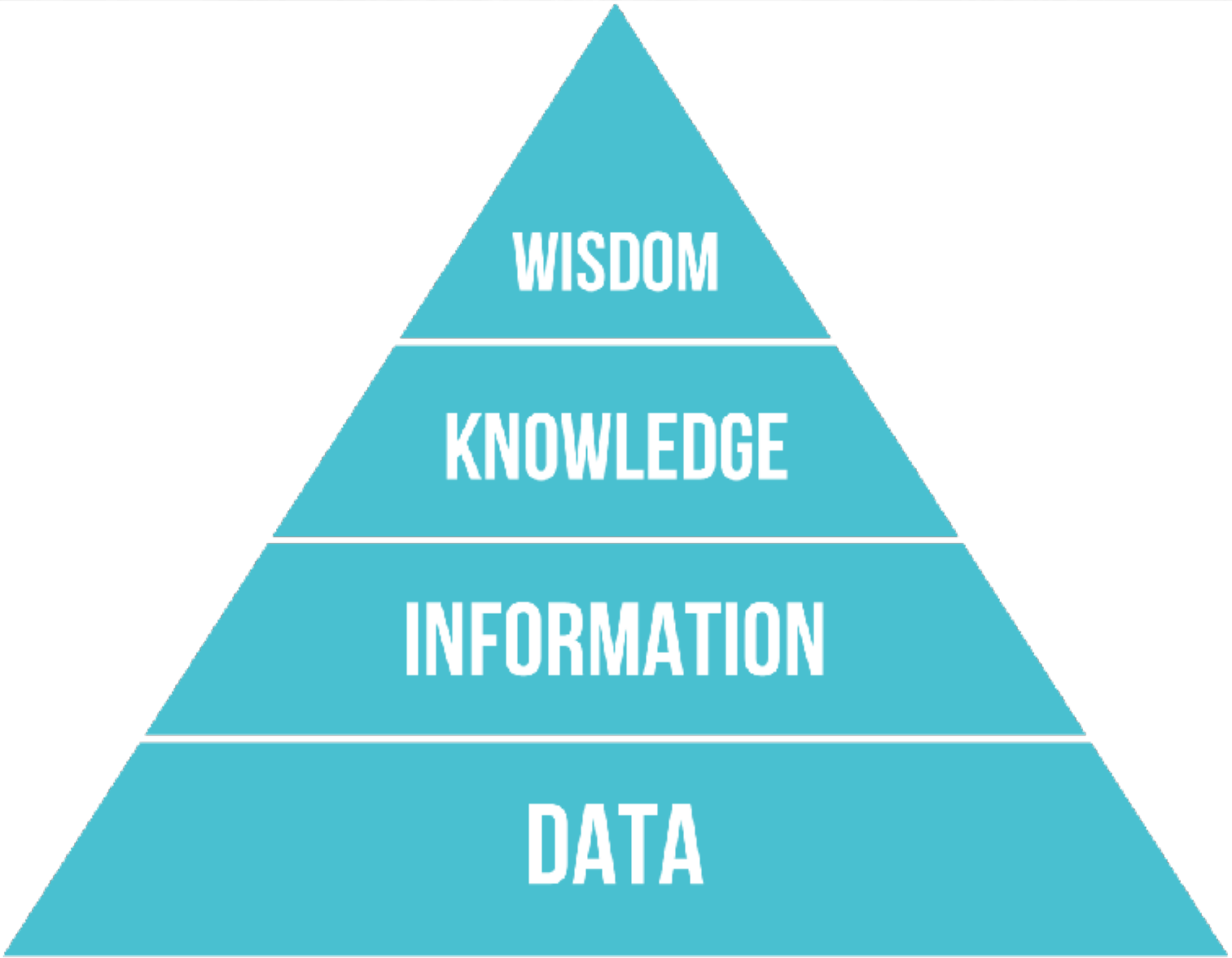
Managing Research Data

Ivaylo Kostadinov, Ph.D.
GFBio e.V.
ivo@gfbio.org
@tigroumaniac

AWI Training on Research Data Management | 2021-10-27

DATA

Data fuels research



https://upload.wikimedia.org/wikipedia/commons/0/06/DIKW_Pyramid.svg

Research Funding in DE

90 Billion €

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

DFG Funding Atlas

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

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

Time effort

.. for discovering and reusing multiple data sources

80%

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

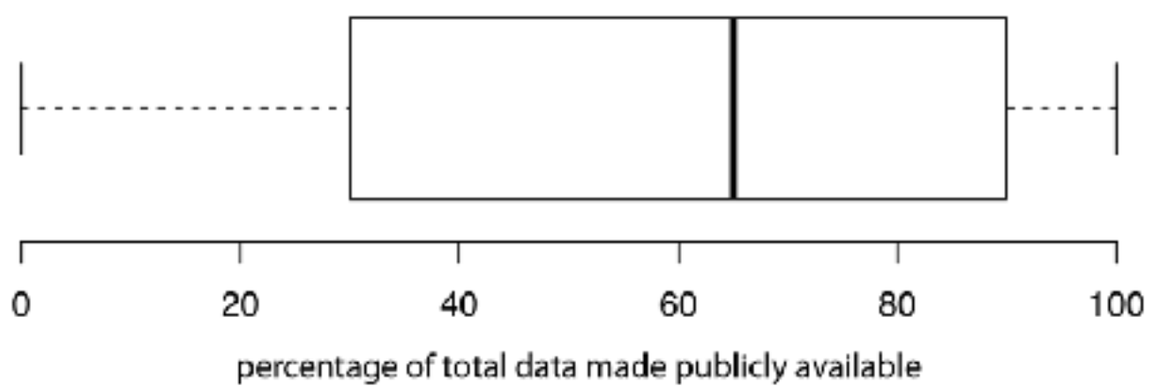
dark data

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

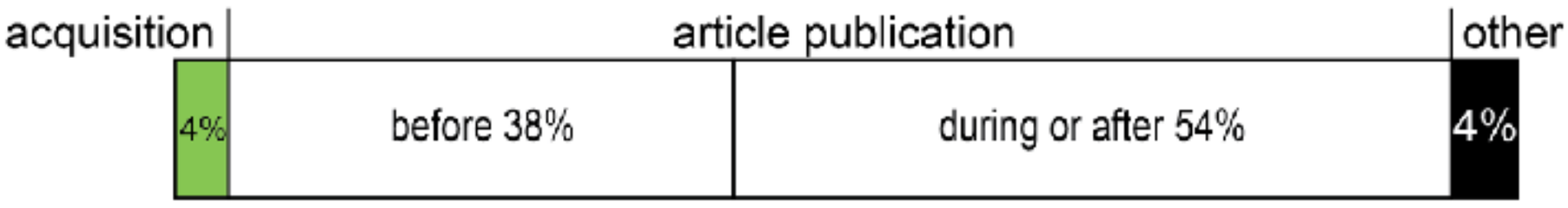
B. P. Heidorn Libr. Trends 57, 280–299; 2008

Challenges

- Not all the data is archived/published



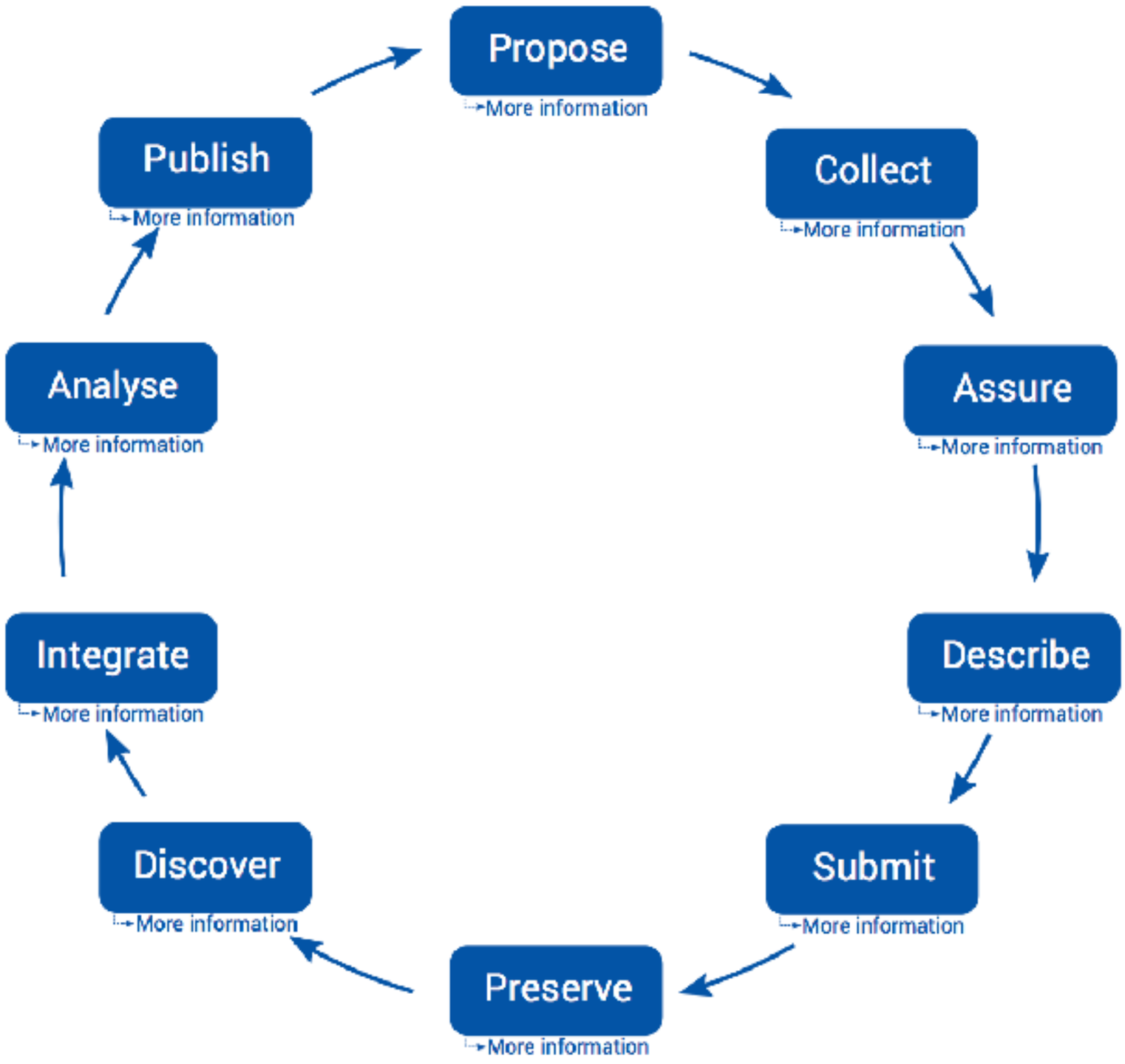
- Data archiving/publication is tied to journal articles



- Contextual data is rarely standardized

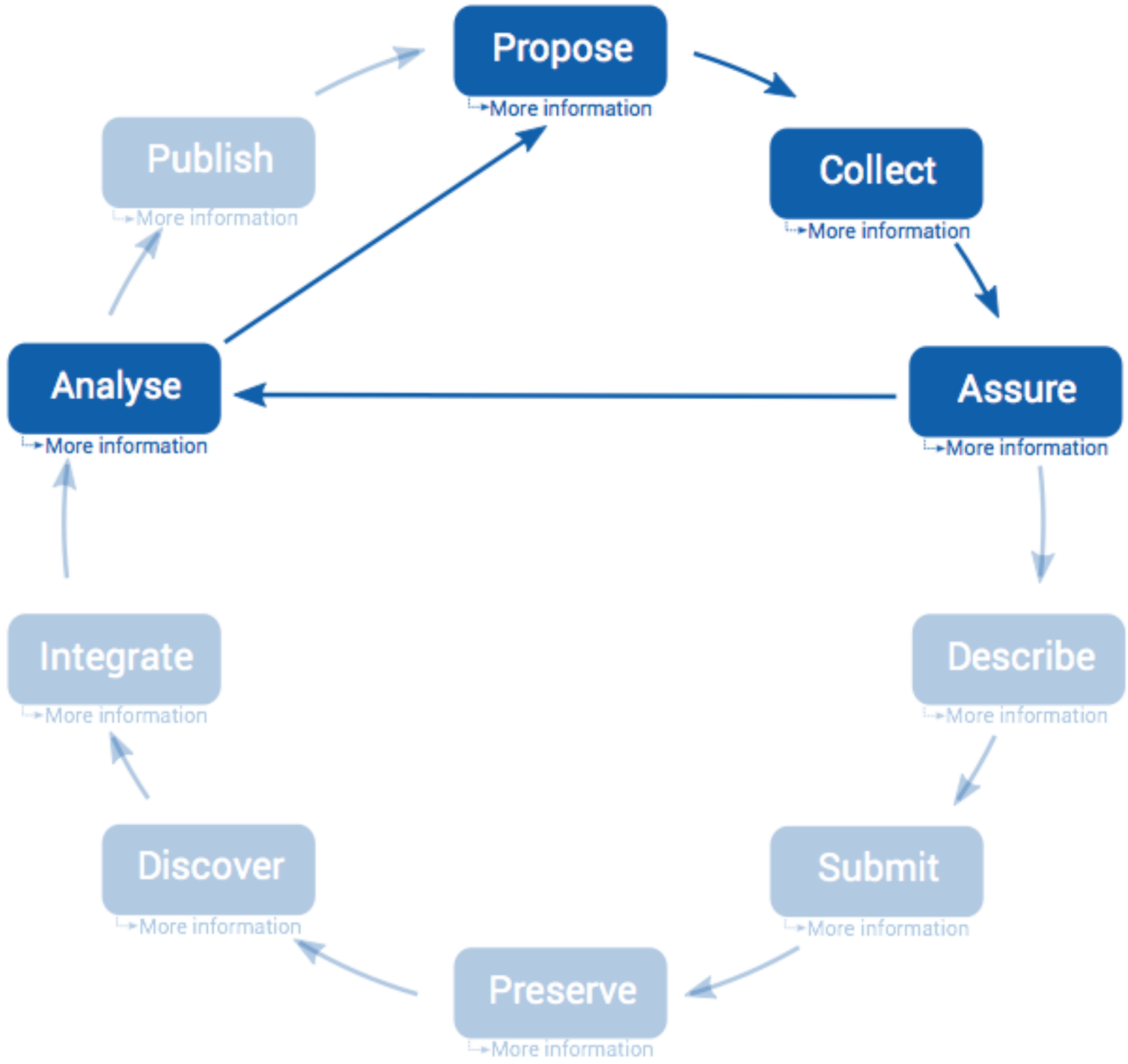
*based on user survey

Data Lifecycle



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

Data Lifecycle



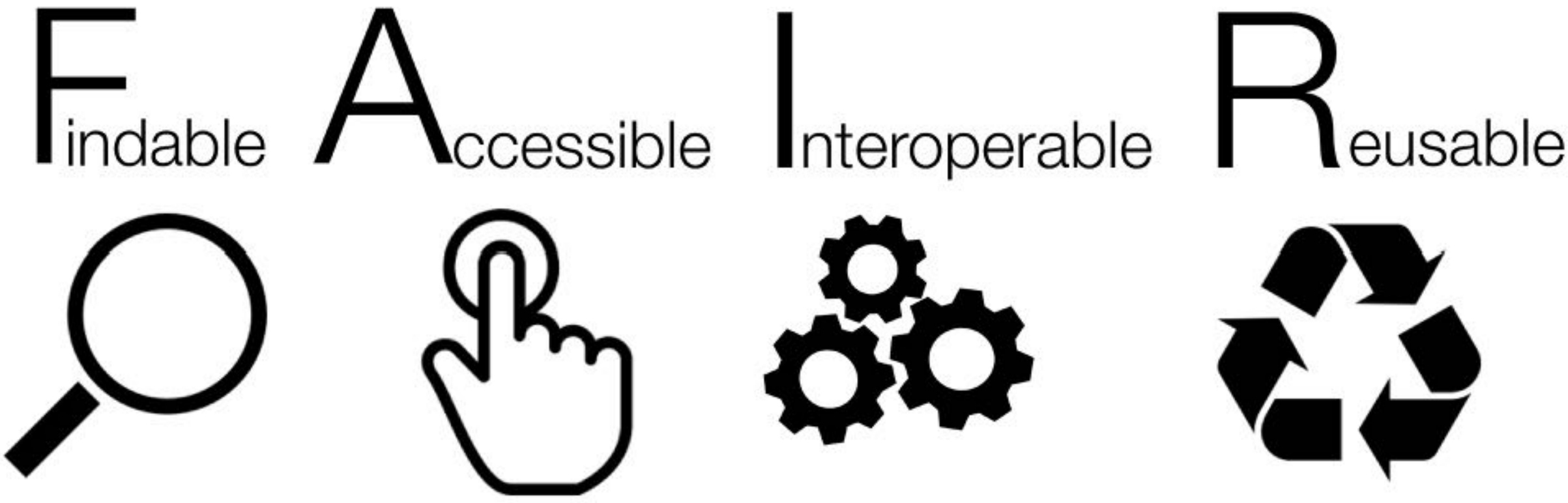
http://www.gfbio.org/training/materials/data-lifecycle

Data challenges

- Data is constantly growing in size and complexity
- Datasets are hard to find and even harder to compare
- No suitable credit for (good) data publication

A lot of primary data gets effectively lost!

Community response



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

www.nature.com/scientificdata

SCIENTIFIC DATA

OPEN

SUBJECT CATEGORIES
» Research data
» Publication characteristics

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

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

Mark D. Wilkinson *et al.*#

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

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

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

B. Mons et al., doi:10.3233/ISU-170824

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)

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

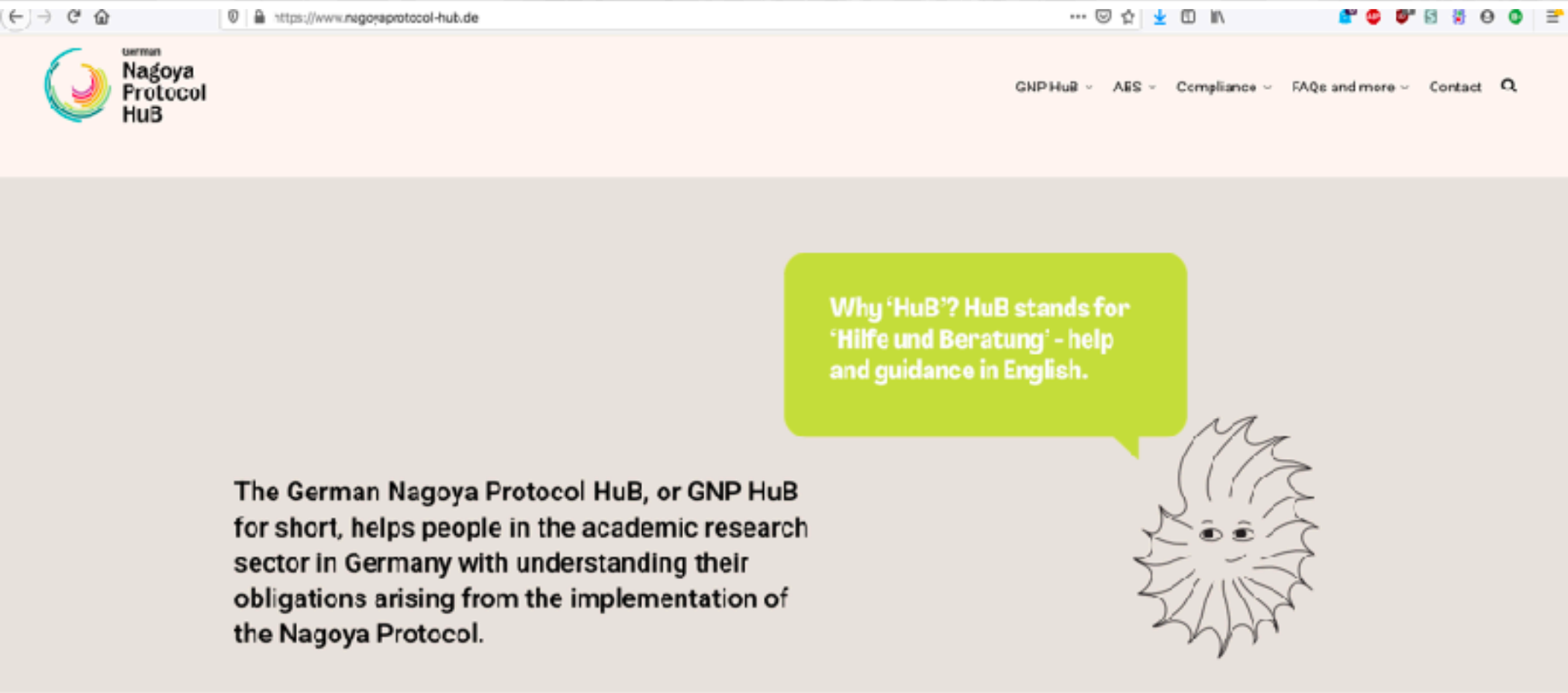
Convention on Biological Diversity

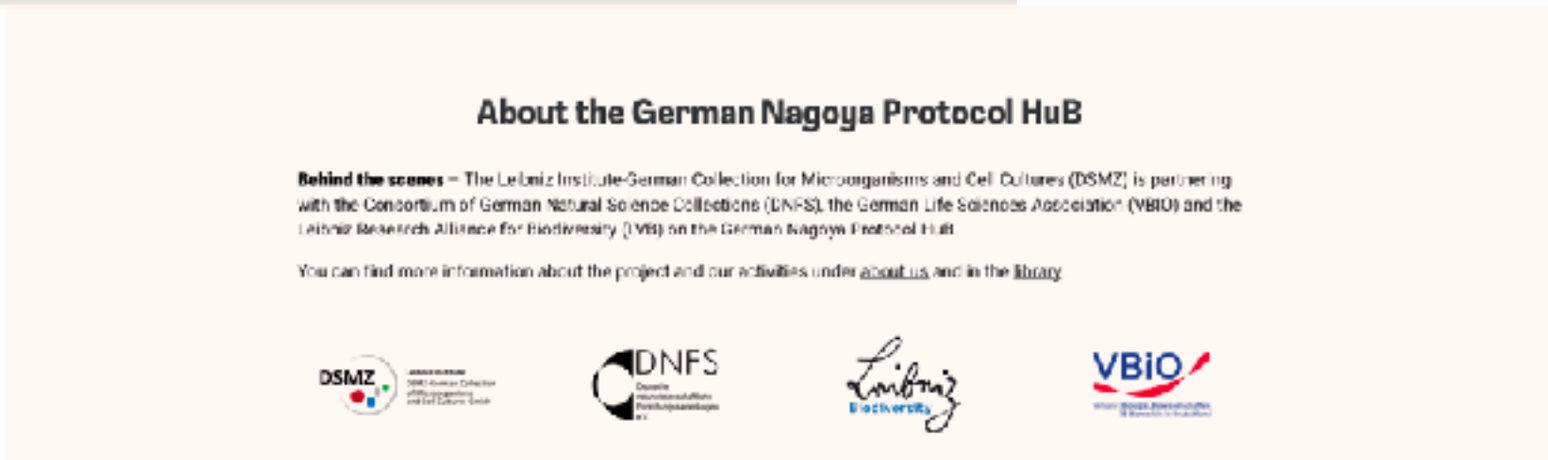
The Nagoya Protocol on Access and Benefit-sharing

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/

Nagoya Hub

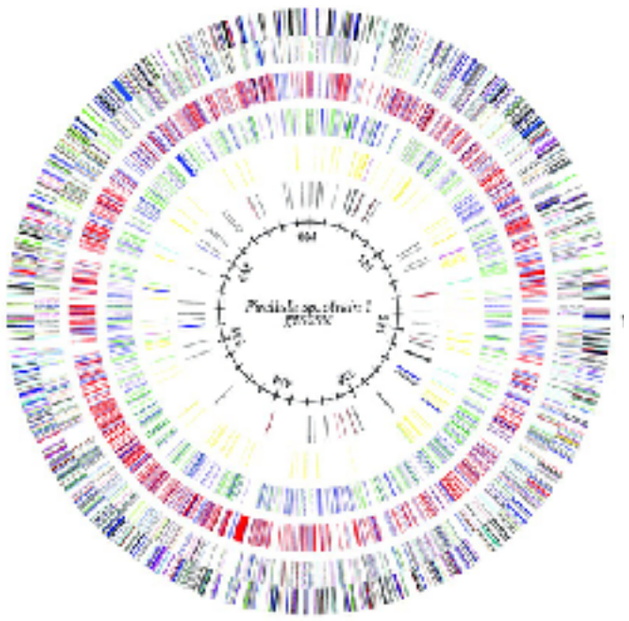




https://www.nagoyaprotocol-hub.de/

METADATA & STANDARDS

Metadata



Glöckner et al., 2003

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

Courtesy: Boyke Bunk, DSMZ

DATA

METADATA

contextual data

Sequence Metadata

MIxS **M**inimal **I**nformation about any(**x**) **S**equence

http://gensc.org/mixs/

developed by:

genomic STANDARDS consortium

supported by:

INSDC International Nucleotide Sequence Database Collaboration

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)

Sequence Metadata

MIxS **M**inimal **I**nformation about any(**x**) **S**equence

	MIGS					MIMS	MIMARKS	
Item	EU	BA	PL	VI	ORG	ME	SU	SP
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
Ploidy	M	-	-	-	-	-	-	-
Estimated size	M	X	X	X	X	-	-	-
Target gene	-	-	-	-	-	-	M	M

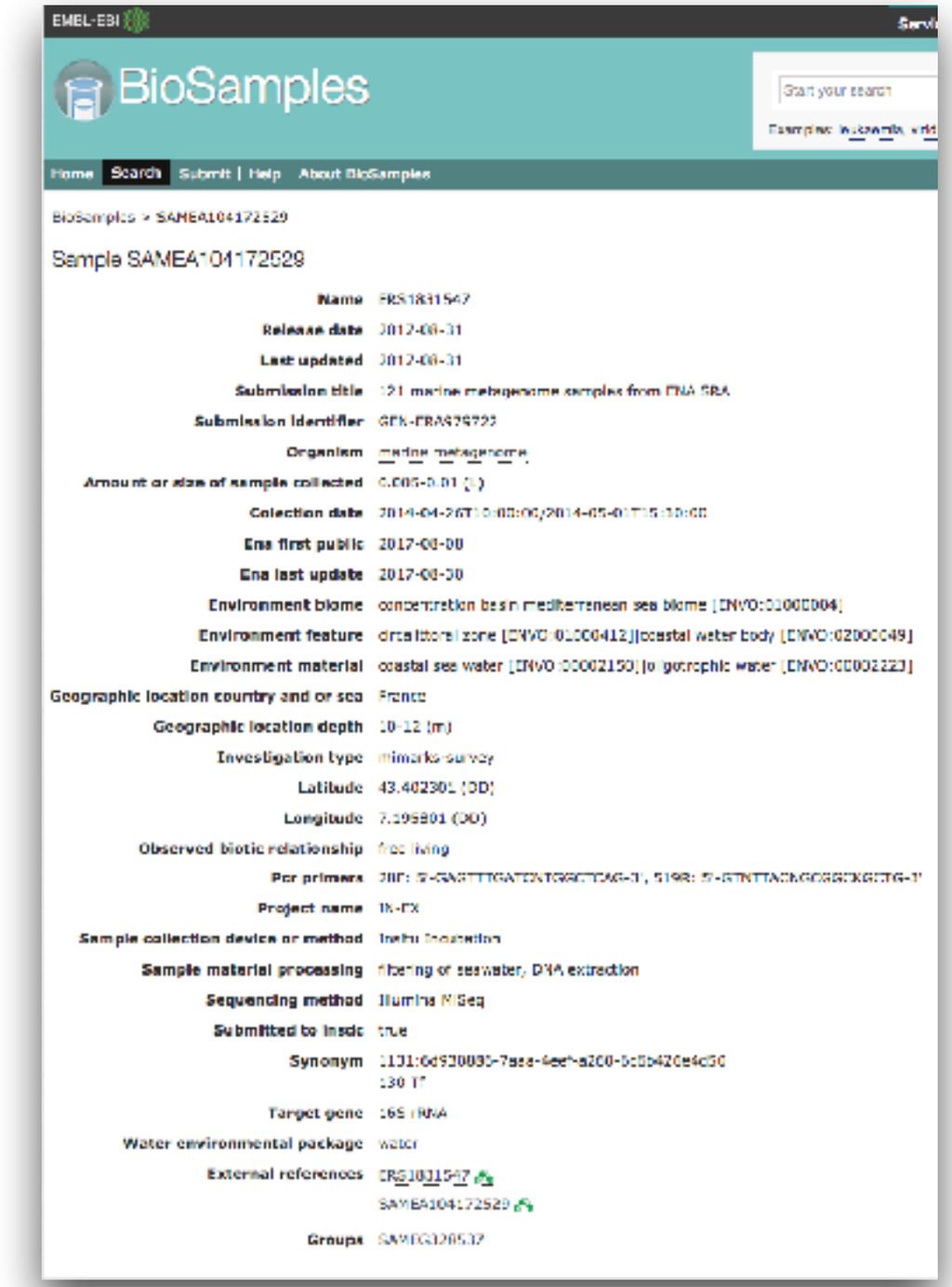
Environmental Packages

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

MIMARKS-Survey-Soil Checklist

Item	SU
Project name	M
Collection date	M
Geographic location	M
Environment	M
Target gene	M
Elevation	M
Depth	M
Horizon	X
pH	X

Yilmaz et al., 2011, Nature Biotechnology, 29, 5, 415-420



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)

What else is out there?

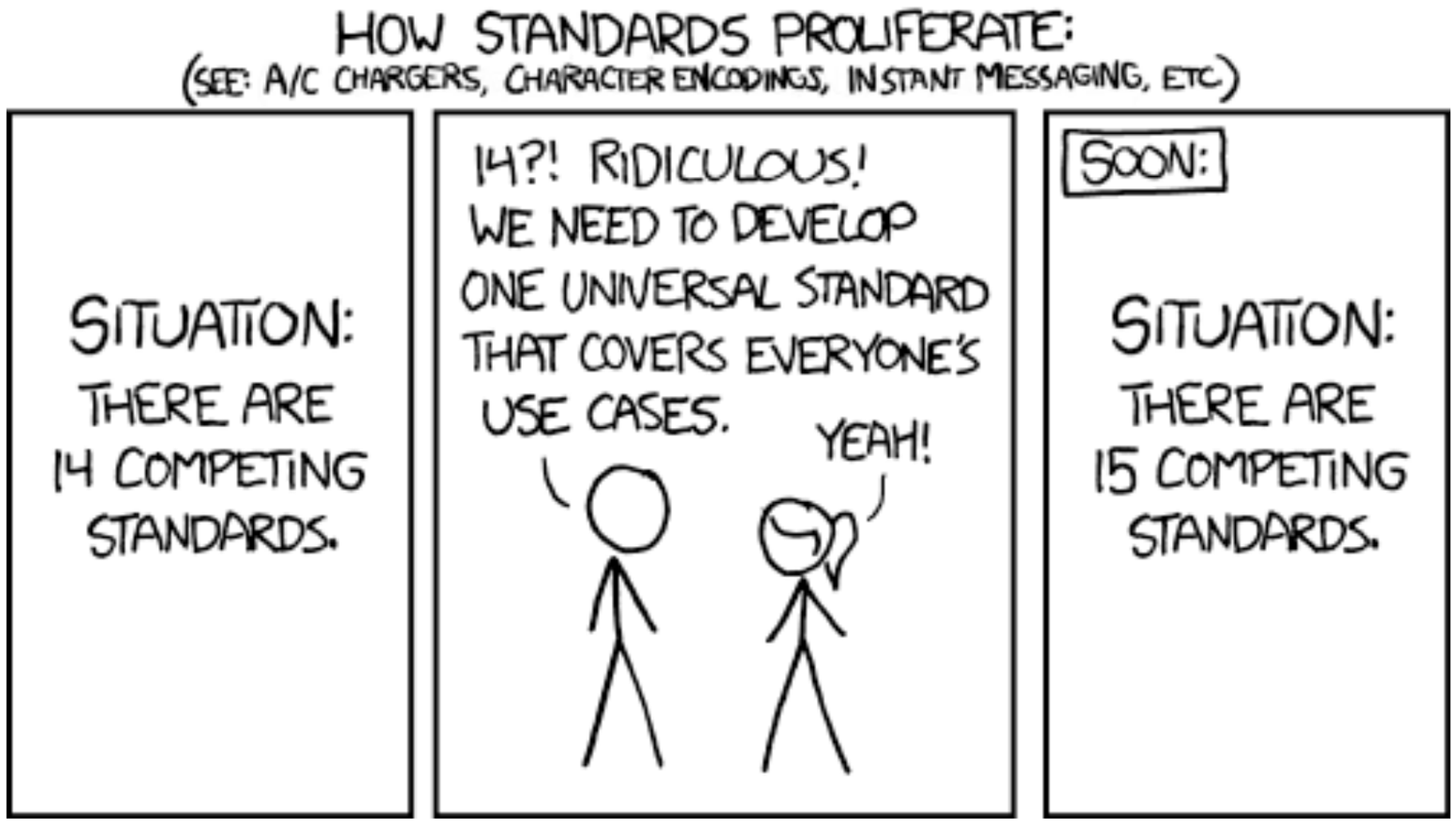
Standards for biodiversity & ecology:

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

Standardization challenges

- Insufficient (acquisition) tooling
- Insufficient training & support

Standardization challenges



http://xkcd.com/927/

ARCHIVAL & PUBLICATION

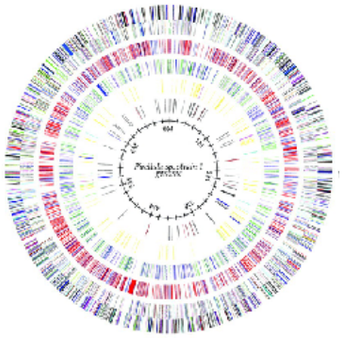
Archival & Publication

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

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

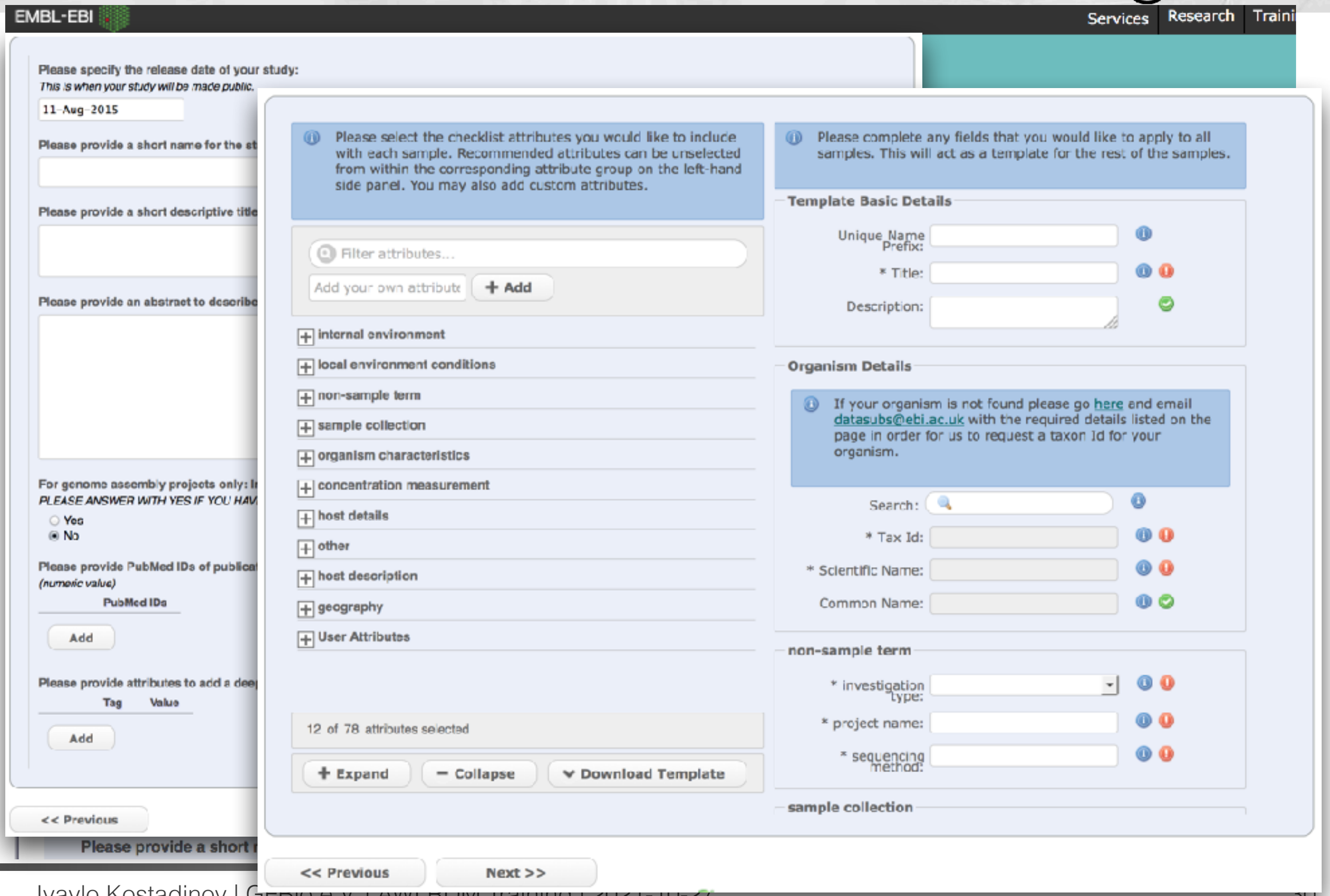
Programmatic submissions

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



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

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

Easy for the user,

not so much for the curator.

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

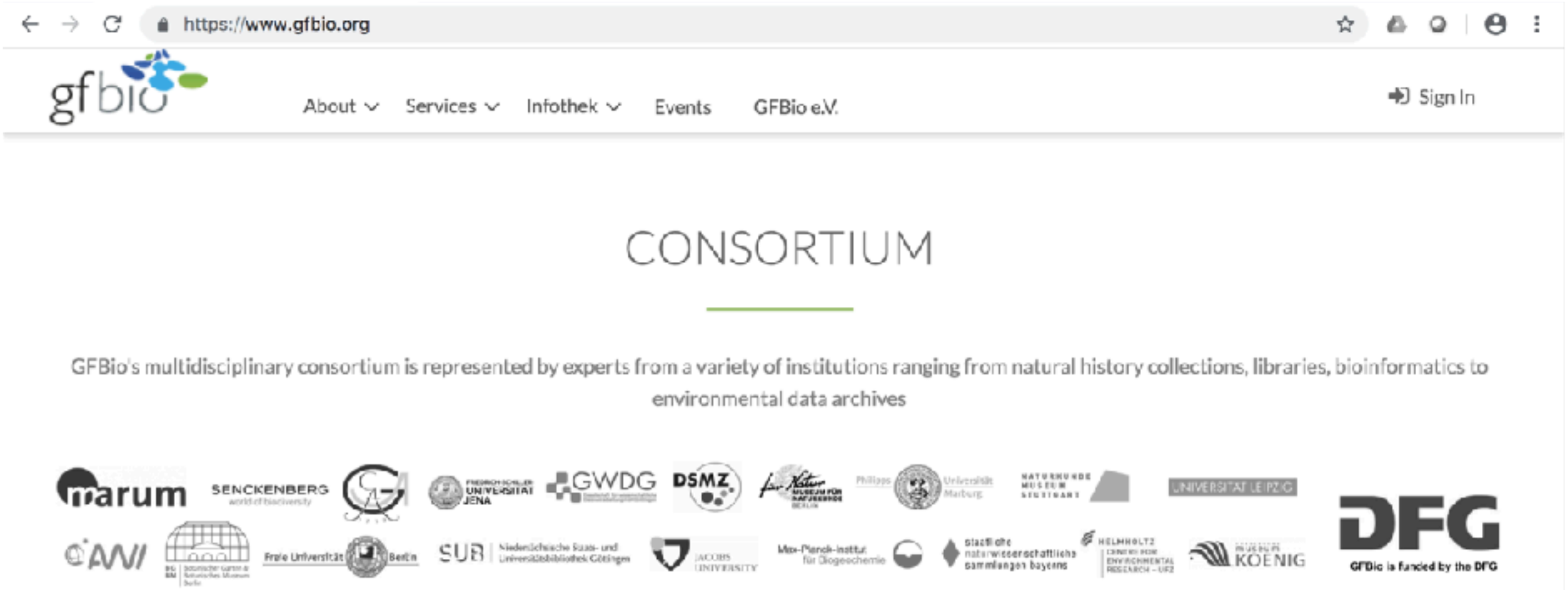
www.gfbio.org

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

Funded by:
DFG Deutsche Forschungsgemeinschaft

www.gfbio.org

Network of experts



data centers and publishers

infrastructure partners

user-representatives

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.

Data Management Plan

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

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

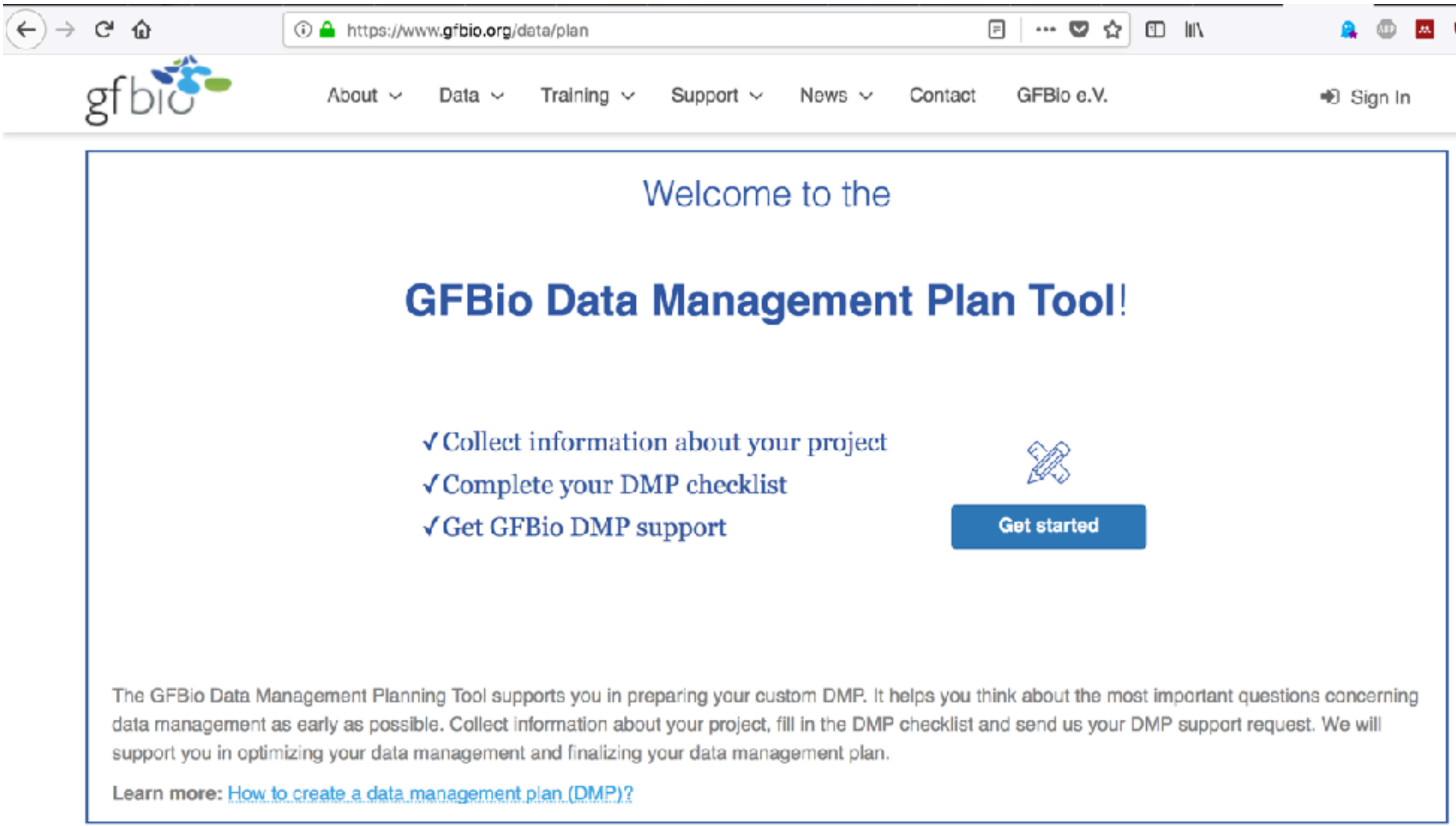
https://www.gfbio.org/training/materials/data-lifecycle/plan

Data Management Plan

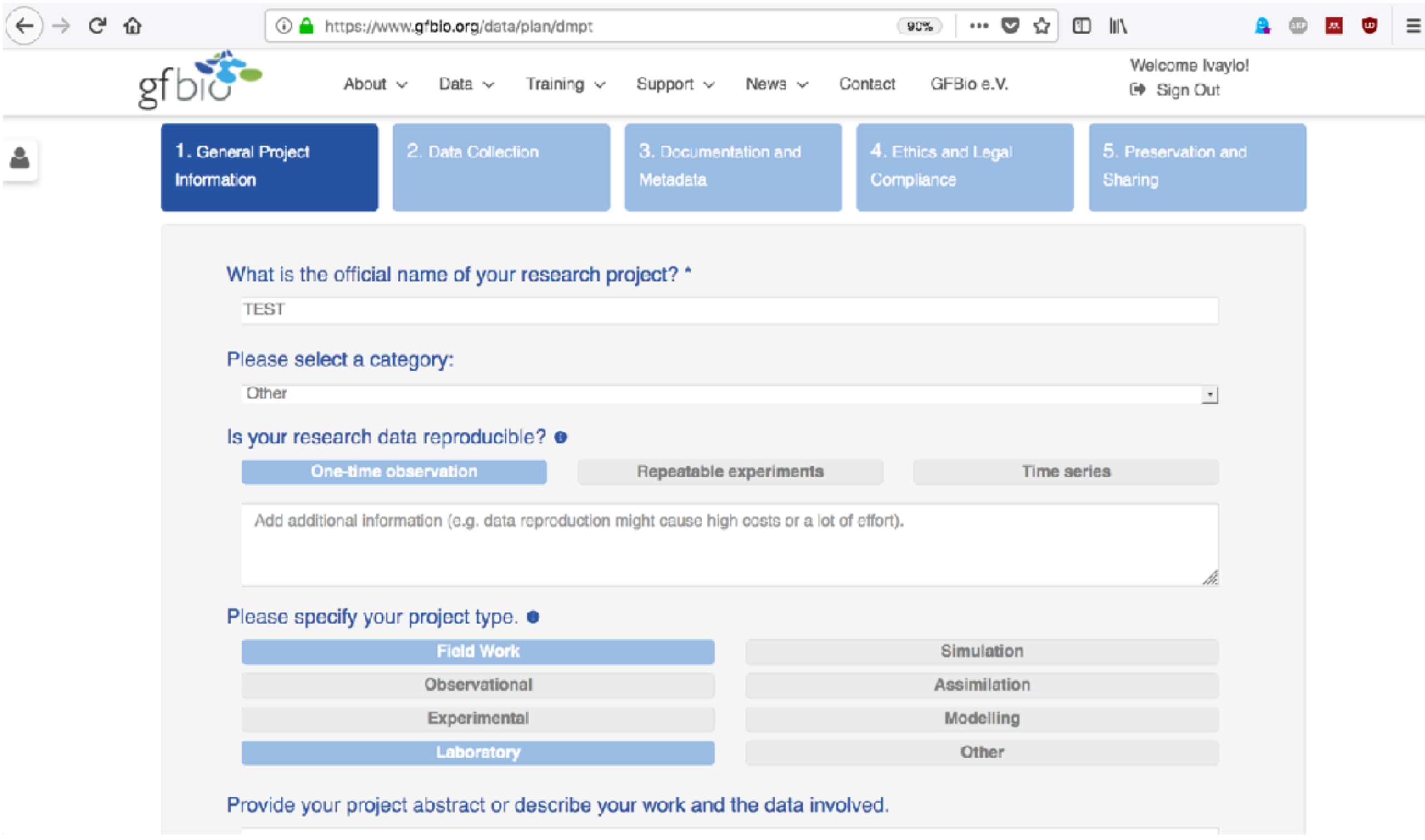
- **Preservation:** selection of data, foreseeable future use, time and location for preservation, costs
- **Data sharing and publication:** modalities, conditions, persistent identifiers
- **Responsibilities:** implementation, roles and responsibilities for each activity, ownership agreement
- **Resources:** 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

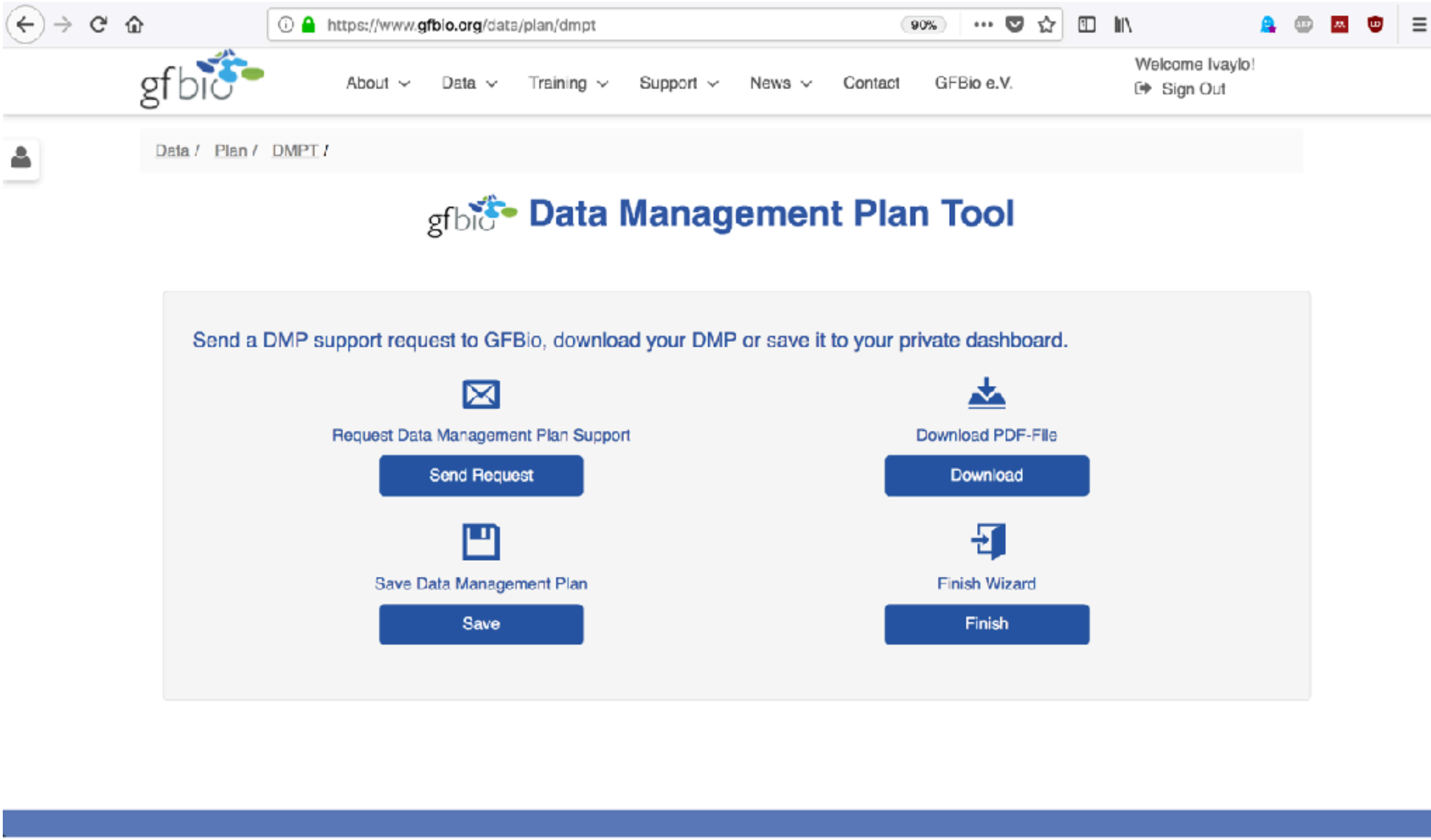
DMP Support


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)?

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.

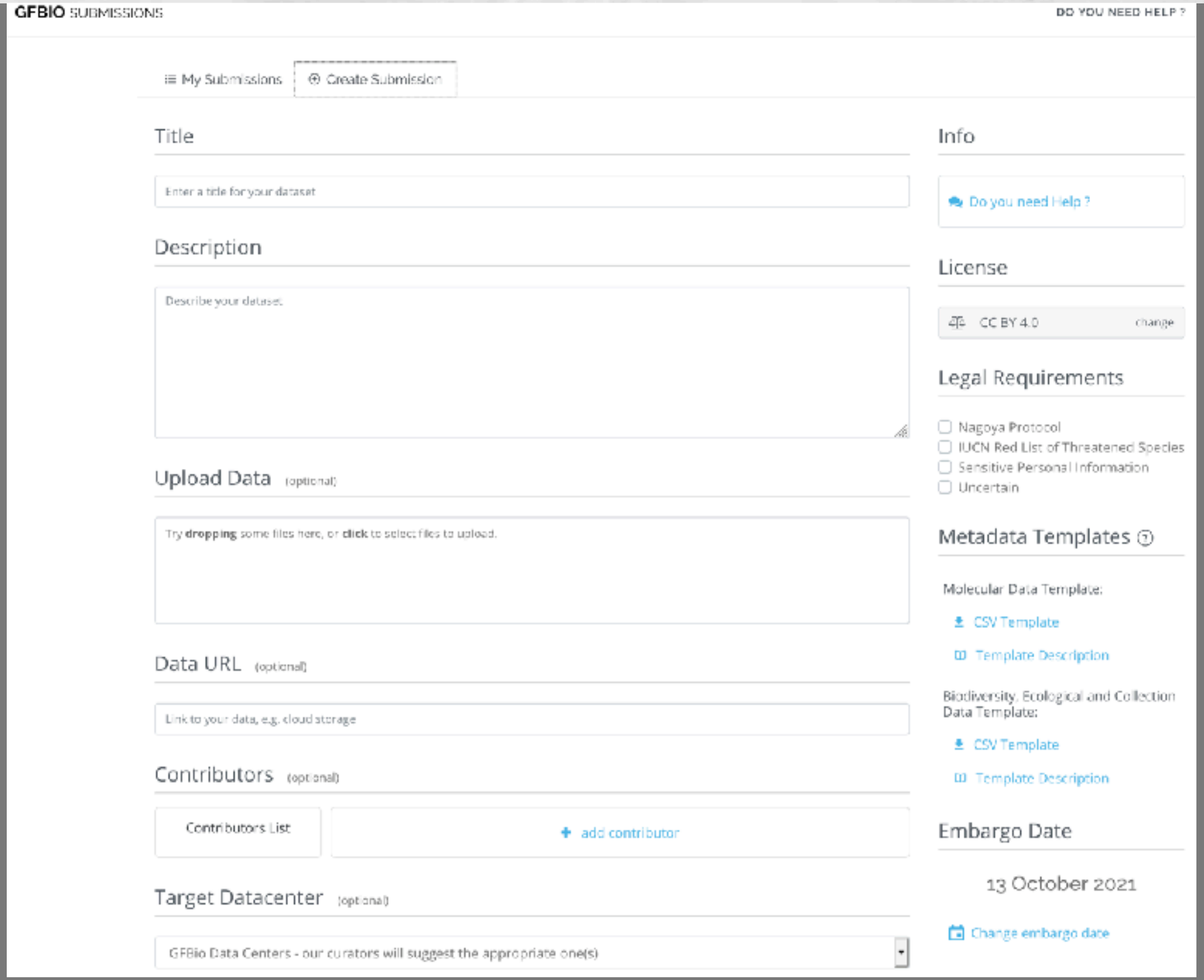
DMP Support


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

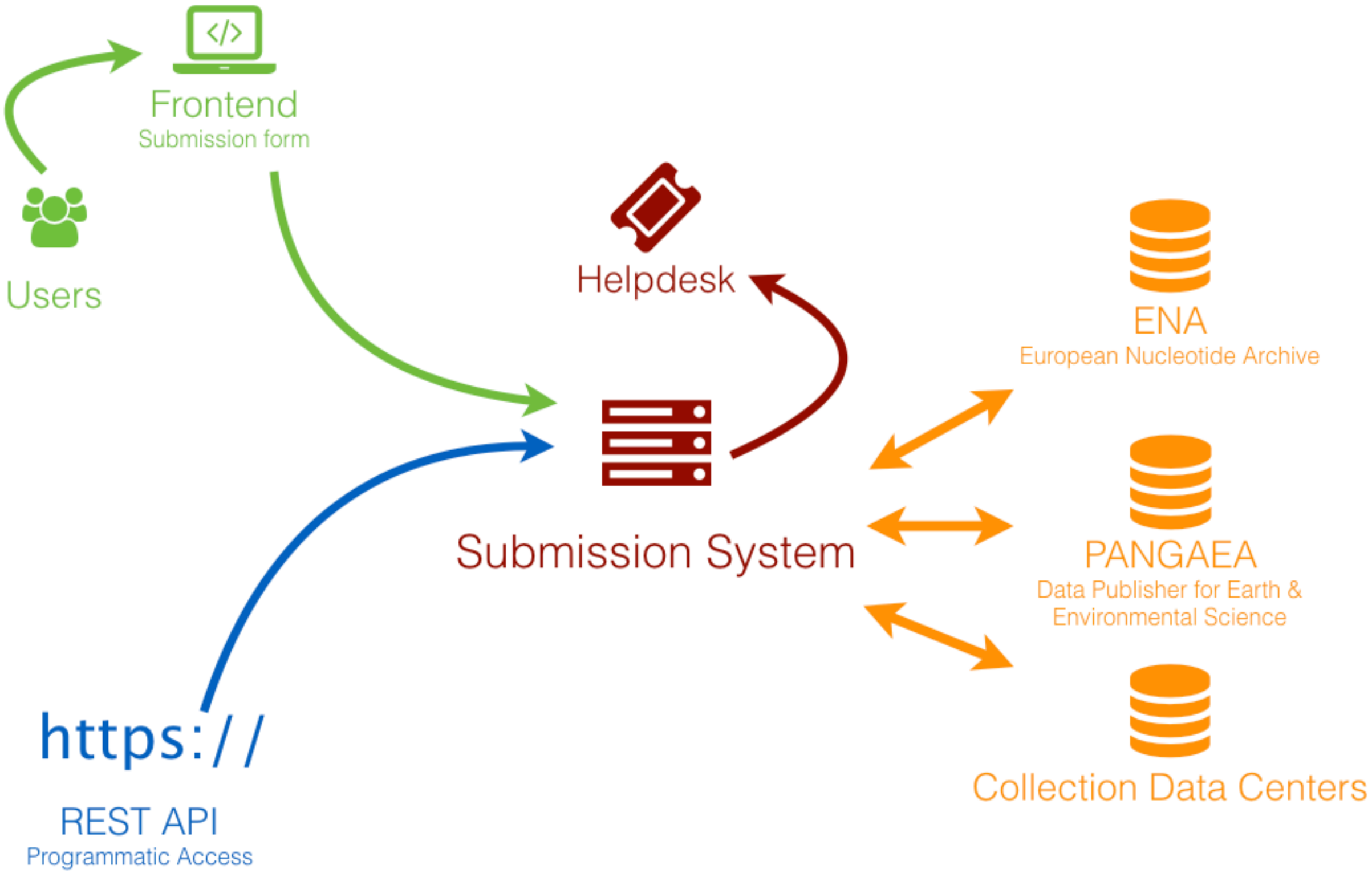
Data Submission

https://submissions.gfbio.org

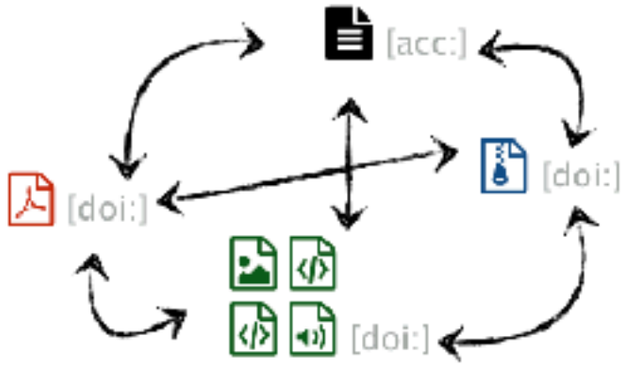
Data Submission



Data Submission



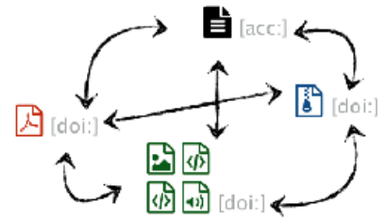
Data Submission



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)

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

Recent Activity | Export

Formats

Abstract | Full Text | PDF

Show annotations in this abstract

☐ Chemicals
☐ Organisms

Data Citations

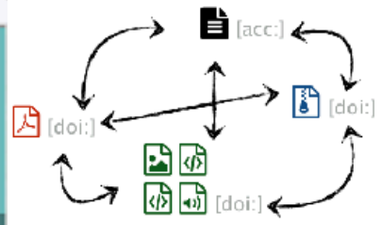
Identified 2 unique Data Citations in the full text

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.

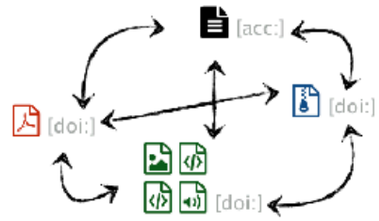
Navigation | Read Files | Portal | Attributes | Publications

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 92835 (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):

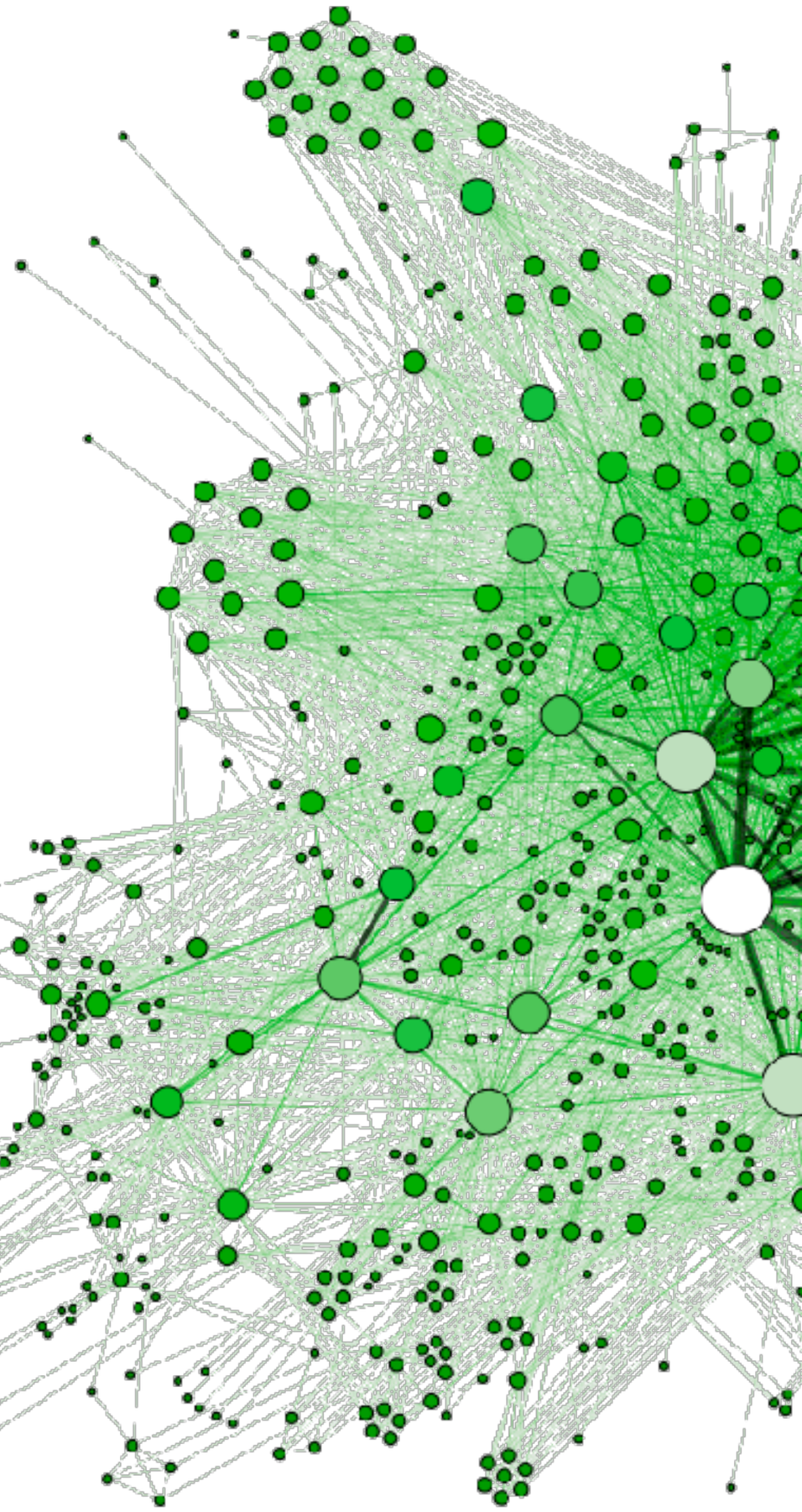
Leibniz Centre for Tropical Marine Research (ZMT)

Outlook

- Major upgrade of DMP support tooling (versioning, collaboration, etc.)
- Linking molecular data sets to ORCID profile
- Establishing brokerage for functional genomics data
- Further data repositories for data deposition

NFDI 4
BIODIVERSITY

NFDI4Biodiveristy Service Portfolio



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

Thank you!

www.nfdi4biodiversity.org

www.gfbio.org

contact@nfdi4biodiveristy.org (collaboration)

info@gfbio.org (service help desk)

@NFDI4Biodiv

NFDI 4 BIODIVERSITY

Questions?