Research Data Management

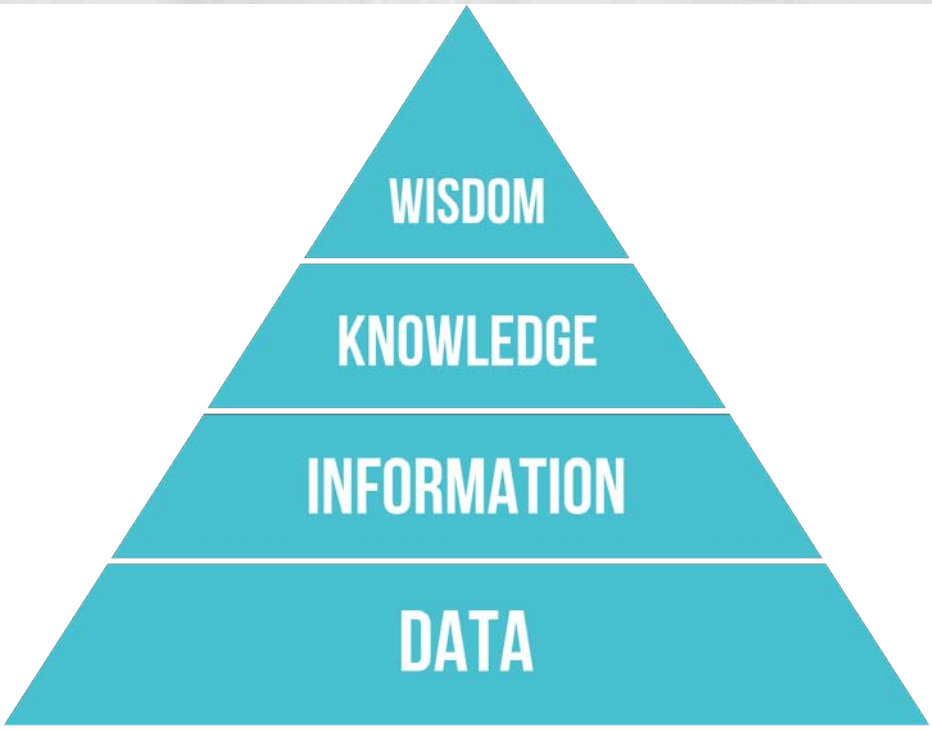
GFBio services for archiving and publishing sequence data

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

gfbio e.V.
@GFBio_Project

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

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

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

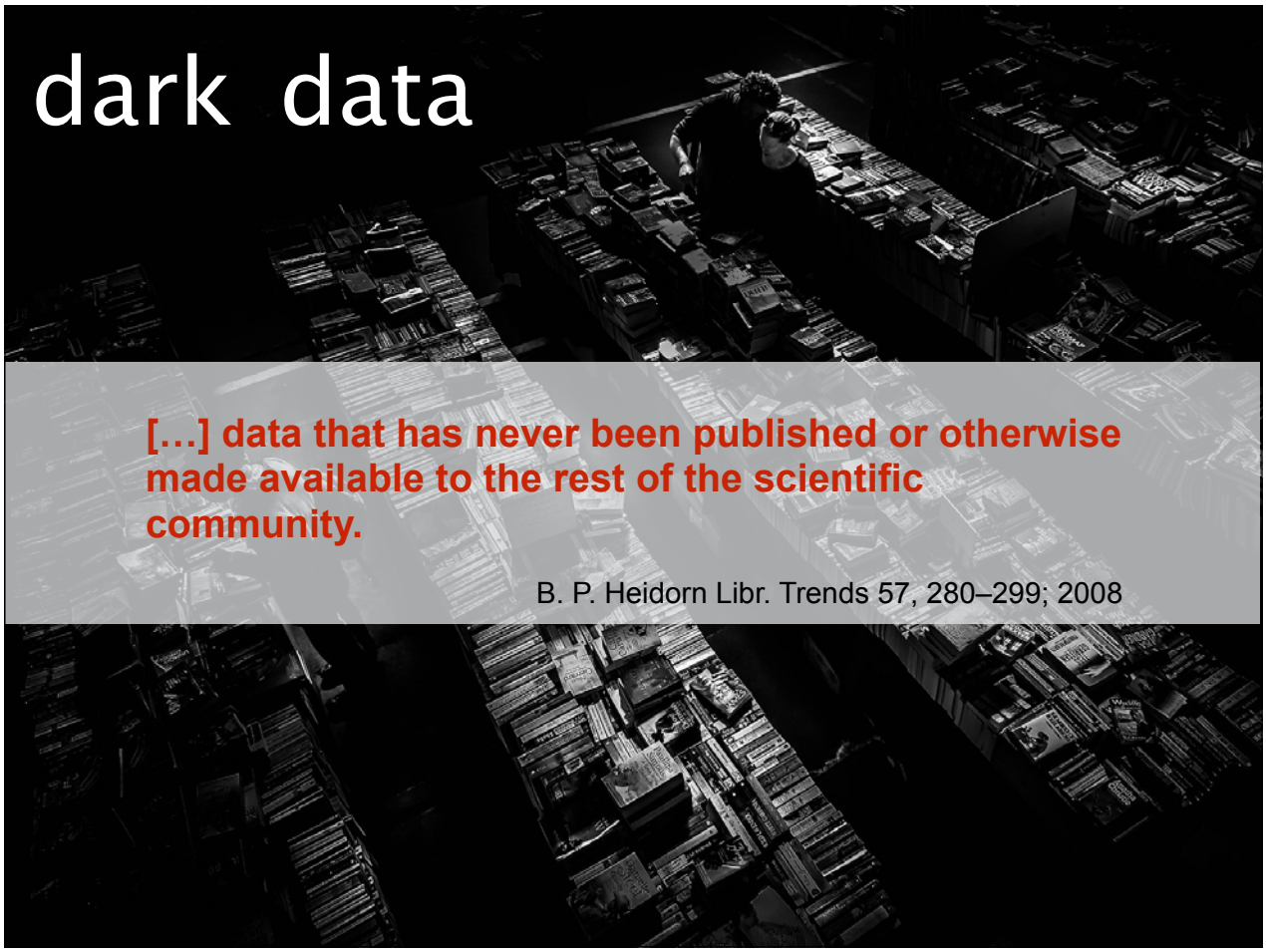
Time effort

.. for discovering and reusing multiple data sources

80%

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

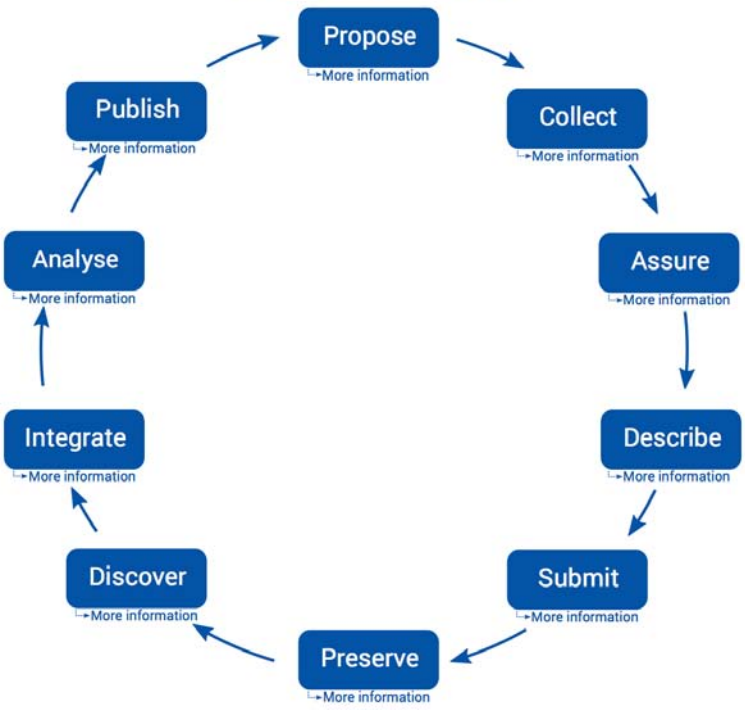
dark data



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

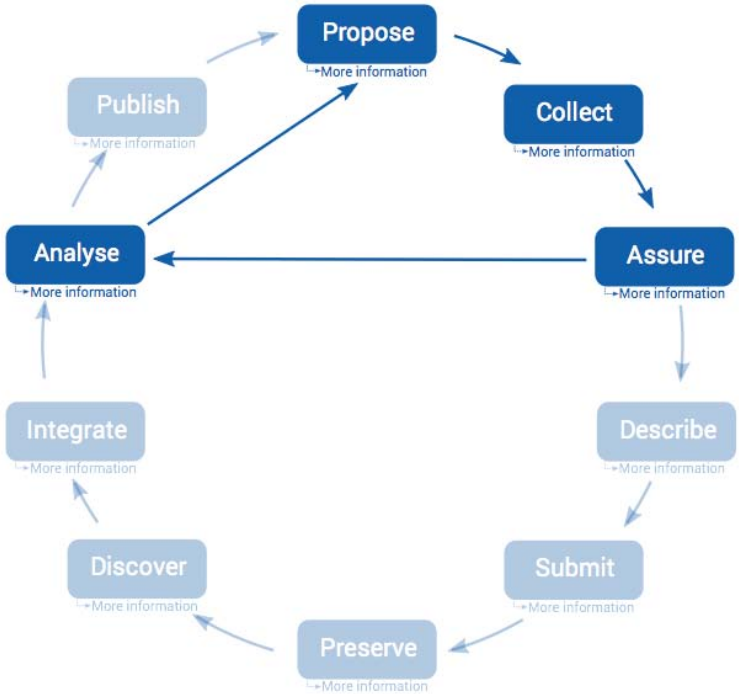
B. P. Heidorn Libr. Trends 57, 280–299; 2008

Data Lifecycle



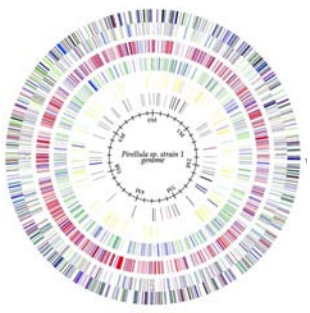
http://www.gfbio.org/training/data-lifecycle

Data Lifecycle

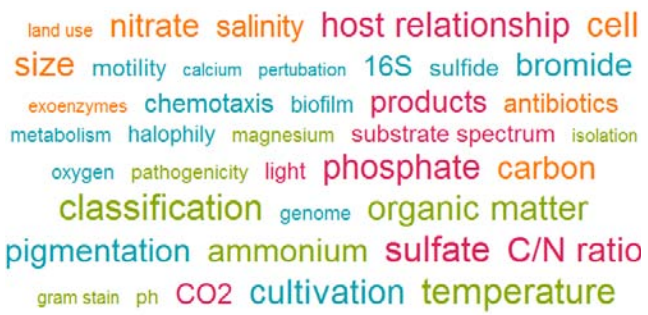


http://www.gfbio.org/training/data-lifecycle

Metadata



Glöckner et al., 2003



Courtesy: Boyke Bunk, DSMZ

DATA

METADATA

contextual data

VIDEO

Data Sharing and Management Snafu in 3 Short Acts by NYU Health Sciences Library

https://youtu.be/N2zK3sAtr-4

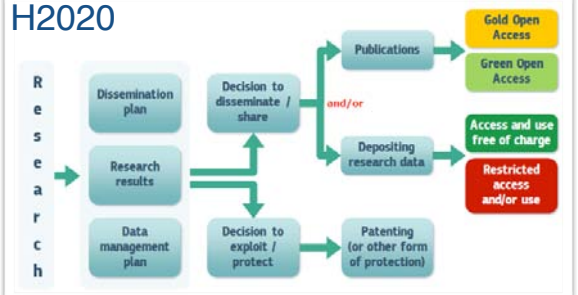
Community response

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

2007

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

H2020



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

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

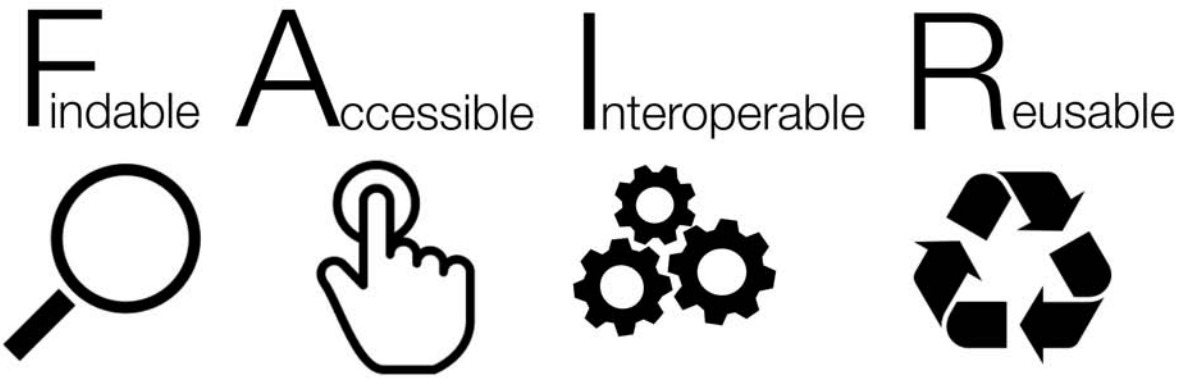
http://www.dfg.de/en/research_funding/proposal_review_decision/applicants/submitting_proposal/research_data/

Rat für Informations Infrastrukturen

http://www.rfii.de

LEISTUNG AUS VIELFALT

Empfehlungen zu Strukturen, Prozessen und Finanzierung des Forschungsdatenmanagements in Deutschland



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

FAIR Data

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

Biopiracy

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/

STANDARDS

Sequence Metadata

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

http://gensc.org/mixs/

developed by:

genomic STANDARDS consortium

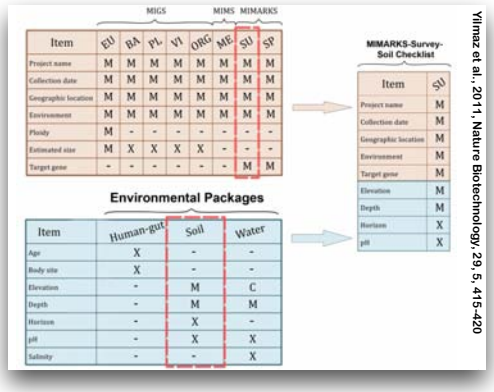
supported by:

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



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)

GFBio

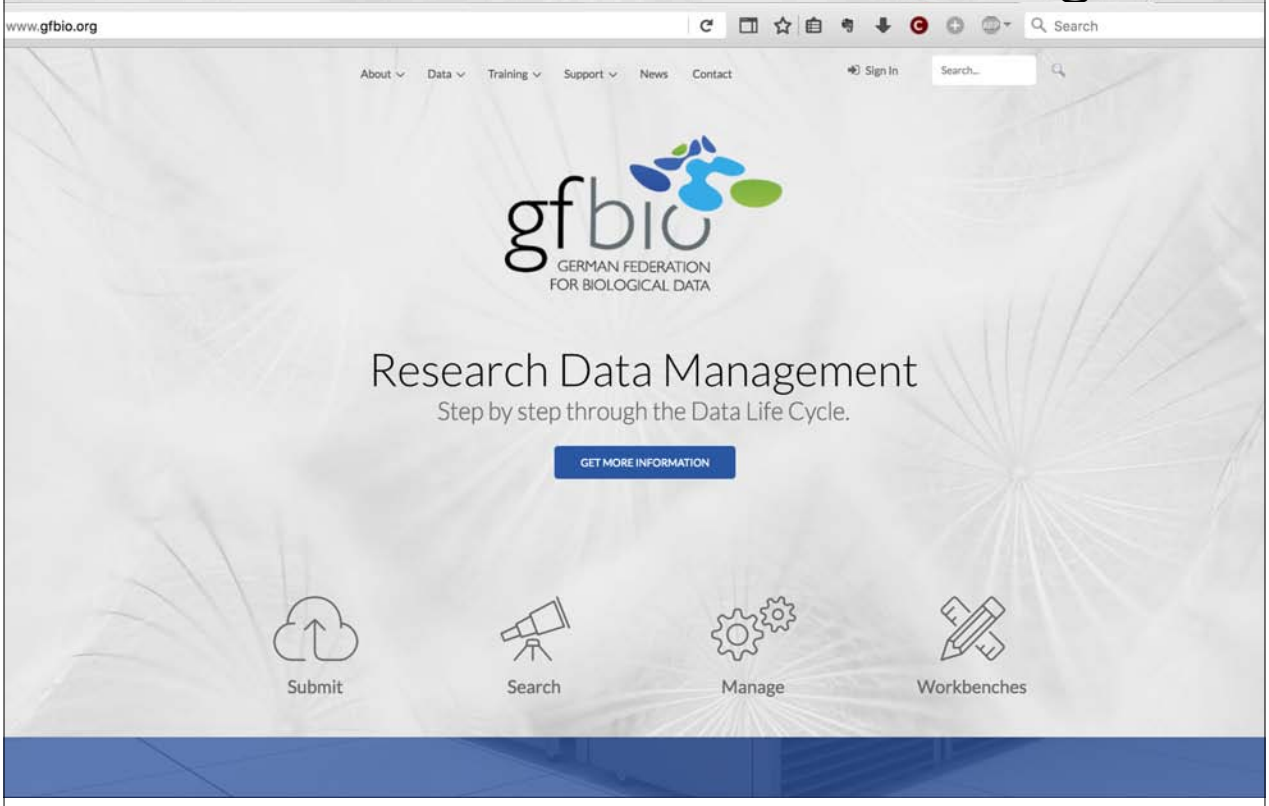
GERMAN FEDERATION FOR BIOLOGICAL DATA

www.gfbio.org

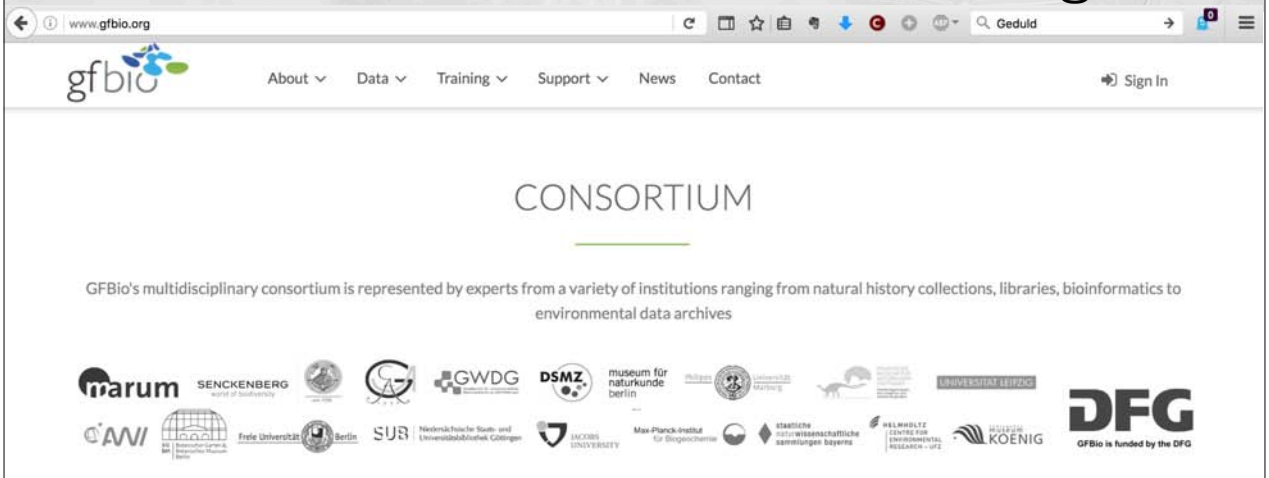
A **sustainable**, **service-oriented** data infrastructure to facilitate **data sharing** and **data-intensive** science in biology and environmental research.

Funded by:
DFG Deutsche Forschungsgemeinschaft

www.gfbio.org



Interdisciplinary team

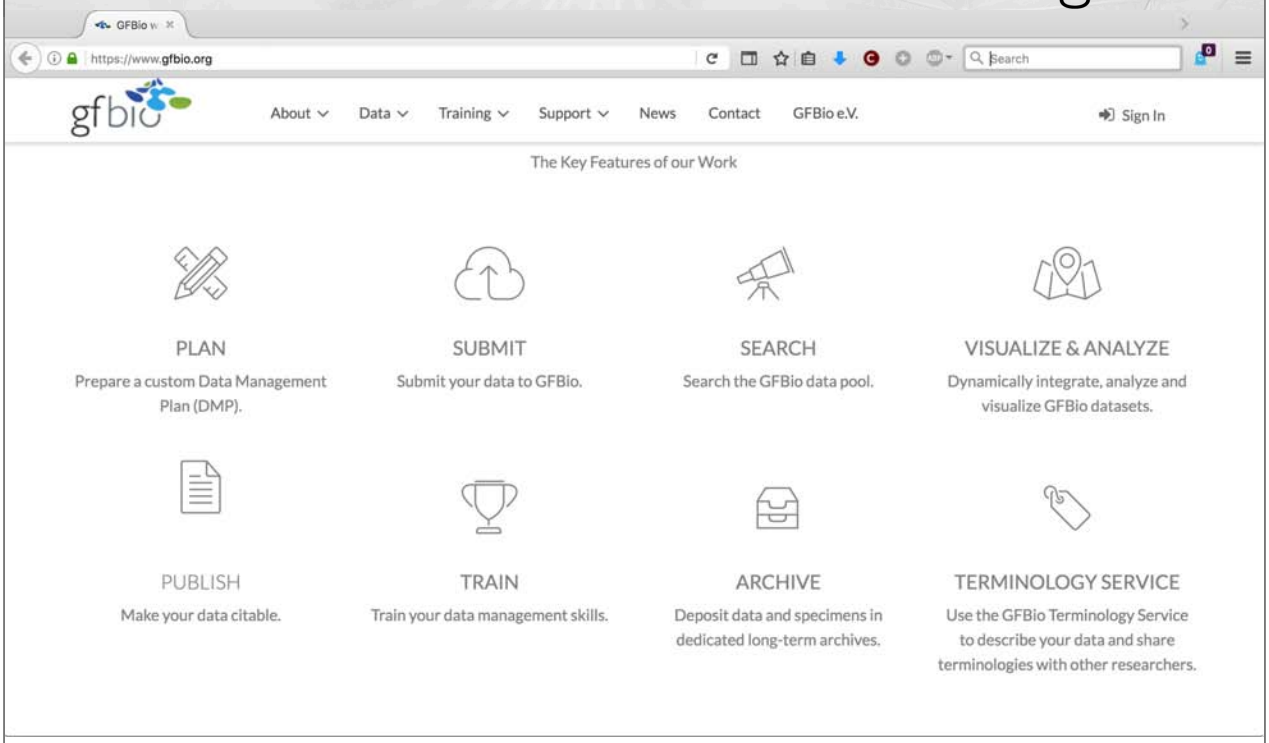


data centers and publishers

infrastructure partners

user-representatives

GFBio Services



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

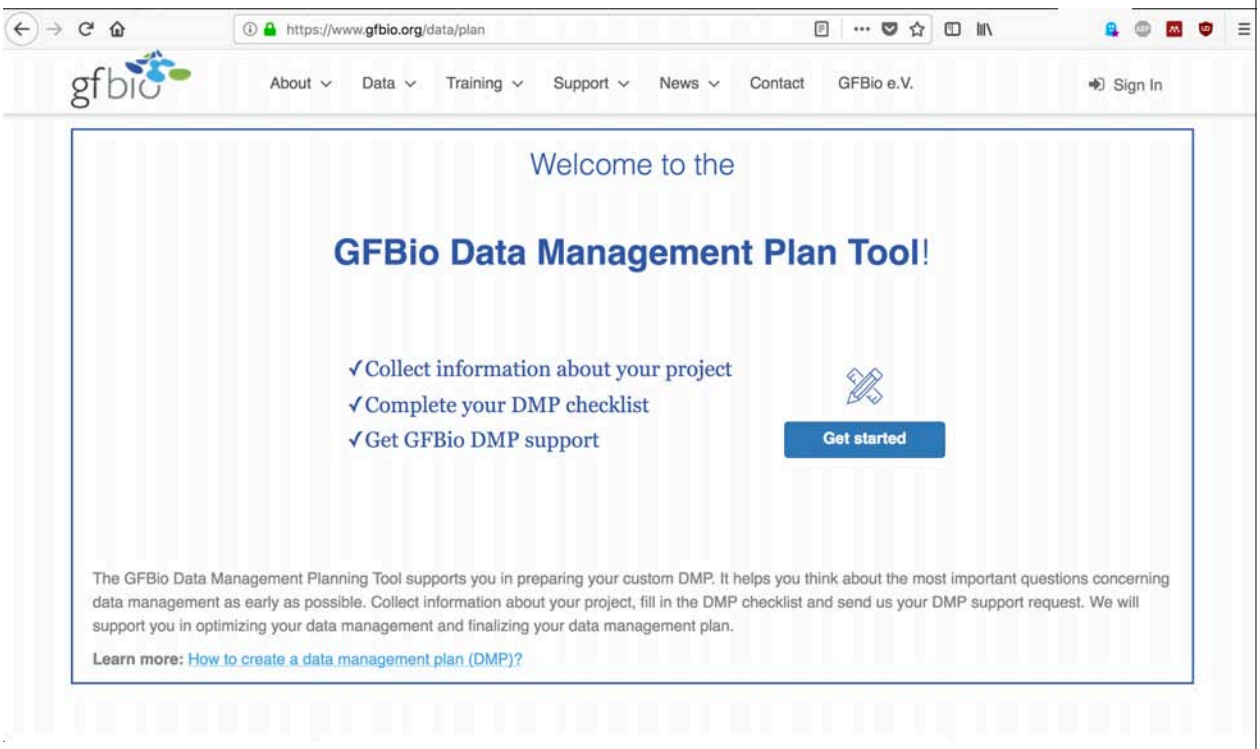
https://www.gfbio.org/training/materials/data-lifecycle/plan

Data Management Plan

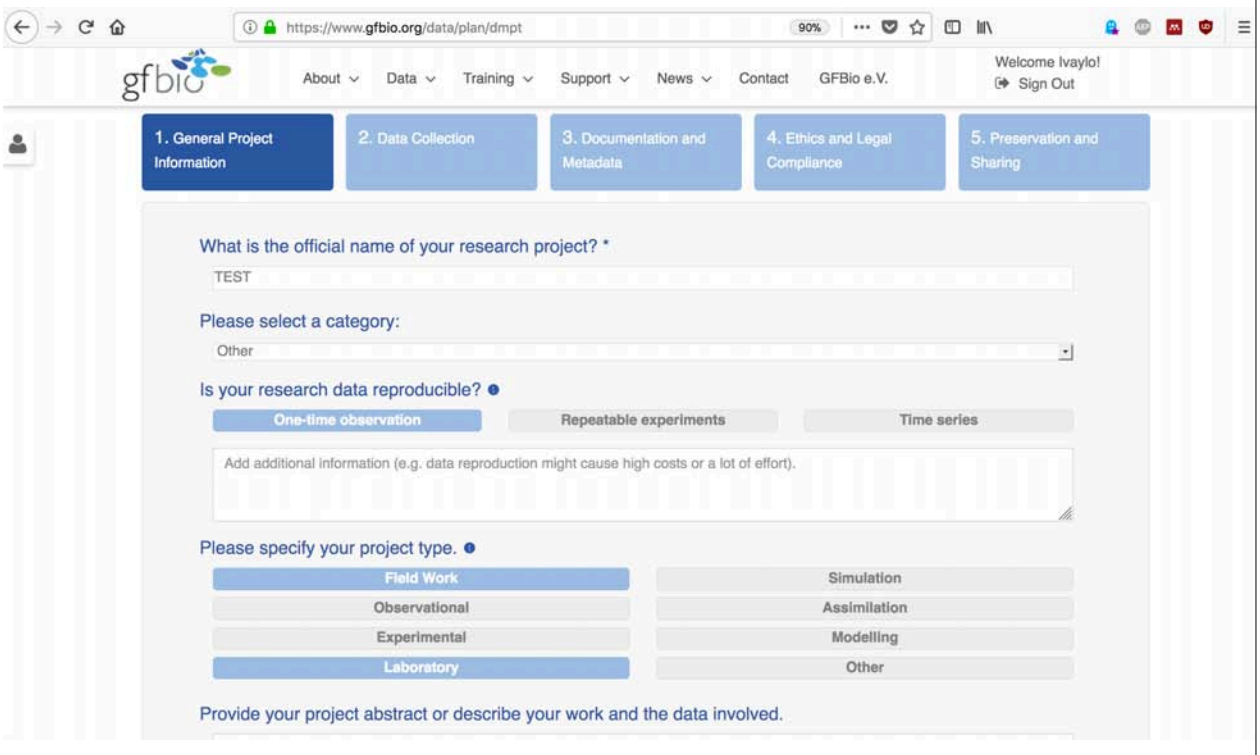
- **Preservation:** selection of data, foreseeable future use, time and location for preservation, costs
- **Data sharing and publication:** modalities, conditions, persistent identifiers
- **Responsibilities:** implementation, roles and responsibilities for each activity, ownership agreement
- **Resources:**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

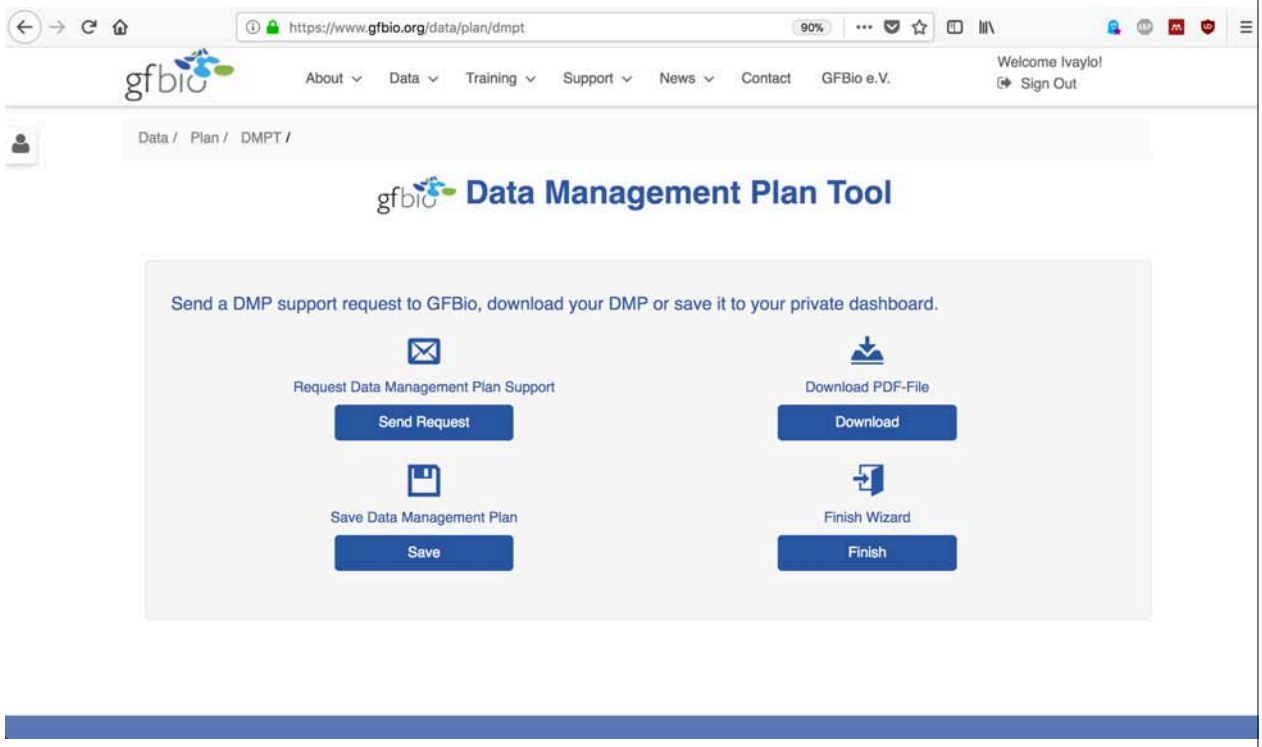
DMP Support



DMP Support

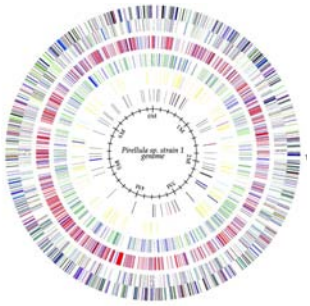


DMP Support



DATA SUBMISSION

Archival & Publication



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

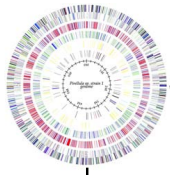
Archival & Publication

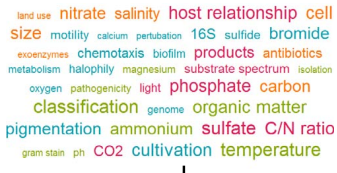
Findable

Accessable

Interoperable

Reproducible





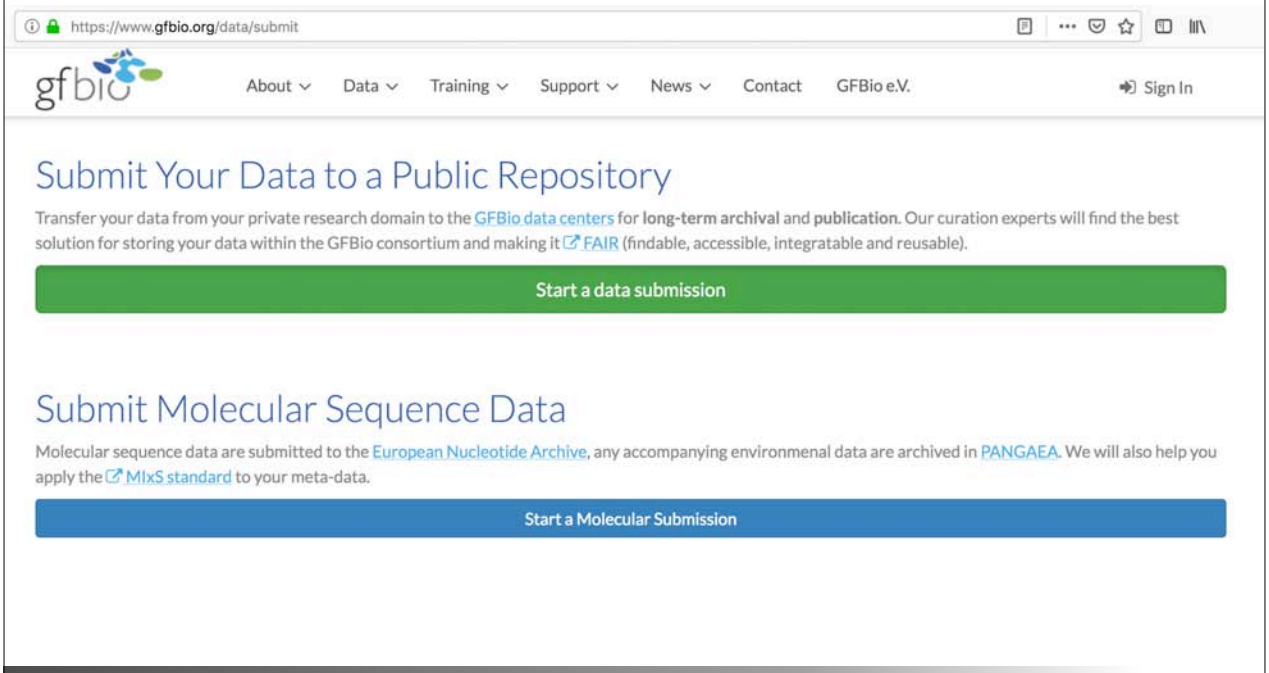
ENA

European Nucleotide Archive

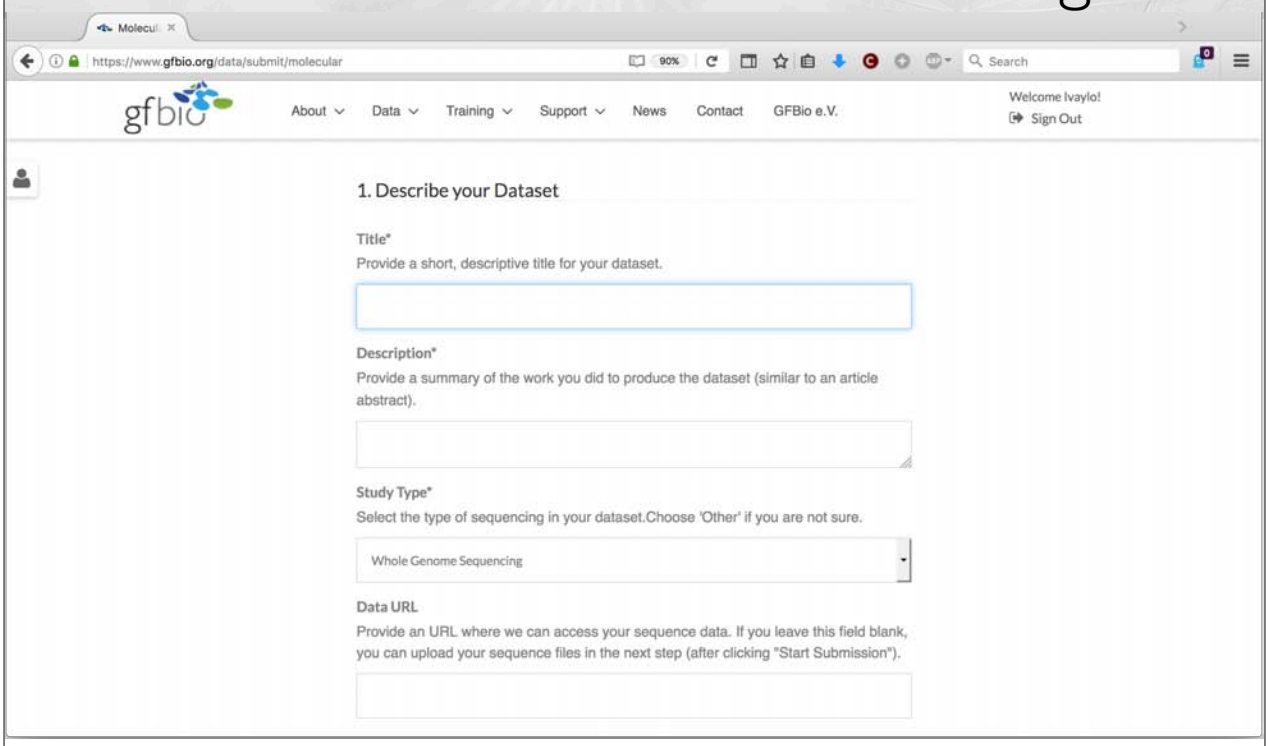
INSDC International Nucleotide Sequence Database Collaboration

dedicated, long-term archives

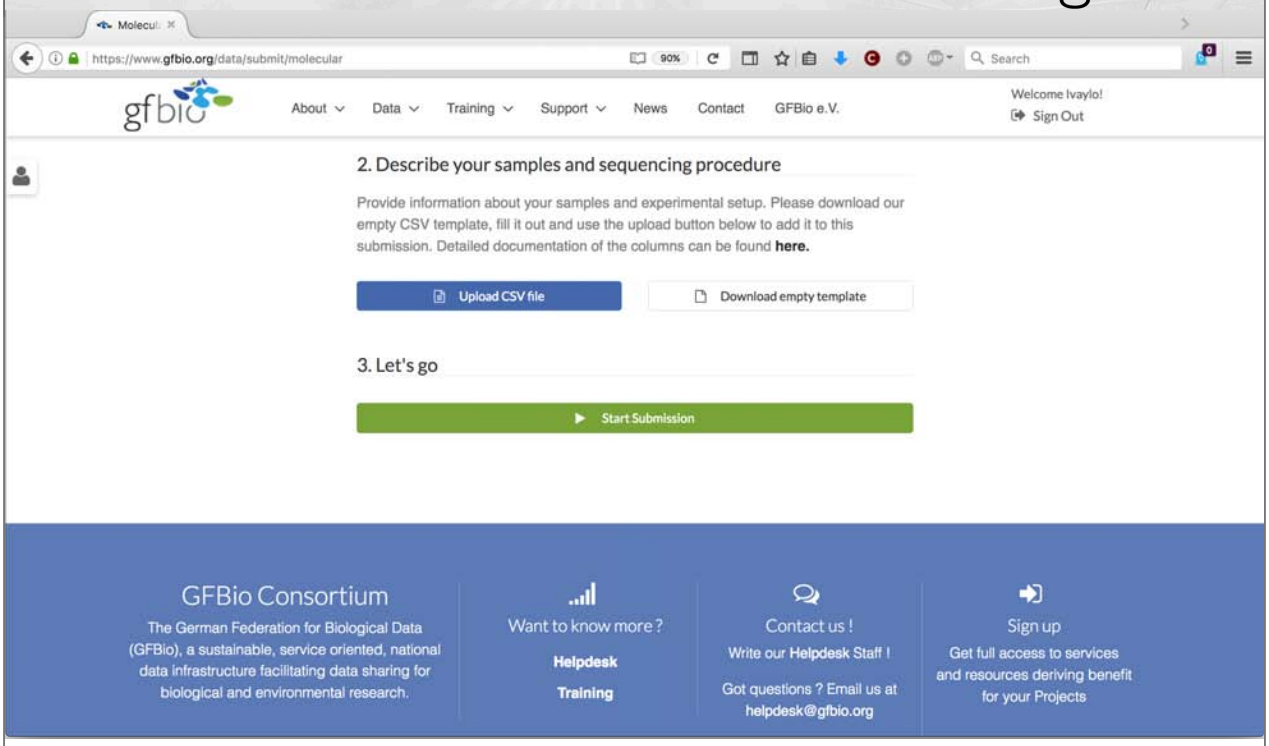
Data Submission



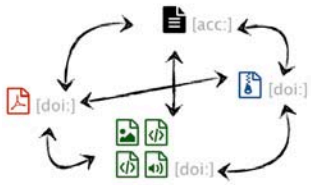
Molecular data submission



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

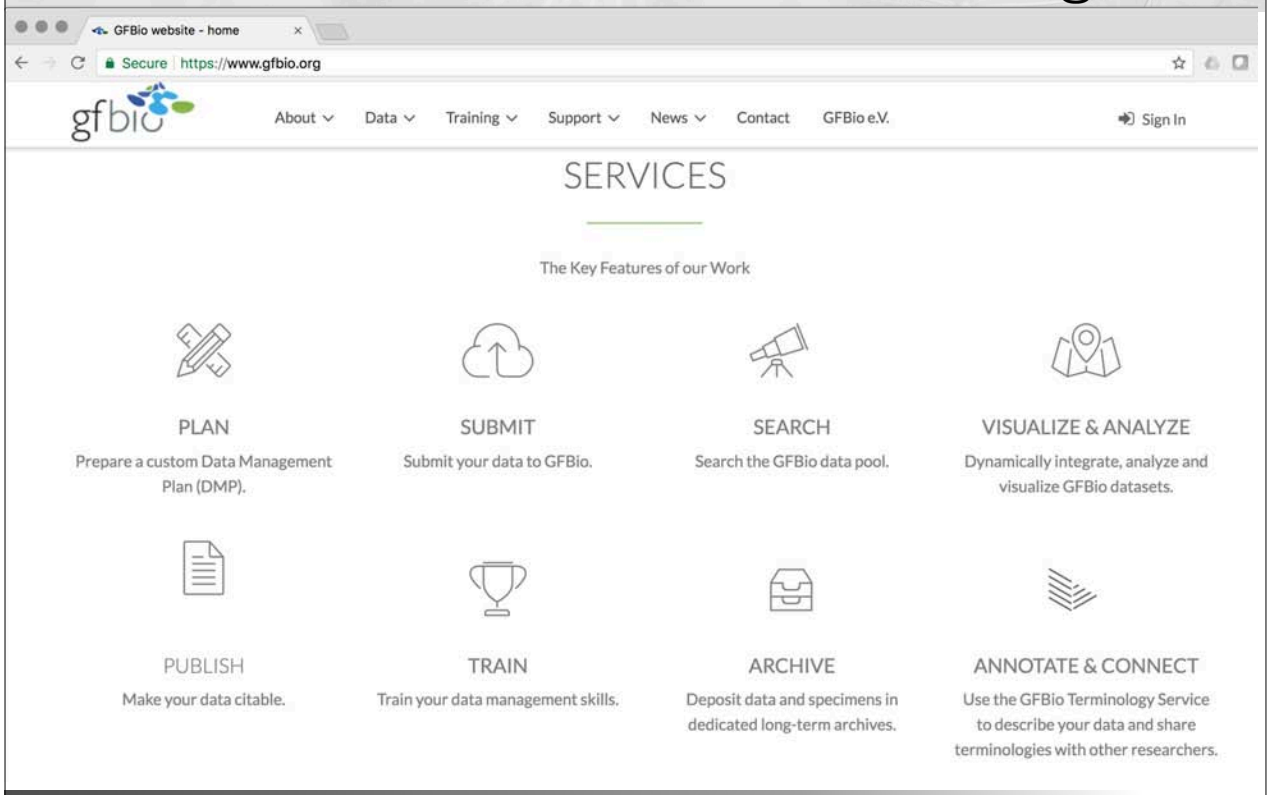
BEXIS RESEARCH DATA MANAGEMENT

http://bexis2.uni-jena.de/

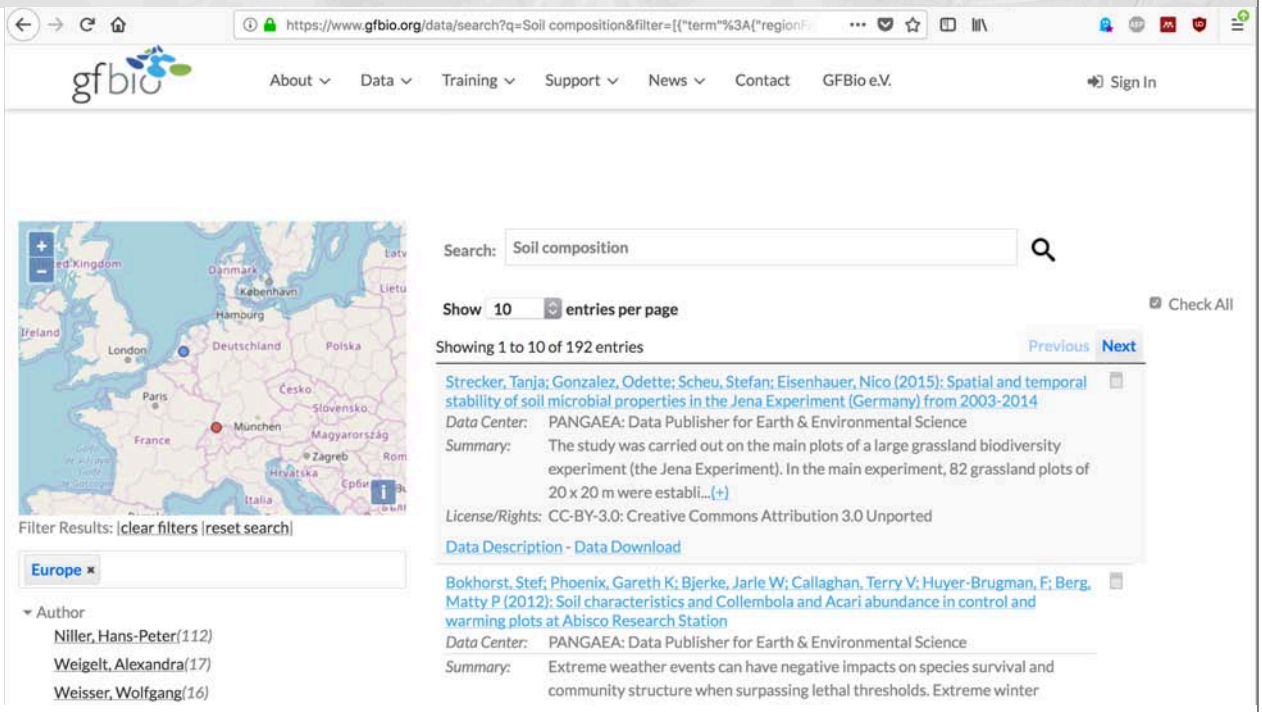
http://diversityworkbench.net

Our Services

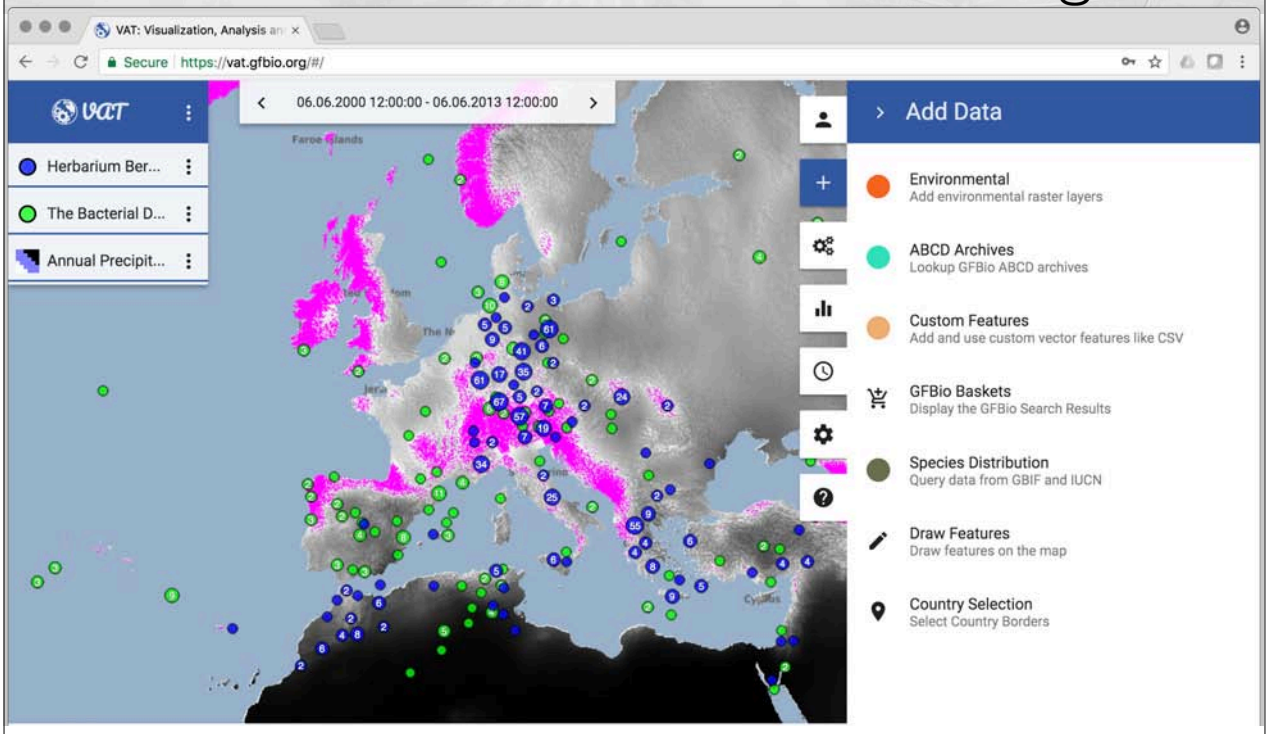
gfbio e.V.



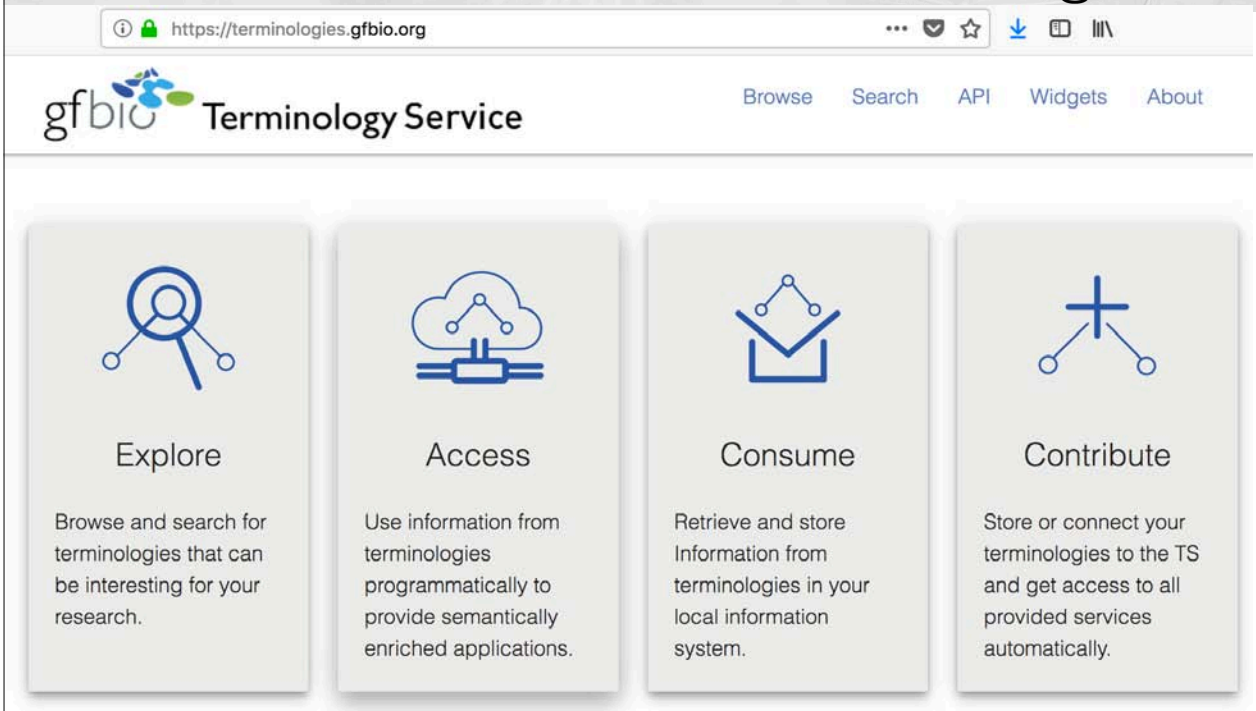
Search



Visualization & Analysis



Terminology Service



We need YOU!

If we build it they will come!?!

Nelson, Nature 461, 160-163 (2009)

Sustainability model

gfbio e.V. [est. 2016]

BG BM Botanischer Garten & Botanisches Museum Berlin

Freie Universität Berlin

Botanische Staatssammlung München

DSMZ Leibniz-Institut DSMZ-Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH

GWDG

Helmholtz Centre for Environmental Research – UFZ

IPK Gatersleben

Museum für Naturkunde Berlin

Senckenberg world of biodiversity

Naturkunde Museum Stuttgart

staatliche naturwissenschaftliche sammlungen bayerns

Friedrich-Schiller-Universität Jena

SUB Niedersächsische Staats- und Universitätsbibliothek Göttingen

Forschungsmuseum Koenig

gfbio e.V.

GERMAN FEDERATION FOR BIOLOGICAL DATA

DFG Deutsche Forschungsgemeinschaft

info@gfbio.org
@GFBio_Project
www.gfbio.org