

# Research Data Management

## GFBio services for archiving and publishing sequence data

Ivaylo Kostadinov, Ph.D.

GFBio e.V.

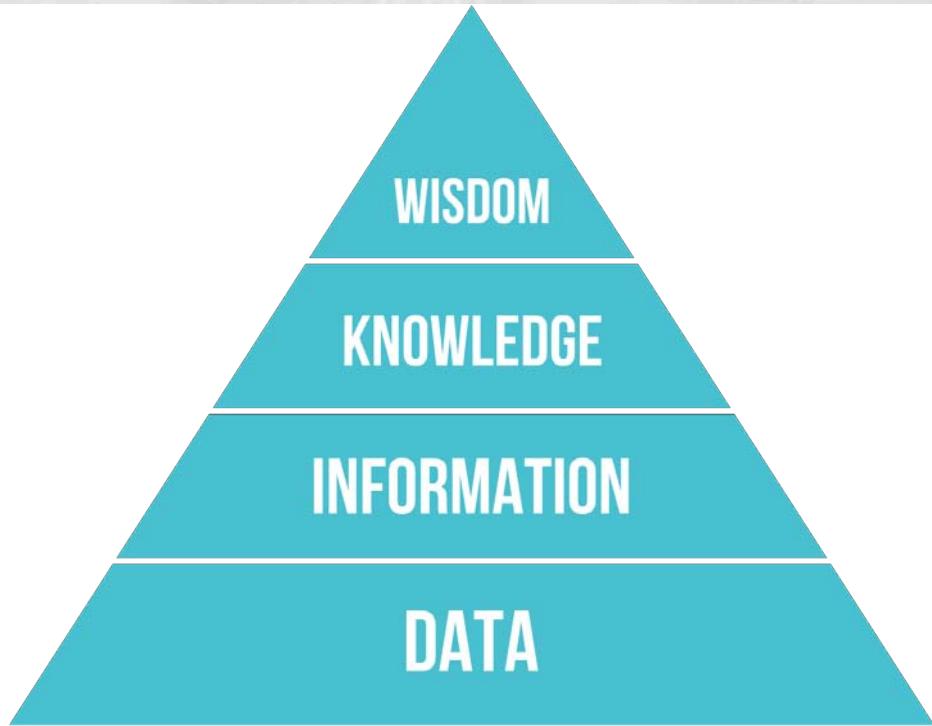
✉ ivo@gfbio.org

🐦 @tigroumaniac



# DATA

# Data fuels research



[https://upload.wikimedia.org/wikipedia/commons/0/06/DIKW\\_Pyramid.svg](https://upload.wikimedia.org/wikipedia/commons/0/06/DIKW_Pyramid.svg)

Ivaylo Kostadinov | GFBio e.V. | ZMT NGS Workshop 2019-03-11

3

## Research Funding in DE



# 90 Billion €

<http://www.dfg.de/sites/foerderatlas2018>

Ivaylo Kostadinov | GFBio e.V. | ZMT NGS Workshop 2019-03-11

4

# DFG Funding Atlas



**7.4** Billion Euro of third party funding was shared between Universities in Germany in 2015

<http://www.dfg.de/sites/foerderatlas2018>

Förderatlas 2018 KARTENANSICHT PUBLIKATION THEMEN DOWNLOADS KONTAKT DFG Deutsche Forschungsgemeinschaft

## Förderatlas 2018

Kennzahlen zur öffentlich finanzierten  
Forschung in Deutschland

ALS E-PAPER ÖFFNEN

Ivaylo Kostadinov | GFBio e.V. | ZMT NGS Workshop 2019-03-11

5

## Time effort



.. for discovering and reusing multiple data sources

**80%**

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

Ivaylo Kostadinov | GFBio e.V. | ZMT NGS Workshop 2019-03-11

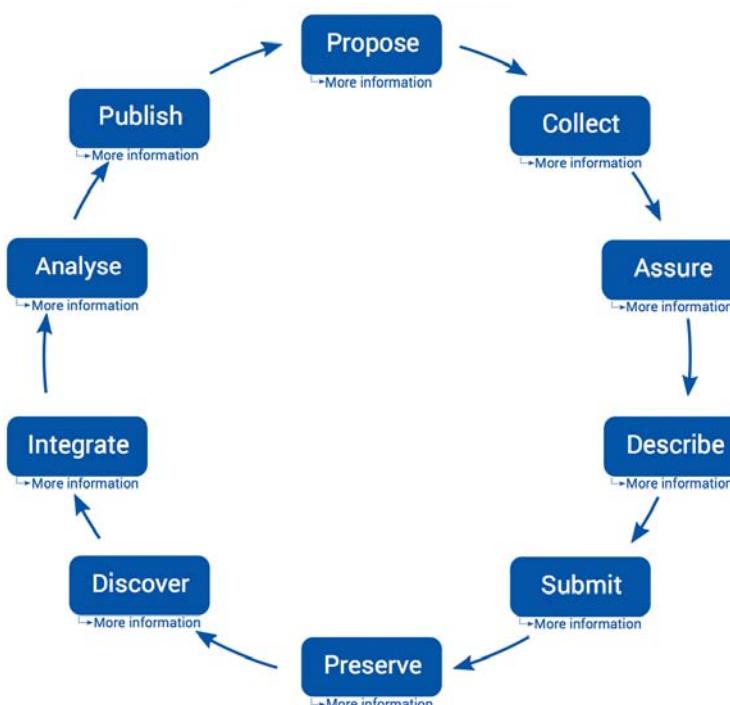
6

# dark data

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

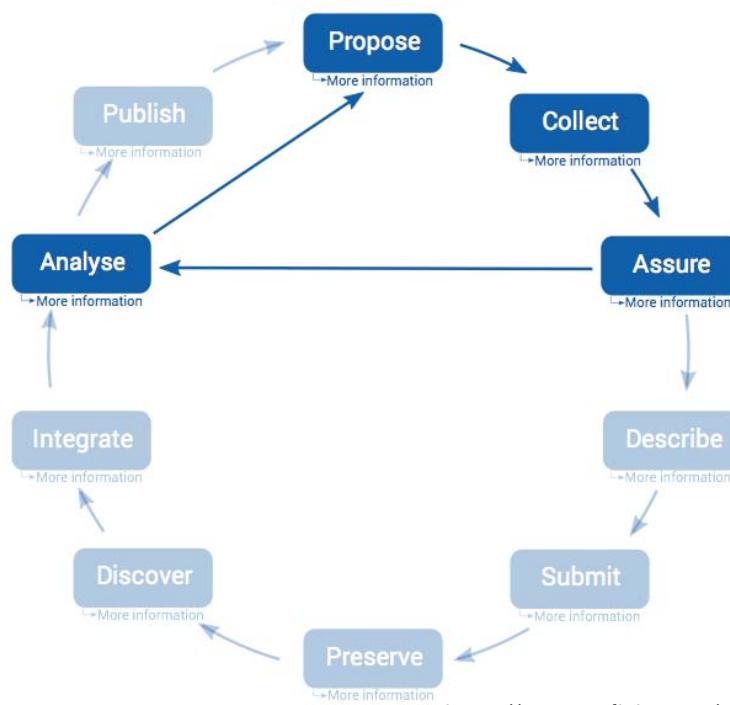
B. P. Heidorn Libr. Trends 57, 280–299; 2008

## Data Lifecycle



<http://www.gfbio.org/training/data-lifecycle>

# Data Lifecycle

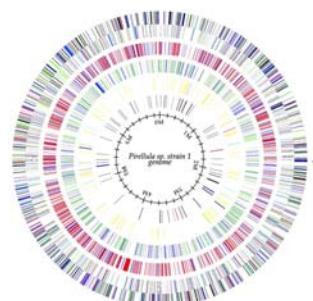


<http://www.gfbio.org/training/data-lifecycle>

Ivaylo Kostadinov | GFBio e.V. | ZMT NGS Workshop 2019-03-11

9

## Metadata



Glöckner et al., 2003

DATA

land use nitrate salinity host relationship cell  
size motility calcium perturbation 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 CO<sub>2</sub> cultivation temperature

Courtesy: Boyke Bunk, DSMZ

METADATA  
contextual data

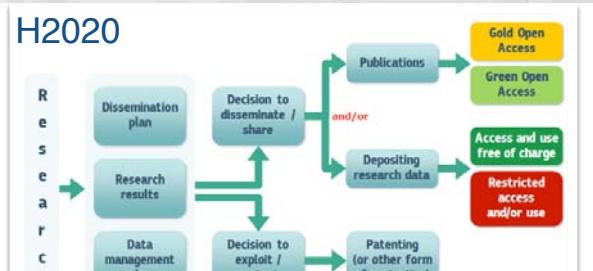
# VIDEO

Data Sharing and Management Snafu in 3 Short Acts by [NYU Health Sciences Library](#)  
<https://youtu.be/N2zK3sAtr-4>

Ivaylo Kostadinov | GFBio e.V. | ZMT NGS Workshop 2019-03-11

11

## Community response

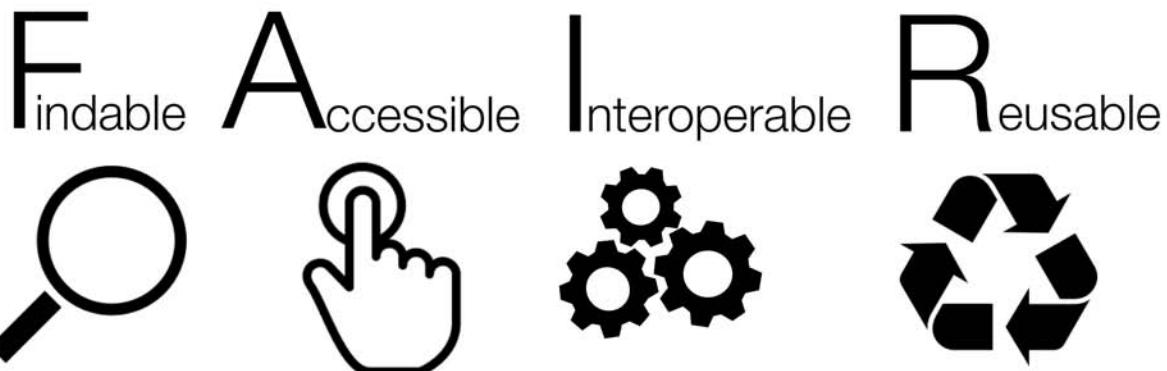


The following general guidelines apply for applicants submitting proposals to the DFG:

- ▶ 1. Project planning and submission of proposal
- ▶ 2. Accessibility
- ▶ 3. Long-term archiving



[http://www.dfg.de/en/research\\_funding/  
proposal\\_review\\_decision/applicants/  
submitting\\_proposal/research\\_data/](http://www.dfg.de/en/research_funding/proposal_review_decision/applicants/submitting_proposal/research_data/)



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

Ivaylo Kostadinov | GFBio e.V. | ZMT NGS Workshop 2019-03-11

13

## FAIR Data

# SCIENTIFIC DATA



**OPEN**  
SUBJECT CATEGORIES

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

Mark D. Wilkinson et al.\*

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 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 exemplary implementations in the community.

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

**Findable  
Accessible  
Interoperable  
Reusable**

## 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>

## FAIR is not

- a standard
- equal to open data
- a quality, but a quantity
- only for humans or only for machines
- only for life sciences
- equal to RDF, Linked Data, or Semantic Web

# Your incentives to be FAIR



- Good scientific practice
- Career boost
  - article acceptance
  - data reuse & citation
  - proposal funding
  - compatibility with future infrastructures
- Career opportunities as a data scientist, manager, steward, custodian, librarian, etc.
- Keep your research legal (i.e. avoid biopiracy)

Ivaylo Kostadinov | GFBio e.V. | ZMT NGS Workshop 2019-03-11

17

## Biopiracy



Image: <https://www.flickr.com/photos/ciat/3887465932>

Biopiracy happens when researchers or research organisations take biological resources without official sanction, largely from less affluent countries or marginalised people.

<http://theconversation.com/biopiracy-when-indigenous-knowledge-is-patented-for-profit-55589>

Ivaylo Kostadinov | GFBio e.V. | ZMT NGS Workshop 2019-03-11

18



## The Nagoya Protocol on Access and Benefit-sharing

A transparent legal framework for the fair and equitable sharing of benefits arising out of the utilization of genetic resources.

<https://www.cbd.int/abs/about/>

# STANDARDS

# Sequence Metadata



## MIxS Minimal Information about any(x) Sequence

<http://gensc.org/mixs/>

developed by:



supported by:



MIGS/MIMS - Field et al., Nature Biotechnology 26, 541 - 547 (2008)  
MIMARKS & MIxS - Yilmaz et al., Nature Biotechnology 29, 415–420 (2011)  
MIxS - Yilmaz et al., The ISME Journal 5, 1565–1567 (2011)

Ivaylo Kostadinov | GFBio e.V. | ZMT NGS Workshop 2019-03-11

21

# Sequence Metadata



## MIxS Minimal Information about any(x) Sequence

Item	MIGS				MIMS MIMARKS			
	EU	PA	PV	VI	ORG	ME	GU	SR
Project name	M	M	M	M	M	M	M	M
Collection date	M	M	M	M	M	M	M	M
Geographic location	M	M	M	M	M	M	M	M
Environment	M	M	M	M	M	M	M	M
Phylo	M	-	-	-	-	-	-	-
Estimated size	M	X	X	X	X	-	-	-
Target gene	-	-	-	-	-	M	M	M

**Environmental Packages**

Item	Human-gut	Soil	Water
Age	X	-	-
Body site	X	-	-
Elevation	-	M	C
Depth	-	M	M
Horizon	-	X	-
pH	-	X	X
Salinity	-	-	X



<http://gensc.org/mixs/>

MIGS/MIMS - Field et al., Nature Biotechnology 26, 541 - 547 (2008)  
MIMARKS & MIxS - Yilmaz et al., Nature Biotechnology 29, 415–420 (2011)  
MIxS - Yilmaz et al., The ISME Journal 5, 1565–1567 (2011)

Ivaylo Kostadinov | GFBio e.V. | ZMT NGS Workshop 2019-03-11

22

# What else is out there?



Standards for biodiversity & ecology:

- Darwin Core (Wieczorek et al. 2012)
- ABCD (Holetschek et al. 2012)
- GGBN (Dröge et al. 2016)
- EML (<https://knb.ecoinformatics.org/#tools/eml>)

Ivaylo Kostadinov | GFBio e.V. | ZMT NGS Workshop 2019-03-11

23

GFBio



[www.gfbio.org](http://www.gfbio.org)



A **sustainable**, **service-oriented** data infrastructure to facilitate **data sharing** and **data-intensive** science in biology and environmental research.

Funded by:



Ivaylo Kostadinov | GFBio e.V. | ZMT NGS Workshop 2019-03-11

24

[www.gfbio.org](http://www.gfbio.org)



www.gfbio.org | Search

About ▾ Data ▾ Training ▾ Support ▾ News Contact Sign In Search...

  
GERMAN FEDERATION  
FOR BIOLOGICAL DATA

Research Data Management  
Step by step through the Data Life Cycle.

GET MORE INFORMATION

Submit Search Manage Workbenches

Ivaylo Kostadinov | GFBio e.V. | ZMT NGS Workshop 2019-03-11 25

# Interdisciplinary team



www.gfbio.org | Geduld

About ▾ Data ▾ Training ▾ Support ▾ News Contact Sign In

## CONSORTIUM

GFBio's multidisciplinary consortium is represented by experts from a variety of institutions ranging from natural history collections, libraries, bioinformatics to environmental data archives



### data centers and publishers

### infrastructure partners

### user-representatives

Ivaylo Kostadinov | GFBio e.V. | ZMT NGS Workshop 2019-03-11 26

# GFBio Services

The Key Features of our Work

			
<b>PLAN</b> Prepare a custom Data Management Plan (DMP).	<b>SUBMIT</b> Submit your data to GFBio.	<b>SEARCH</b> Search the GFBio data pool.	<b>VISUALIZE &amp; ANALYZE</b> Dynamically integrate, analyze and visualize GFBio datasets.
			
<b>PUBLISH</b> Make your data citable.	<b>TRAIN</b> Train your data management skills.	<b>ARCHIVE</b> Deposit data and specimens in dedicated long-term archives.	<b>TERMINOLOGY SERVICE</b> Use the GFBio Terminology Service to describe your data and share terminologies with other researchers.

Ivaylo Kostadinov | GFBio e.V. | ZMT NGS Workshop 2019-03-11

# Data Management Plan

- A good DMP should cover:
  - Data acquisition
  - Quality assurance
  - Intermediate handling and storage
  - Long-term archiving
  - Analysis
  - Publication (open-access, licensing)

Ivaylo Kostadinov | GFBio e.V. | ZMT NGS Workshop 2019-03-11

# Data Management Plan



- **Basic information:** project title, contact information, motivation for data collection
- **Information on data:** type, format, volume, collection standards, methodologies, quality assurance
- **Documentation and metadata:** readability and interpretability of data, metadata standards
- **Ethical and legal compliance:** agreement on preservation/sharing conditions, sensitive data, intellectual property
- **Storage and backup plan:** responsibility, data recovery, access for collaborators, security

<https://www.gfbio.org/training/materials/data-lifecycle/plan>

# Data Management Plan



- **Preservation:** selection of data, foreseeable future use, time and location for preservation, costs
- **Data sharing and publication:** modalities, conditions, persistent identifiers
- **Responsibilities:** implementation, roles and responsibilities for each activity, ownership agreement
- **Resources:** need for additional hardware/software or expertise for training, efforts and costs for data management and data archiving

<https://www.gfbio.org/training/materials/data-lifecycle/plan>

# 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 \(DMP\)?](#)

Ivaylo Kostadinov | GFBio e.V. | ZMT NGS Workshop 2019-03-11

31

# DMP Support



https://www.gfbio.org/data/plan/dmpt

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	Simulation
Observational	Assimilation
Experimental	Modelling
Laboratory	Other

Provide your project abstract or describe your work and the data involved.

Ivaylo Kostadinov | GFBio e.V. | ZMT NGS Workshop 2019-03-11

32

# DMP Support



https://www.gfbio.org/data/plan/dmpt

About Data Training Support News Contact GFBio e.V. Welcome Ivaylo! Sign Out

Data / Plan / DMPT /

## gfbio 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 Download PDF-File

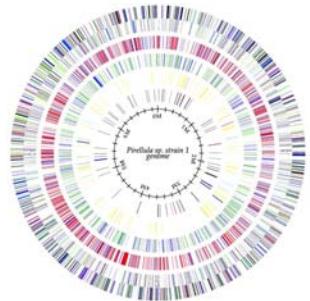
Save Data Management Plan Finish Wizard

Send Request Download Save Finish

Ivaylo Kostadinov | GFBio e.V. | ZMT NGS Workshop 2019-03-11 33

# DATA SUBMISSION

# Archival & Publication



Glöckner et al., 2003

DATA

land use nitrate salinity host relationship cell  
size motility calcium perturbation 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 CO<sub>2</sub> cultivation temperature

Courtesy: Boyke Bunk, DSMZ

METADATA  
contextual data

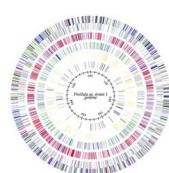
Ivaylo Kostadinov | GFBio e.V. | ZMT NGS Workshop 2019-03-11

35

# Archival & Publication



Findable  
Accessible  
Interoperable  
Reproducible



 ENA  
European Nucleotide Archive  
INSDC International Nucleotide Sequence Database Collaboration

land use nitrate salinity host relationship cell  
size motility calcium perturbation 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 CO<sub>2</sub> cultivation temperature



dedicated, long-term archives

Ivaylo Kostadinov | GFBio e.V. | ZMT NGS Workshop 2019-03-11

36

# Data Submission



① https://www.gfbio.org/data/submit

About Data Training Support News Contact GFBio e.V. Sign In

## Submit Your Data to a Public Repository

Transfer your data from your private research domain to the [GFBio data centers](#) for long-term archival and publication. Our curation experts will find the best solution for storing your data within the GFBio consortium and making it [FAIR](#) (findable, accessible, integratable and reusable).

[Start a data submission](#)

## Submit Molecular Sequence Data

Molecular sequence data are submitted to the [European Nucleotide Archive](#), any accompanying environmental data are archived in [PANGAEA](#). We will also help you apply the [MixS standard](#) to your meta-data.

[Start a Molecular Submission](#)

Ivaylo Kostadinov | GFBio e.V. | ZMT NGS Workshop 2019-03-11 37

# Molecular data submission

Welcome Ivaylo!  
Sign Out

### 1. Describe your Dataset

Title\*  
Provide a short, descriptive title for your dataset.

Description\*  
Provide a summary of the work you did to produce the dataset (similar to an article abstract).

Study Type\*  
Select the type of sequencing in your dataset. Choose 'Other' if you are not sure.

Data URL  
Provide an URL where we can access your sequence data. If you leave this field blank, you can upload your sequence files in the next step (after clicking "Start Submission").

Ivaylo Kostadinov | GFBio e.V. | ZMT NGS Workshop 2019-03-11 38

# Molecular data submission

The screenshot shows a web browser window with the URL <https://www.gfbio.org/data/submit/molecular>. The page has a header with the GFBio logo and navigation links for About, Data, Training, Support, News, Contact, and GFBio e.V. A user profile icon shows "Welcome Ivaylo!" and "Sign Out".

**Step 2: Describe your samples and sequencing procedure**

Provide information about your samples and experimental setup. Please download our empty CSV template, fill it out and use the upload button below to add it to this submission. Detailed documentation of the columns can be found [here](#).

[Upload CSV file](#) | [Download empty template](#)

**Step 3: Let's go**

[Start Submission](#)

**GFBio Consortium**  
The German Federation for Biological Data (GFBio), a sustainable, service oriented, national data infrastructure facilitating data sharing for biological and environmental research.

**Want to know more?**  
[Helpdesk](#) | [Training](#)

**Contact us!**  
Write our [Helpdesk Staff](#)! | Got questions? Email us at [helpdesk@gfbio.org](mailto:helpdesk@gfbio.org)

**Sign up**  
Get full access to services and resources deriving benefit for your Projects

Ivaylo Kostadinov | GFBio e.V. | ZMT NGS Workshop 2019-03-11

# 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, MIxS, ENVO
- Manual and programmatic (API) operation
- Integration with local RDM systems

<http://bexit2.uni-jena.de/>

<http://diversityworkbench.net>

Ivaylo Kostadinov | GFBio e.V. | ZMT NGS Workshop 2019-03-11



“Working with GFBio was excellent. It significantly aided and improved my data upload experience...”

# Our Services

The screenshot shows the GFBio website's "SERVICES" page. At the top right is the GFBio logo. The navigation bar includes links for About, Data, Training, Support, News, Contact, GFBio e.V., and Sign In. Below the navigation is a section titled "The Key Features of our Work" with eight service icons arranged in two rows of four. Each service has a title and a brief description.

PLAN	SUBMIT	SEARCH	VISUALIZE & ANALYZE
Prepare a custom Data Management Plan (DMP).	Submit your data to GFBio.	Search the GFBio data pool.	Dynamically integrate, analyze and visualize GFBio datasets.
PUBLISH	TRAIN	ARCHIVE	ANNOTATE & CONNECT
Make your data citable.	Train your data management skills.	Deposit data and specimens in dedicated long-term archives.	Use the GFBio Terminology Service to describe your data and share terminologies with other researchers.

Ivaylo Kostadinov | GFBio e.V. | ZMT NGS Workshop 2019-03-11

42

# Search

gfbio e.V.

About Data Training Support News Contact GFBio e.V. Sign In

Search: Soil composition

Show 10 entries per page

Showing 1 to 10 of 192 entries

**Strecker, Tanja; Gonzalez, Odette; Scheu, Stefan; Eisenhauer, Nico (2015): Spatial and temporal stability of soil microbial properties in the Jena Experiment (Germany) from 2003-2014**

Data Center: PANGAEA: Data Publisher for Earth & Environmental Science

Summary: The study was carried out on the main plots of a large grassland biodiversity experiment (the Jena Experiment). In the main experiment, 82 grassland plots of 20 x 20 m were established. The study focused on the spatial and temporal stability of soil microbial properties in the Jena Experiment. The data set includes measurements of soil microbial biomass, activity, and functional diversity at different depths (0-10 cm, 10-20 cm, 20-30 cm) and times (2003-2014).

License/Rights: CC-BY-3.0: Creative Commons Attribution 3.0 Unported

Data Description - Data Download

Bokhorst, Stef, Phoenix, Gareth K, Bjerke, Jarle W, Callaghan, Terry V, Huyer-Brugman, F, Berg, Matty P (2012): Soil characteristics and Collembola and Acari abundance in control and warming plots at Abisko Research Station

Data Center: PANGAEA: Data Publisher for Earth & Environmental Science

Summary: Extreme weather events can have negative impacts on species survival and community structure when surpassing lethal thresholds. Extreme winter

Filter Results: [clear filters |reset search]

Europe x

Author

- Niller, Hans-Peter(112)
- Weigelt, Alexandra(17)
- Weisser, Wolfgang(16)

Ivaylo Kostadinov | GFBio e.V. | ZMT NGS Workshop 2019-03-11

43

# Visualization & Analysis

gfbio e.V.

VAT: Visualization, Analysis and Tools

Secure https://vat.gfbio.org/#/

06.06.2000 12:00:00 - 06.06.2013 12:00:00

Faroe Islands

Herbarium Ber...  
The Bacterial D...  
Annual Precipit...

Add Data

- Environmental Add environmental raster layers
- ABCD Archives Lookup GFBio ABCD archives
- Custom Features Add and use custom vector features like CSV
- GFBio Baskets Display the GFBio Search Results
- Species Distribution Query data from GBIF and IUCN
- Draw Features Draw features on the map
- Country Selection Select Country Borders

Ivaylo Kostadinov | GFBio e.V. | ZMT NGS Workshop 2019-03-11

44

# Terminology Service



https://terminologies.gfbio.org

... ⌂ ⌂ ⌂ ⌂ ⌂ ⌂



Browse Search API Widgets About



## Explore

Browse and search for terminologies that can be interesting for your research.



## Access

Use information from terminologies programmatically to provide semantically enriched applications.



## Consume

Retrieve and store information from terminologies in your local information system.



## Contribute

Store or connect your terminologies to the TS and get access to all provided services automatically.

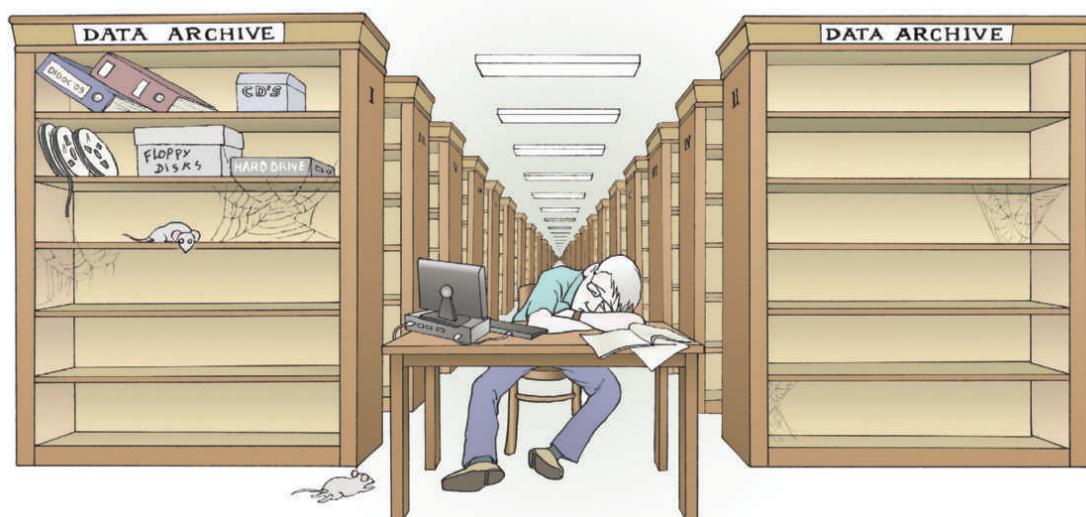
Ivaylo Kostadinov | GFBio e.V. | ZMT NGS Workshop 2019-03-11

45

# We need YOU!



If we build it they will come!?!?



Nelson, Nature 461, 160-163 (2009)

Ivaylo Kostadinov | GFBio e.V. | ZMT NGS Workshop 2019-03-11

46

# Sustainability model



**gfbio** e.V. [est. 2016]



BG | Botanischer Garten &  
BM | Botanisches Museum  
Berlin



SENCKENBERG  
world of biodiversity



SUB | Niedersächsische Staats- und  
Universitätsbibliothek Göttingen



Ivaylo Kostadinov | GFBio e.V. | ZMT NGS Workshop 2019-03-11

47

**gfbio** e.V.  
GERMAN FEDERATION  
FOR BIOLOGICAL DATA



Deutsche  
Forschungsgemeinschaft

✉ info@gfbi.org  
🐦 @GFBio\_Project  
💻 www.gfbi.org