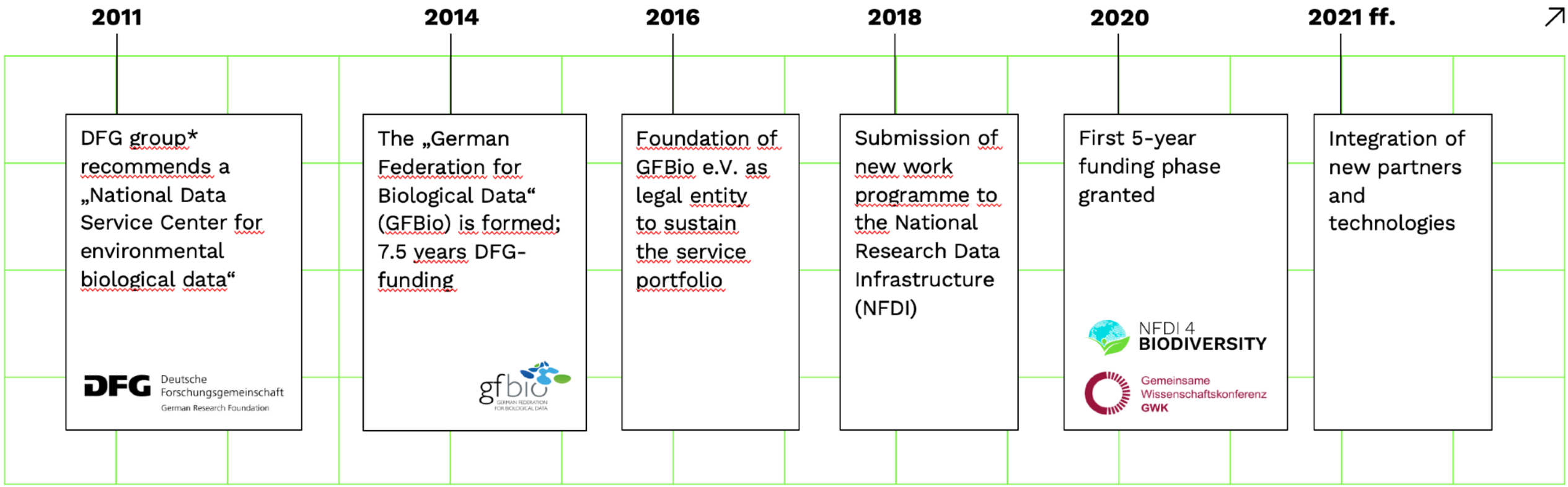


www.nfdi4biodiversity.de

info@gfbio.org

@NFDI4BioDiv

A short history



NFDI 4
BIODIVERSITY

Data submission, versioning and publication
Including long-term preservation (extended GFBio services)

Helpdesk
Individual support for researchers and data centers (extended GFBio service)

Support with integration and harmonization of data
(GFBio data centers)

Provision of collaborative workspaces
With support for scientific workflows and provenance management

Education and Training
Tailored events, tools and materials for teaching

Basic tools for data managers
Validation, transformation, automated quality checks various de.NBI tools.

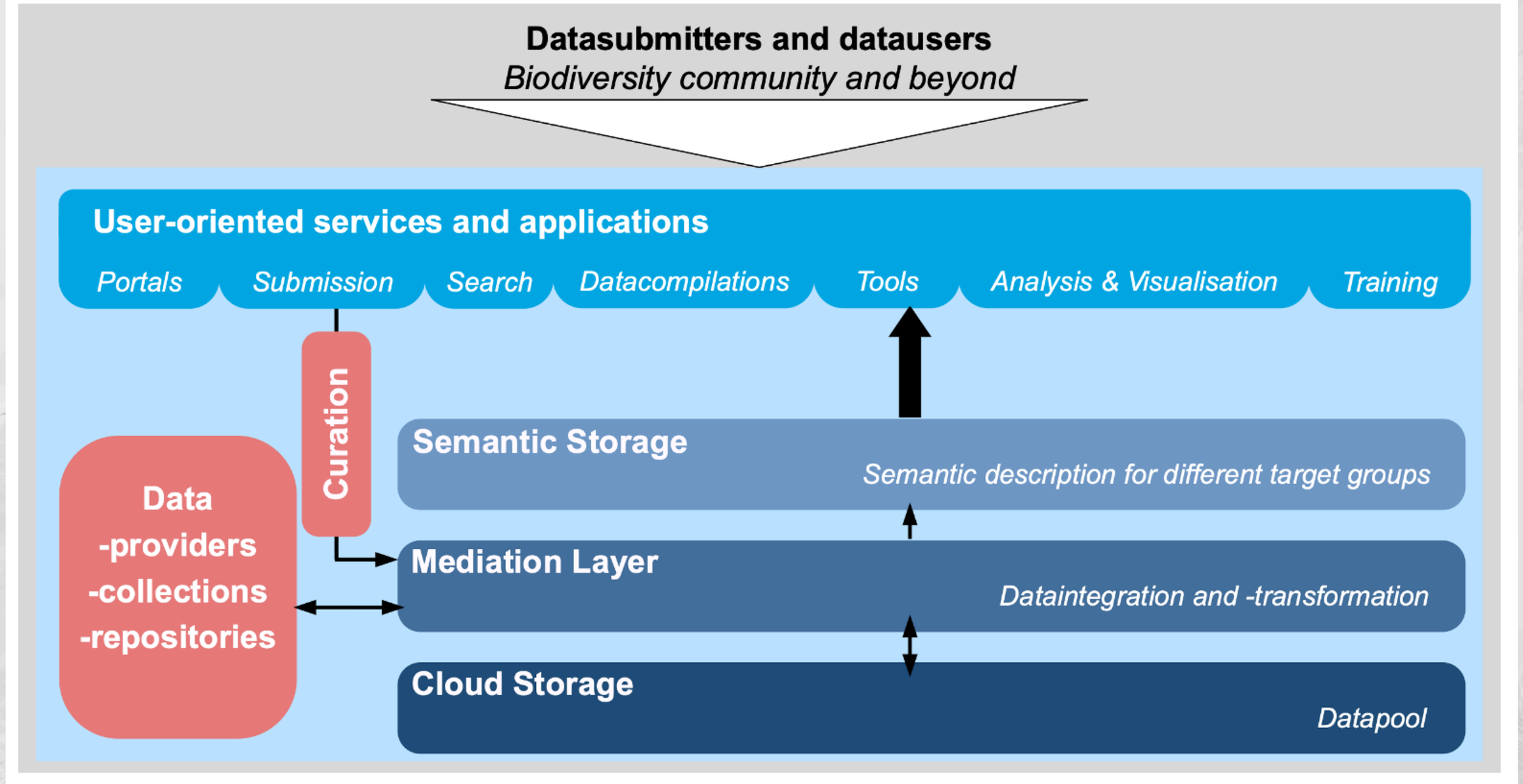
Terminology service
(extended GFBio service)

Search portals and API for data and tools
(extended GFBio services)

Elastic compute service
(Infrastructure-as-a-service)
In future: Research Data Commons

NFDI 4 BIODIVERSITY

Vision: **R**esearch **D**ata **C**ommons
Cloud-based platform for applications and data



Thank you!

www.nfdi4biodiversity.org

www.gfbio.org

contact@nfdi4biodiveristy.org (collaboration)

info@gfbio.org (service help desk)

@NFDI4Biodiv

NFDI 4 BIODIVERSITY

Questions?