gfbio
GERMAN FEDERATION
FOR BIOLOGICAL DATA

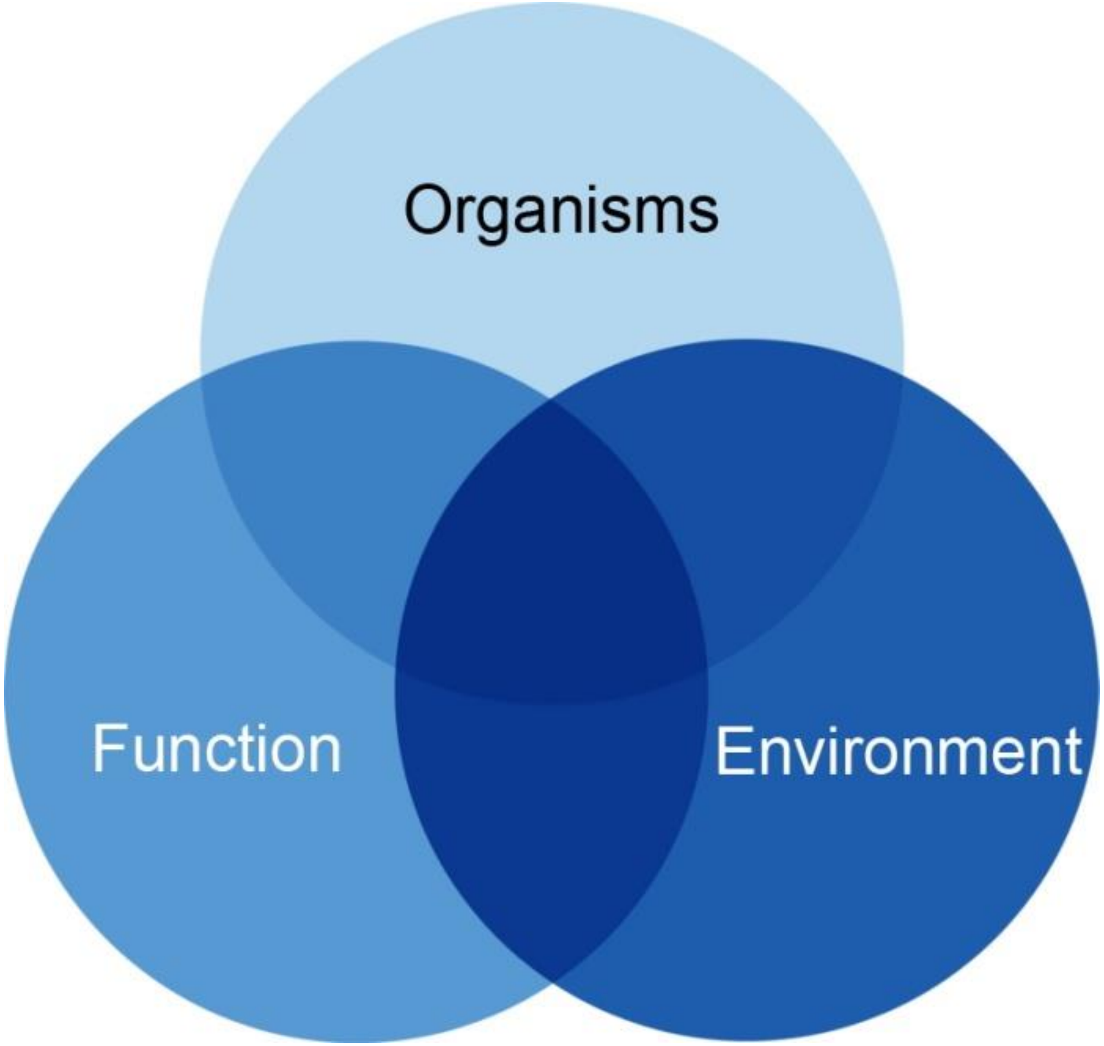
Research Data Management

Step by step through the Data Life Cycle

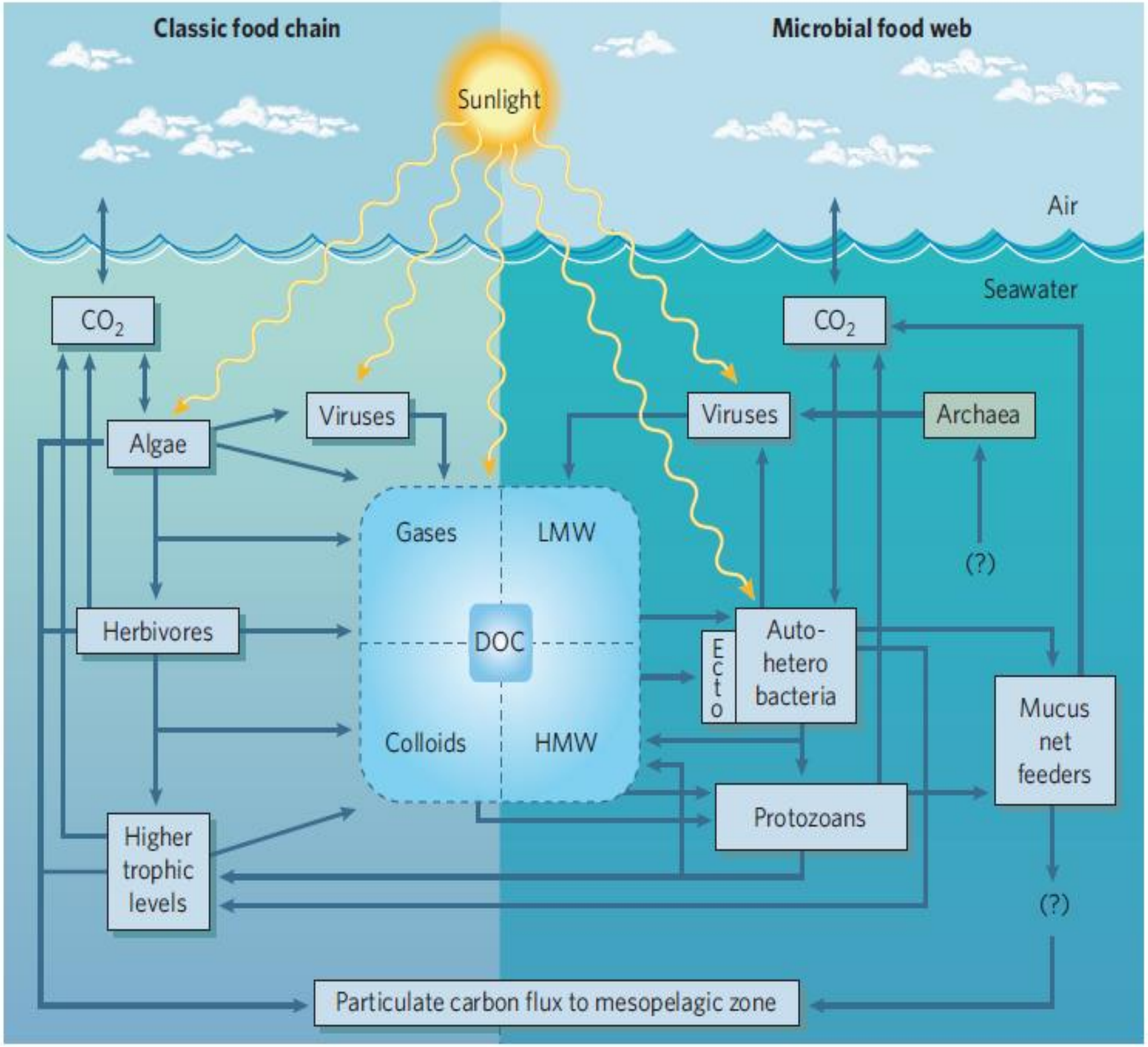
DFG Deutsche
Forschungsgemeinschaft

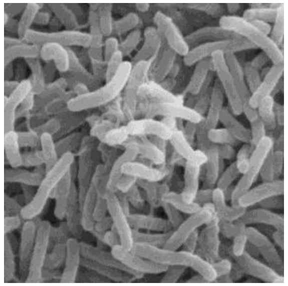
The Vision

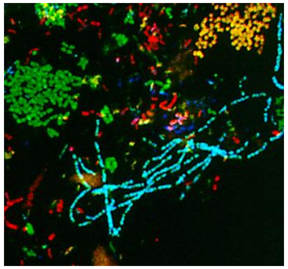
Ecosystems Biology

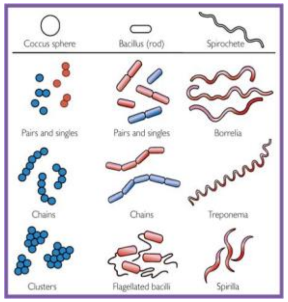


The Marine Foodweb







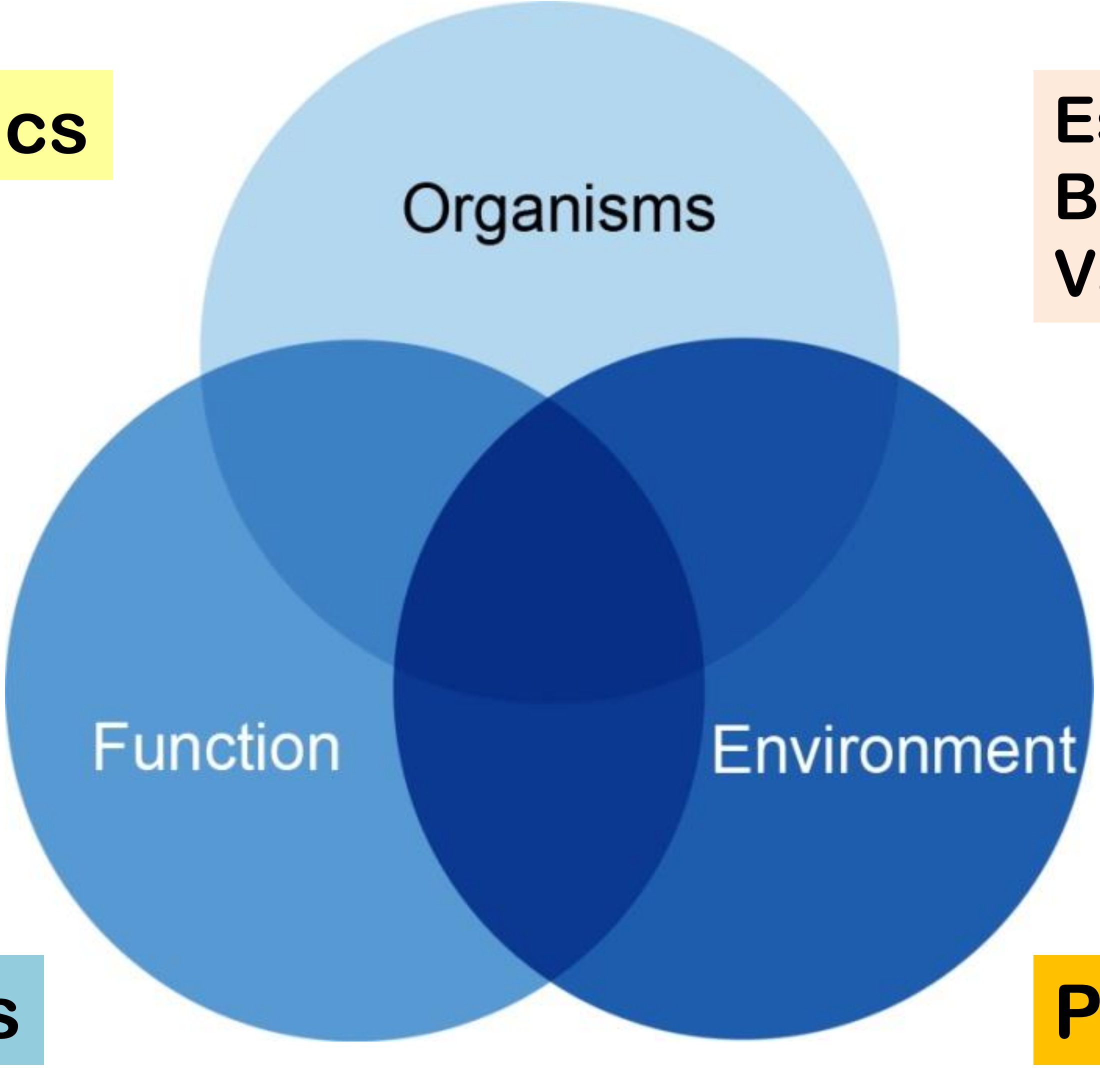


DeLong et al., Nature, Vol. 437, 2005

Ecosystems Biology

Statistics

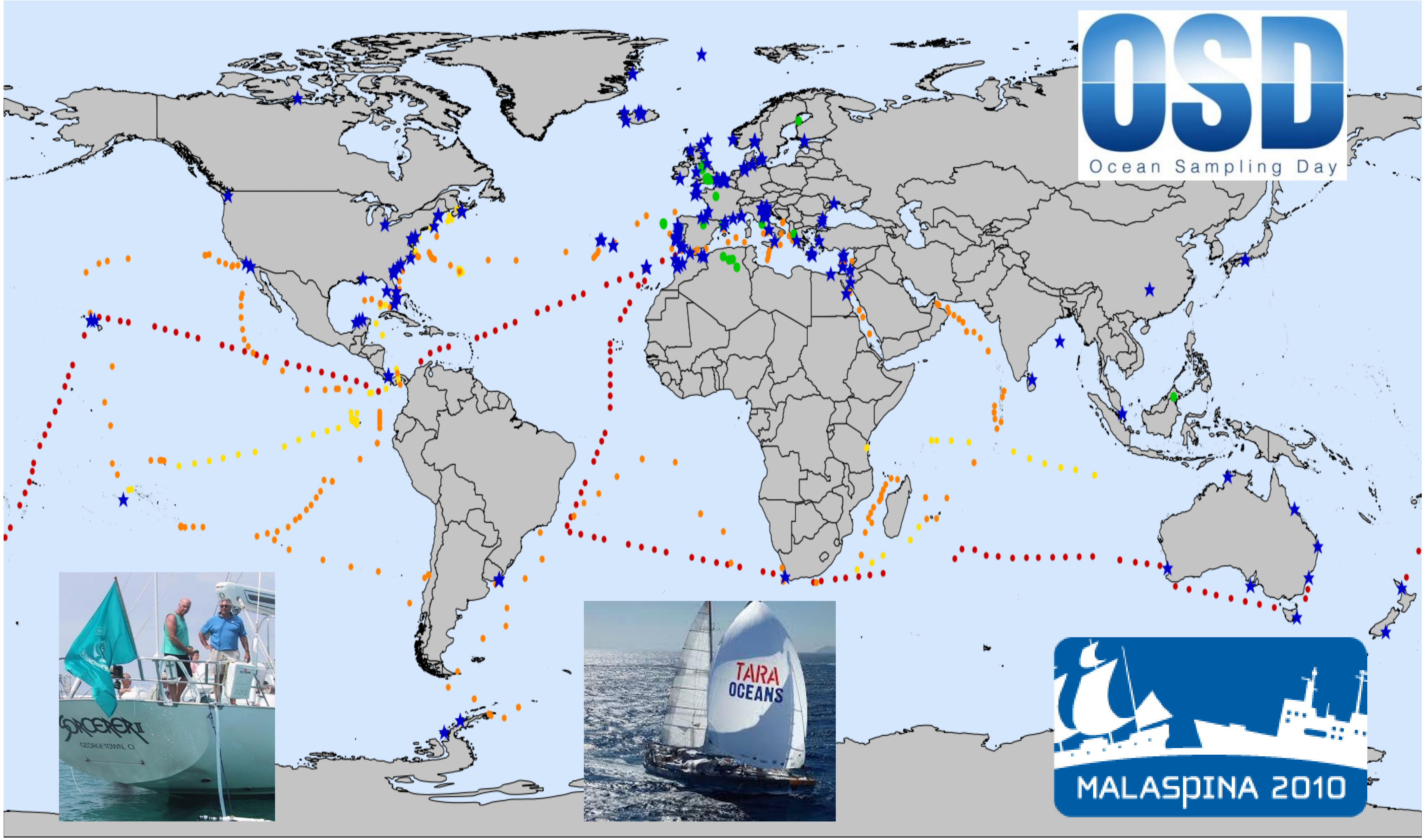
Essential Biodiversity Variables



Models

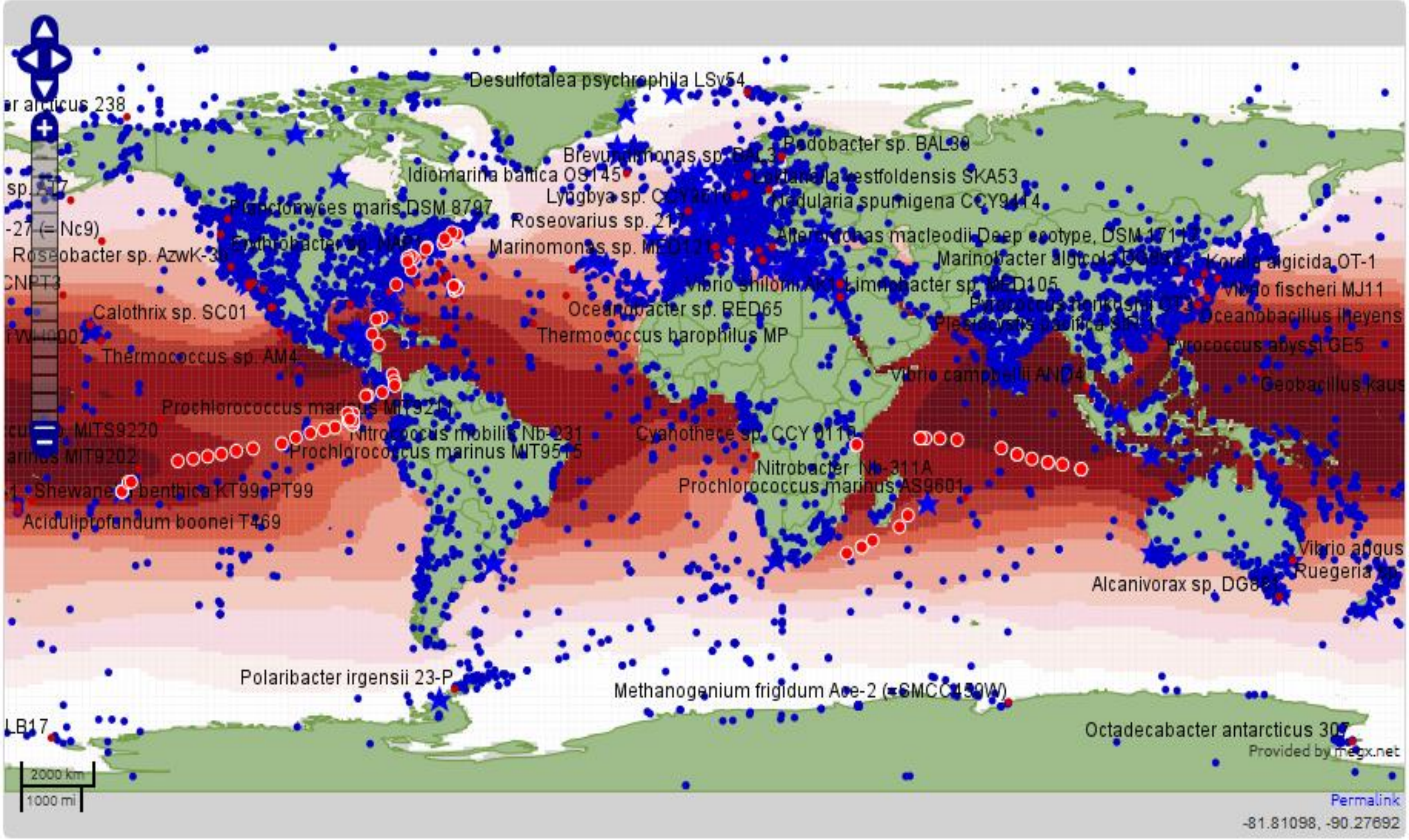
Predictions

Marine Megasequencing Projects



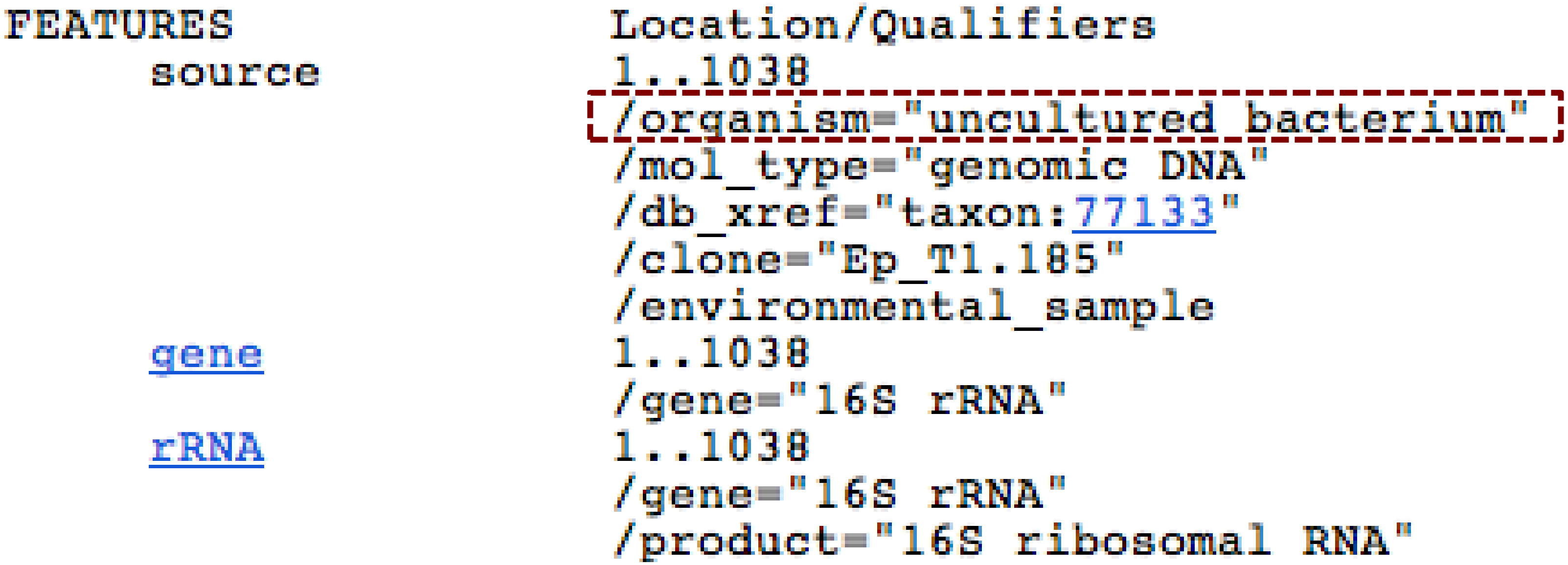
OSD: blue stars, RSD: green dots, Tara Oceans: orange dots, Malaspina cruise: red dots, Global Ocean Sampling (GOS): yellow dots.

Data Integration



The Reality

‘Abandoned’ sequences in INSDC databases



```
FEATURES             Location/Qualifiers
     source          1..1038
                     /organism="uncultured bacterium"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:77133"
                     /clone="Ep_T1.185"
                     /environmental_sample
     gene            1..1038
                     /gene="16S rRNA"
     rRNA            1..1038
                     /gene="16S rRNA"
                     /product="16S ribosomal RNA"
```

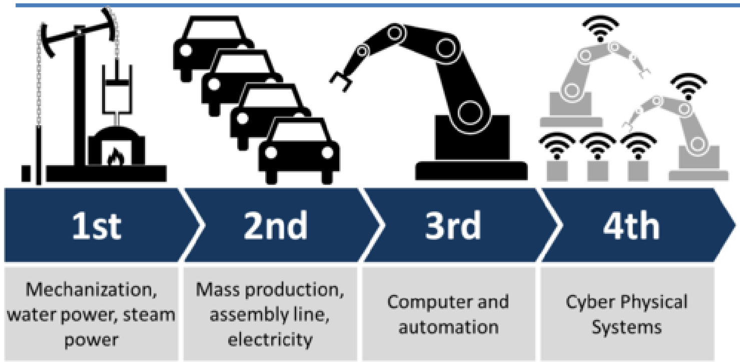
8% with coordinates (latitude/longitude)

9% with collection date

41% with taxonomic assignment

Pelin Yilmaz

Big Data


1st
Mechanization, water power, steam power
2nd
Mass production, assembly line, electricity
3rd
Computer and automation
4th
Cyber Physical Systems

INDUSTRY
4.0

Google

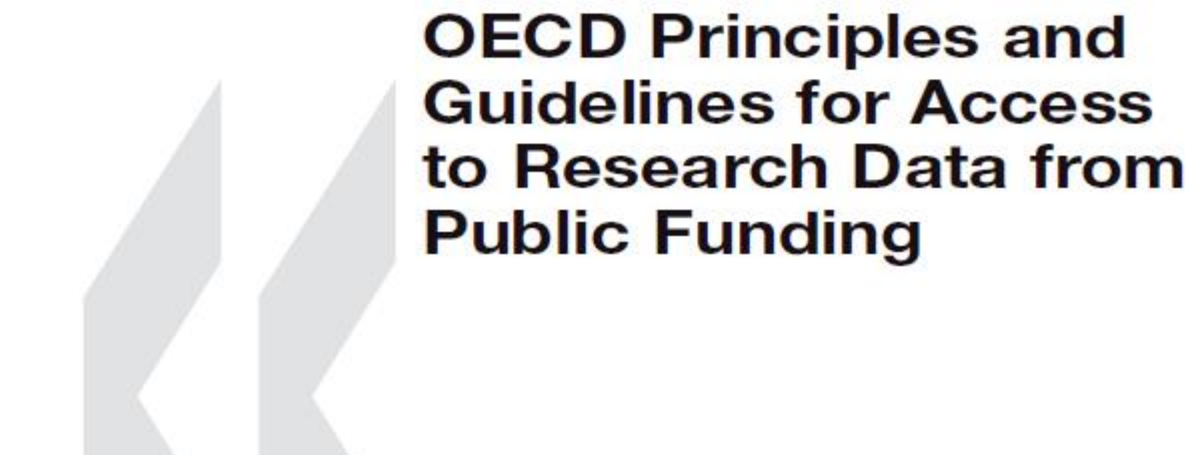
facebook

INTERNET of
THINGS

access
open
organisations
research
science
European
infrastructures
researchers
data funding
innovation
States
stakeholders
Member
knowledge
development
conference
companies
faster
re-use
Council
Agenda
plans
national
perspective
making
need
interest

science 2.0
JOIN THE REVOLUTION

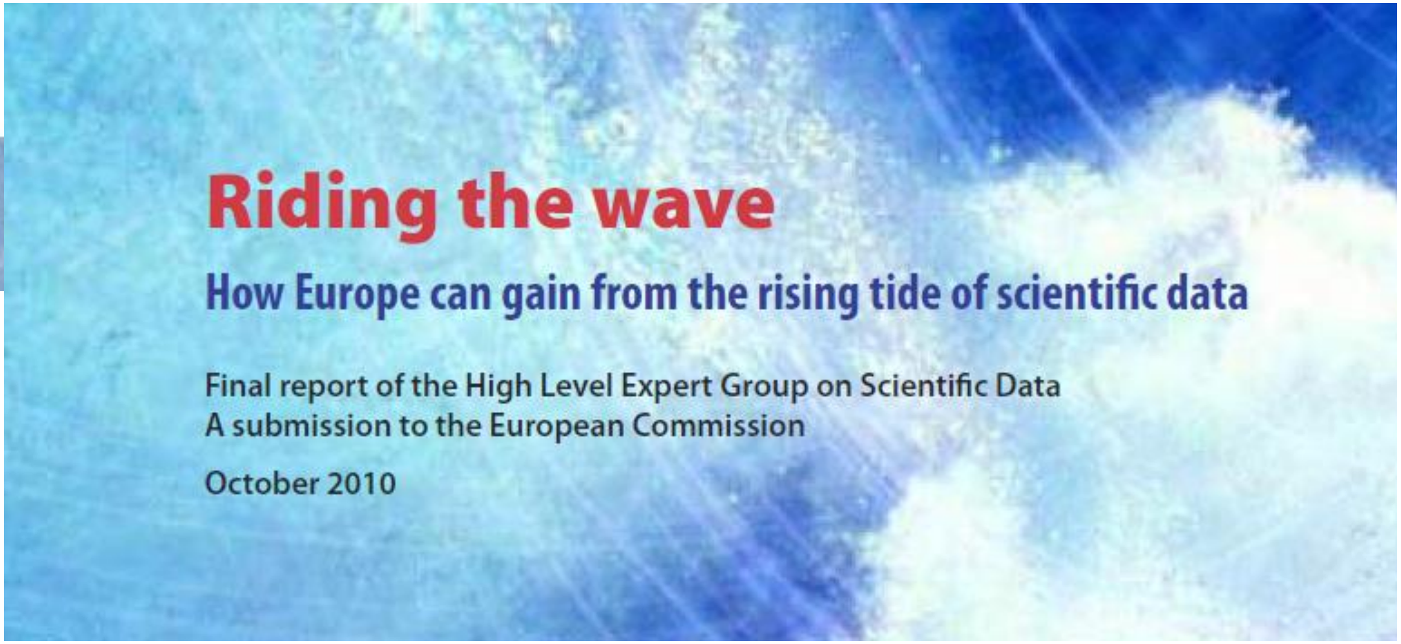
Value of Research Data



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

2007

2010



Riding the wave

How Europe can gain from the rising tide of scientific data

Final report of the High Level Expert Group on Scientific Data
A submission to the European Commission

October 2010

Value of Research Data

KE

Knowledge Exchange

A SURFBOARD
FOR RIDING THE WAVE

TOWARDS A FOUR COUNTRY ACTION PROGRAMME
ON RESEARCH DATA

2011

The Data Harvest:

How sharing research data can yield knowledge, jobs and growth

An RDA Europe Report

December 2014

2014

Sowing the seed: Incentives and motivations for sharing research data, a researcher's perspective

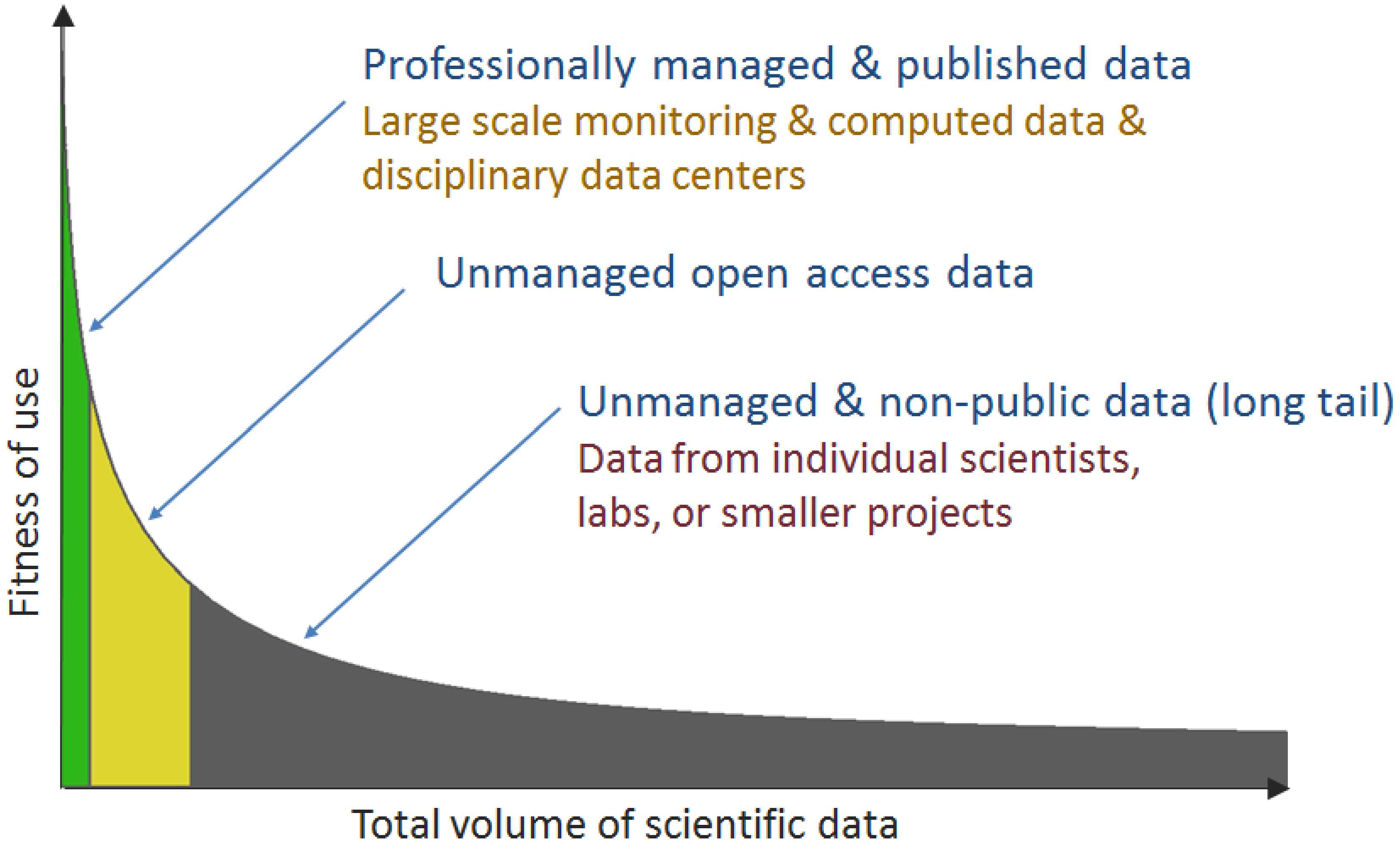
DFG Jisc CSC DEff SURF

Summary

Daten
Finanzierung
Infrastrukturen
Forschung
verfügbar
Austausch
Nutzer
data
Zugang
Förderer
Standards
Ausbildung
Forscher
Training
Bessere
öffentlich
Verlage
Norm
Credits
Geringe
Unterstützung
Wissenschaft
Zusammenarbeit
Erwartungshaltung
Universitäten
Wirtschaft
Bewusstsein
essentieller
Teil
Karriere
Sichtbarkeit
vorhanden
Länder
Policy
Umgang
Institute
Fähigkeiten
Deutschland
Infrastruktur
infrastructure
Konsortium
Anerkennung
e-Infrastrukturen
EU
Bestanteil
policy
Wertschätzung
Bedarf
of
Metadaten
mehr
stellen
Entscheidungen
müssen
Datenaustausch
the
Nutzung
Befeuerung
wissenschaftliche
scientists
nachhaltige
Rahmenbedingungen
Karrierewege
Qualitätssicherung
Wann
FDZ
kommerzielle
E-Infrastrukturen
Data
Bestandteil
Internationale
Mikrodaten
Archivierung
Mehrwert
nationale
Förderung
Zentren
Hindernisse
Omics
Verfügung
Integrierte
zugänglich
Etablierung
Datenmanagement
DSZ
Aufbau
Archive
Lebenswissenschaften
Forschungsprozesses
becomes
Nationales
akademischen
Ebenen
z.B
klare
Zugangs
wissenschaftlichen
Kriterien
Forschungsdaten
e-infrastructure
Conduct
FDZ/DSZ
Fehlende
inter
Akkreditierung
access
national
rechtliche
Big
sichtbar
Zitate
Umfang
sollten
Aufgaben
and
Speicherung
systems
Geographic
The
Sichtbarer
England

http://www.wordle.net/

Reality



Professionally managed & published data
Large scale monitoring & computed data & disciplinary data centers

Unmanaged open access data

Unmanaged & non-public data (long tail)
Data from individual scientists, labs, or smaller projects

Graphic by Michael Diepenbroek (PANGAEA)

Dark Data (the long tail)

When asked, almost all scientists will quickly acknowledge that they are holding dark data, data that has never been published or otherwise made available to the rest of the scientific community. An example of dark data is the type of data that exists only in the **bottom left-hand desk drawer** of scientists on some media that is quickly aging and soon will be unreadable by commonly available devices. The data remains in this dark desk drawer, inaccessible to the scientific community until the scientist retires. At the point of retirement some scientists rush to find a more suitable home for their data, be they in the form of slides, photographs, specimens, or electronic media files. More often than not, even in a well-planned retirement the desk drawer is eventually emptied into a dumpster because no one, including the scientist, knows exactly what the data is since it **lacks adequate documentation.**

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

Dark Data (the long tail)

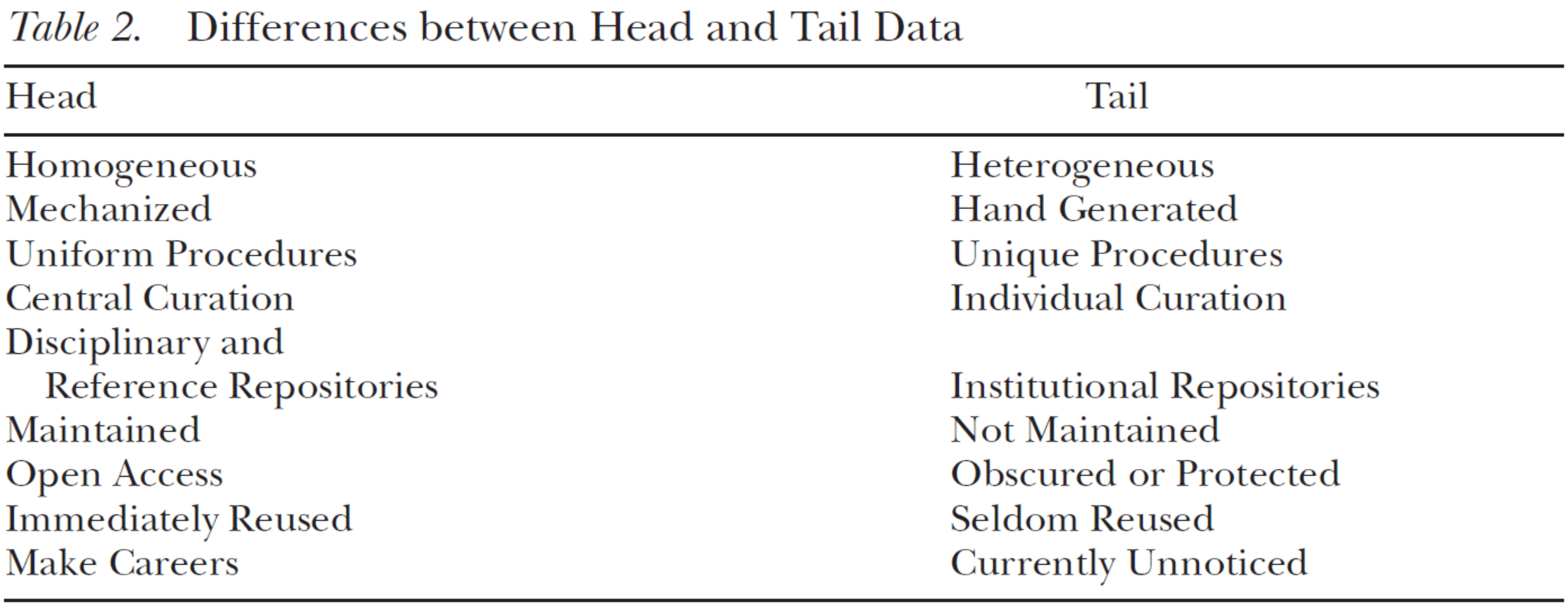


Table 2. Differences between Head and Tail Data

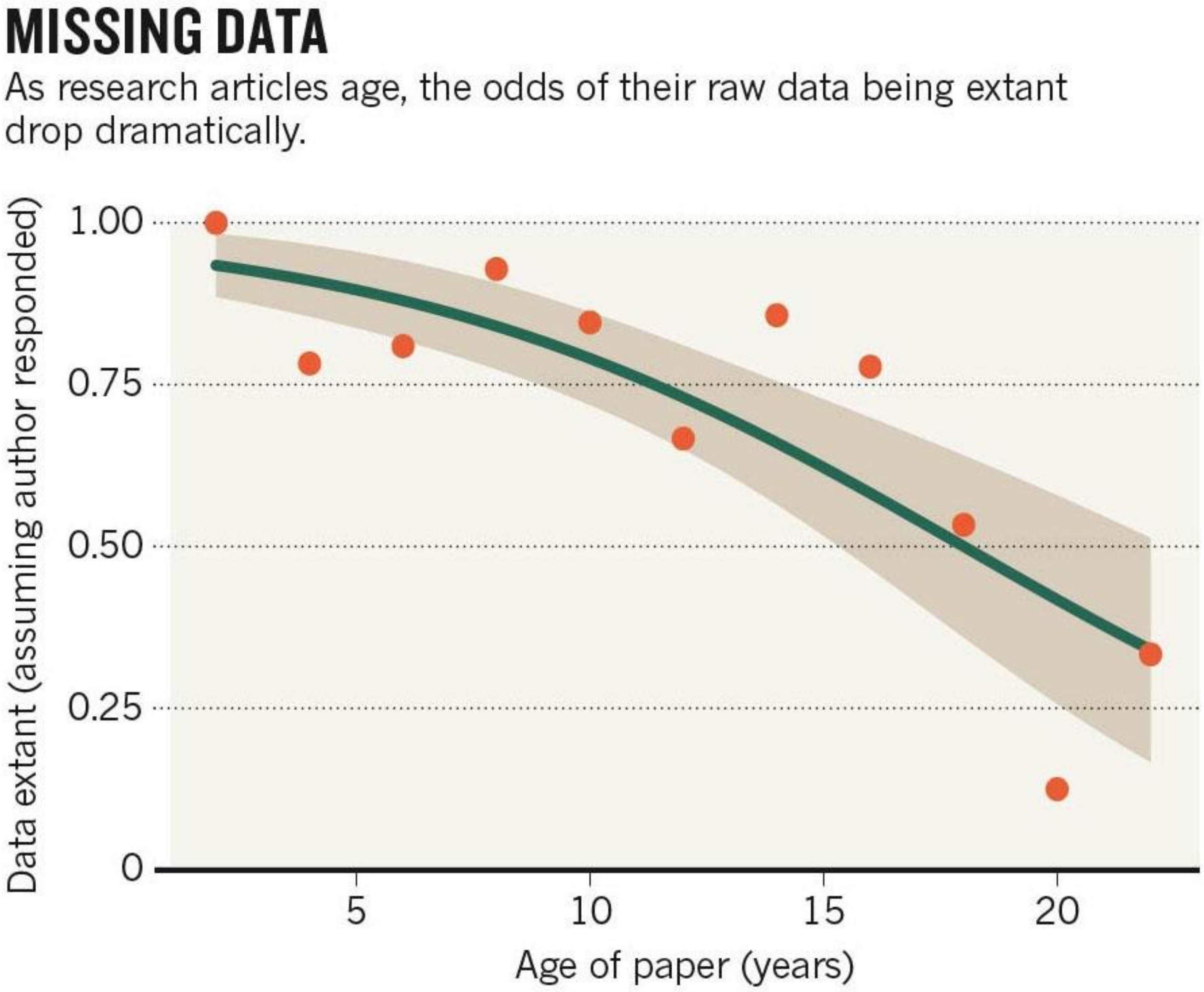
Head	Tail
Homogeneous	Heterogeneous
Mechanized	Hand Generated
Uniform Procedures	Unique Procedures
Central Curation	Individual Curation
Disciplinary and Reference Repositories	Institutional Repositories
Maintained	Not Maintained
Open Access	Obscured or Protected
Immediately Reused	Seldom Reused
Make Careers	Currently Unnoticed

20% by number of grants

80% by number of grants

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

Availability of Research Data with Time



Odds of data being lost are estimated to increase by 17% in every year after publication.

Find a working e-mail address for the first, last, or corresponding author fell by 7% per year.

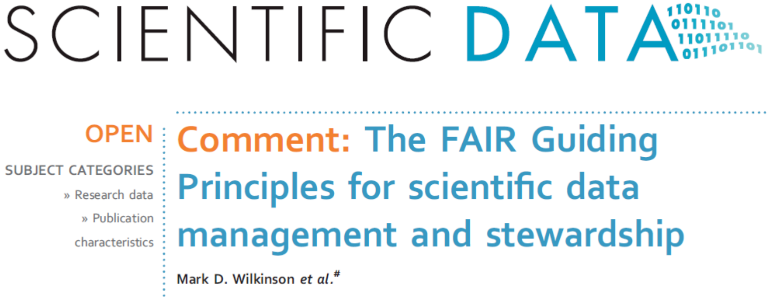
Overall, we only received 19.5% of the requested data sets, and only 11% for articles published before 2000.

Vines, Timothy H. et al. Current Biology , 2014, Volume 24 , Issue 1 , 94 - 97

The Solution?

FAIR Data

Findable, Accessible, Interoperable, Re-usable



SCIENTIFIC DATA

OPEN

SUBJECT CATEGORIES

» Research data

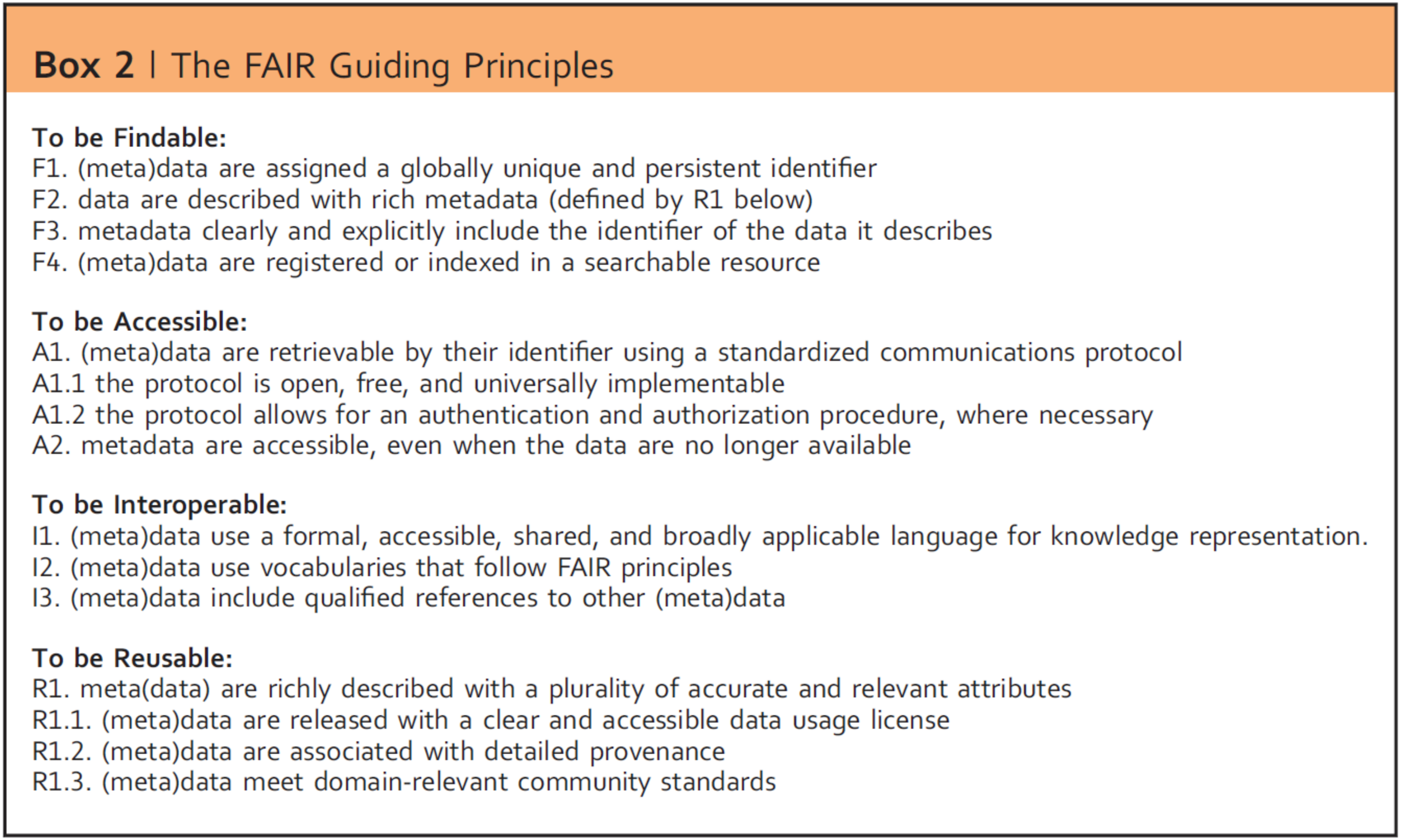
» Publication characteristics

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

Mark D. Wilkinson *et al.*#

http://www.nature.com/articles/sdata201618

FAIR Principles



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

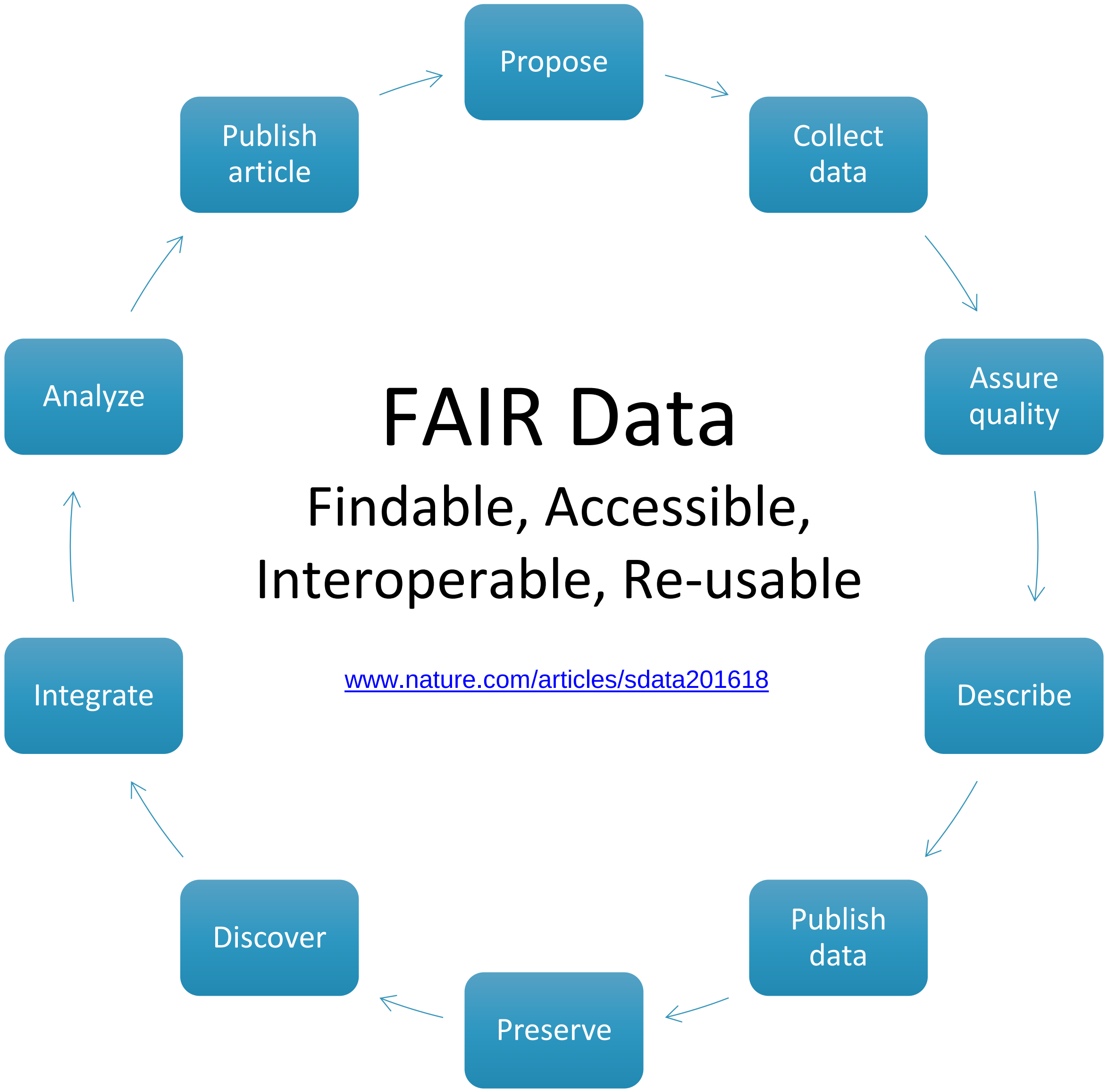
To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even 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

http://www.nature.com/articles/sdata201618

Data Life Cycle



Propose

Collect data

Assure quality

Describe

Publish data

Preserve

Discover

Integrate

Analyze

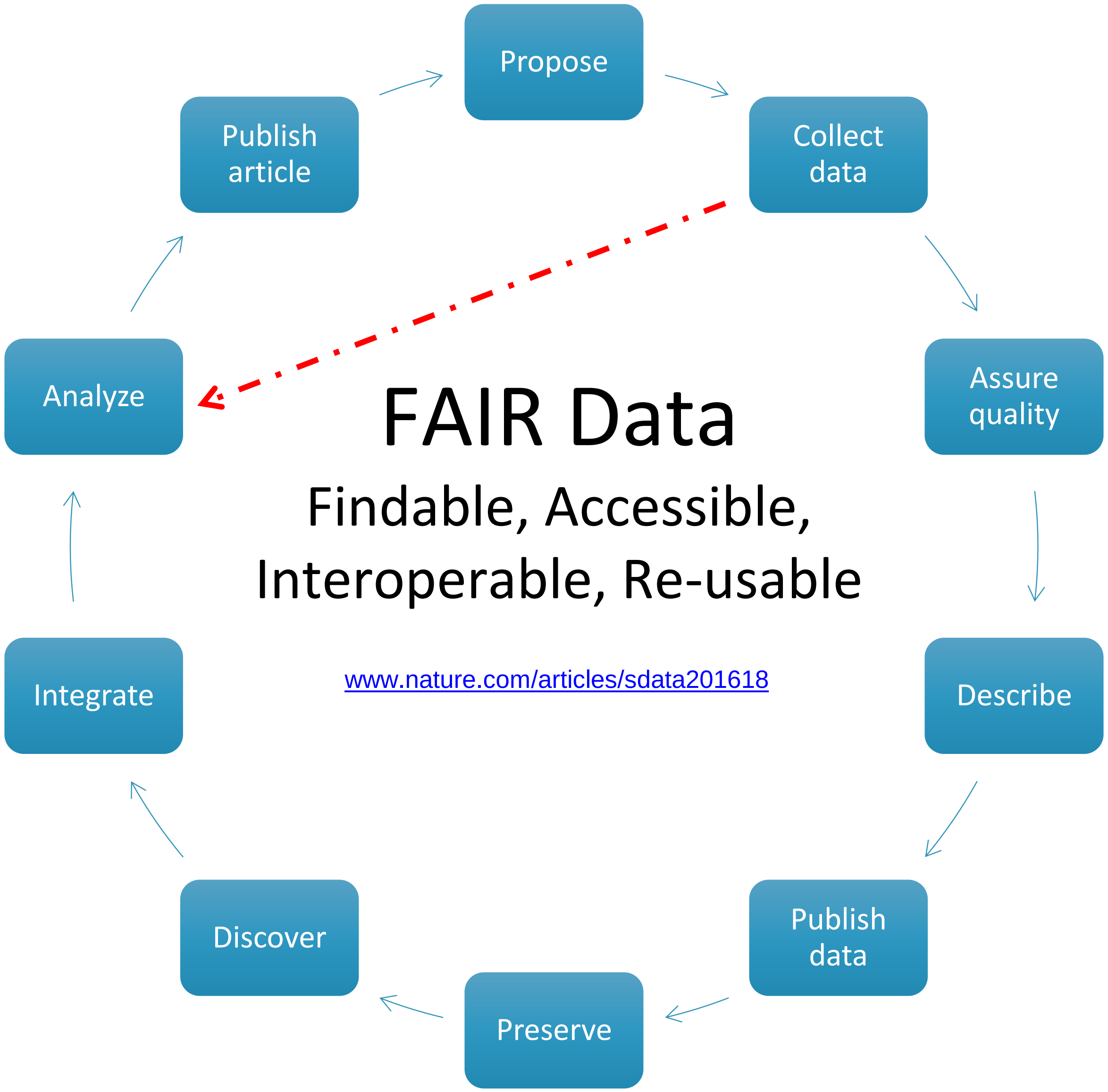
Publish article

FAIR Data

Findable, Accessible, Interoperable, Re-usable

www.nature.com/articles/sdata201618

Data Life Cycle



FAIR Data

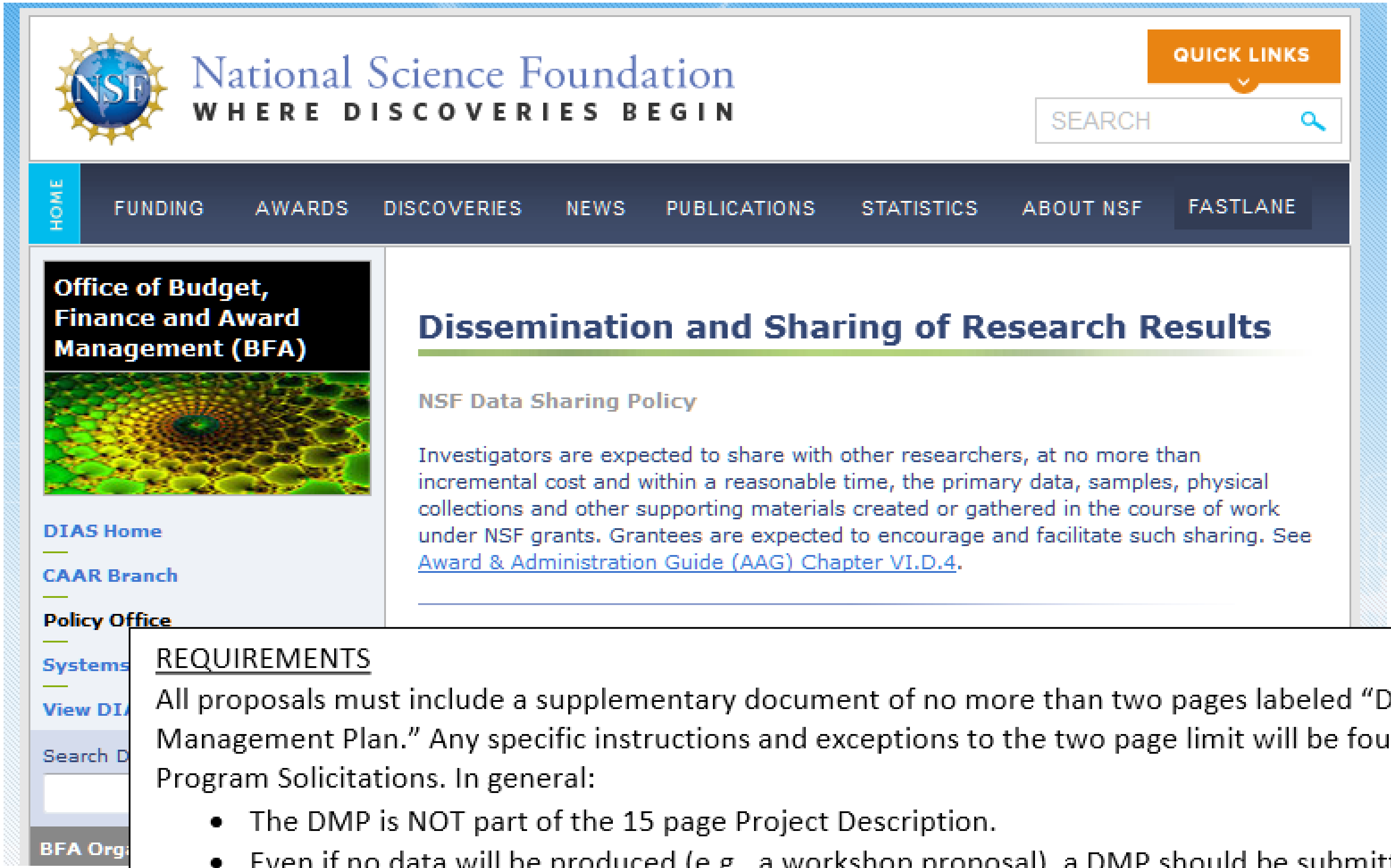
Findable, Accessible, Interoperable, Re-usable

www.nature.com/articles/sdata201618

Incentives

- Making data available is an essential part of the research process
  - It must be in the culture – the norm
- Career
  - Visibility – more citations
  - Credibility – more credits
  - Exchange – improve accessibility
- Standards
- Financial and legal framework
- Expectation “policy” by funders and publishers
- Adequate support and infrastructures

Example USA/NSF



Dissemination and Sharing of Research Results

NSF Data Sharing Policy

Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants. Grantees are expected to encourage and facilitate such sharing. See Award & Administration Guide (AAG) Chapter VI.D.4.

October 2015

REQUIREMENTS

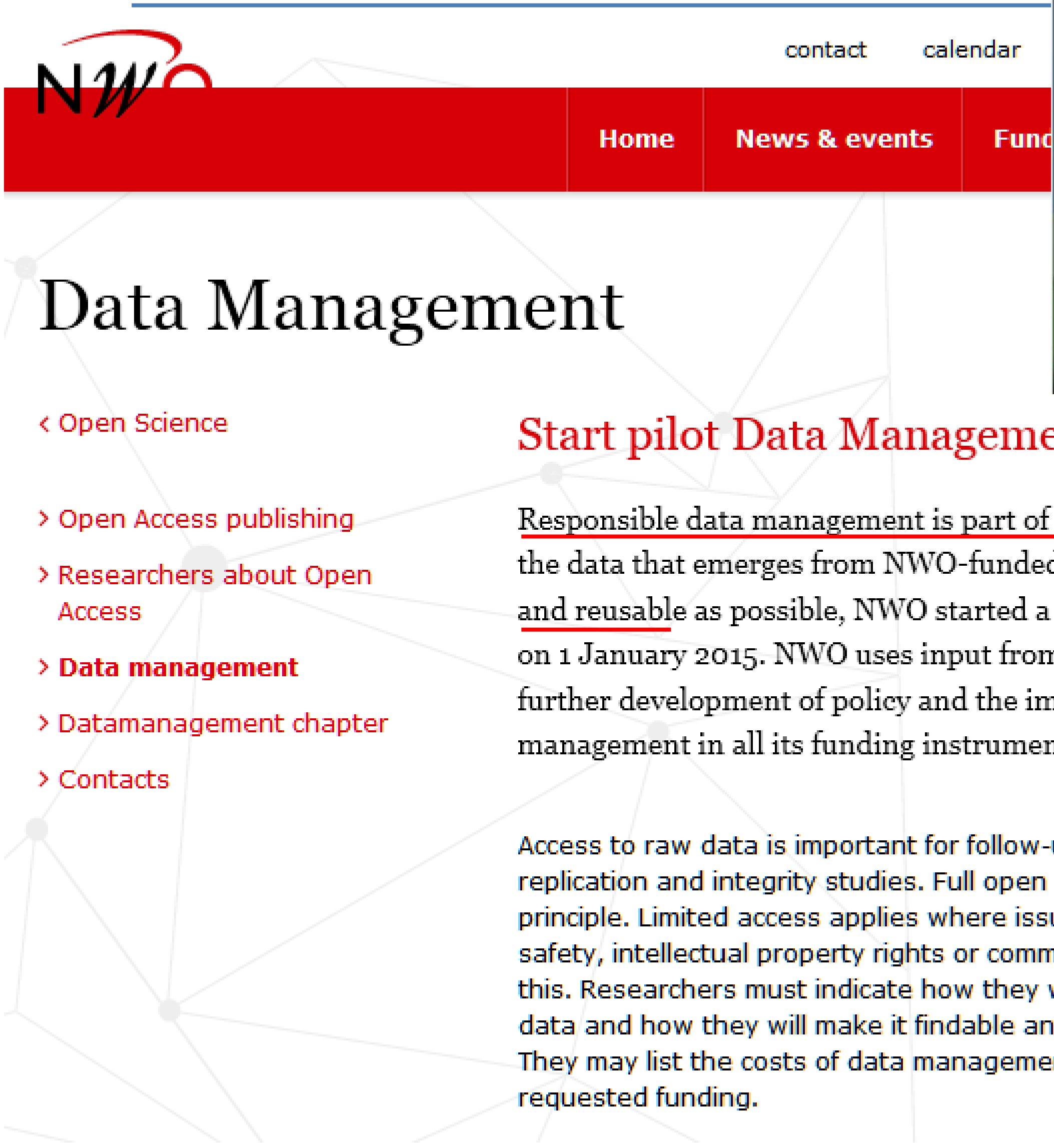
All proposals must include a supplementary document of no more than two pages labeled “Data Management Plan.” Any specific instructions and exceptions to the two page limit will be found in specific Program Solicitations. In general:

- The DMP is NOT part of the 15 page Project Description.
- Even if no data will be produced (e.g., a workshop proposal), a DMP should be submitted that states: “No data are expected to be produced from this project.”
- Proposals that do not include a Data Management Plan will be returned without review.

http://www.nsf.gov/bio/pubs/BIODMP_Guidance.pdf

http://www.nsf.gov/bfa/dias/policy/dmp.jsp

Example Netherlands



contact calendar

Home News & events Func

Data Management

- ‹ Open Science
- › Open Access publishing
- › Researchers about Open Access
- › **Data management**
- › Datamanagement chapter
- › Contacts

Start pilot Data Management

Responsible data management is part of good research. To make the data that emerges from NWO-funded research as accessible and reusable as possible, NWO started a pilot Data Management on 1 January 2015. NWO uses input from this pilot for the further development of policy and the implementation of data management in all its funding instruments.

Access to raw data is important for follow-up research and for replication and integrity studies. Full open access is the operating principle. Limited access applies where issues of privacy, public safety, intellectual property rights or commercial interests require this. Researchers must indicate how they will store their research data and how they will make it findable and suitable for re-use. They may list the costs of data management as part of the requested funding.

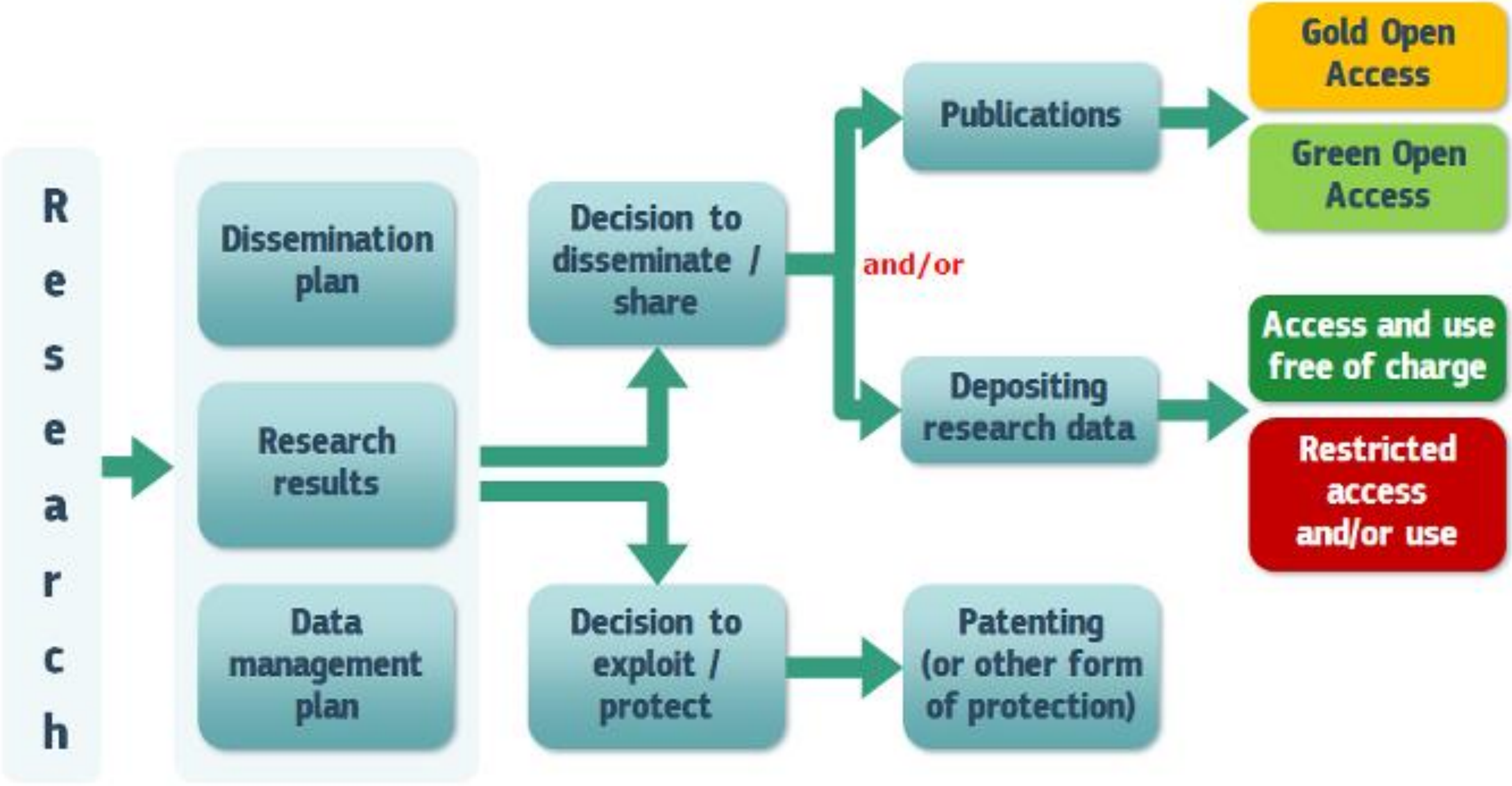
Is there any record of what these field names mean?

Yes! My co-author knows what the content of Sam2 is...

FAIR Data

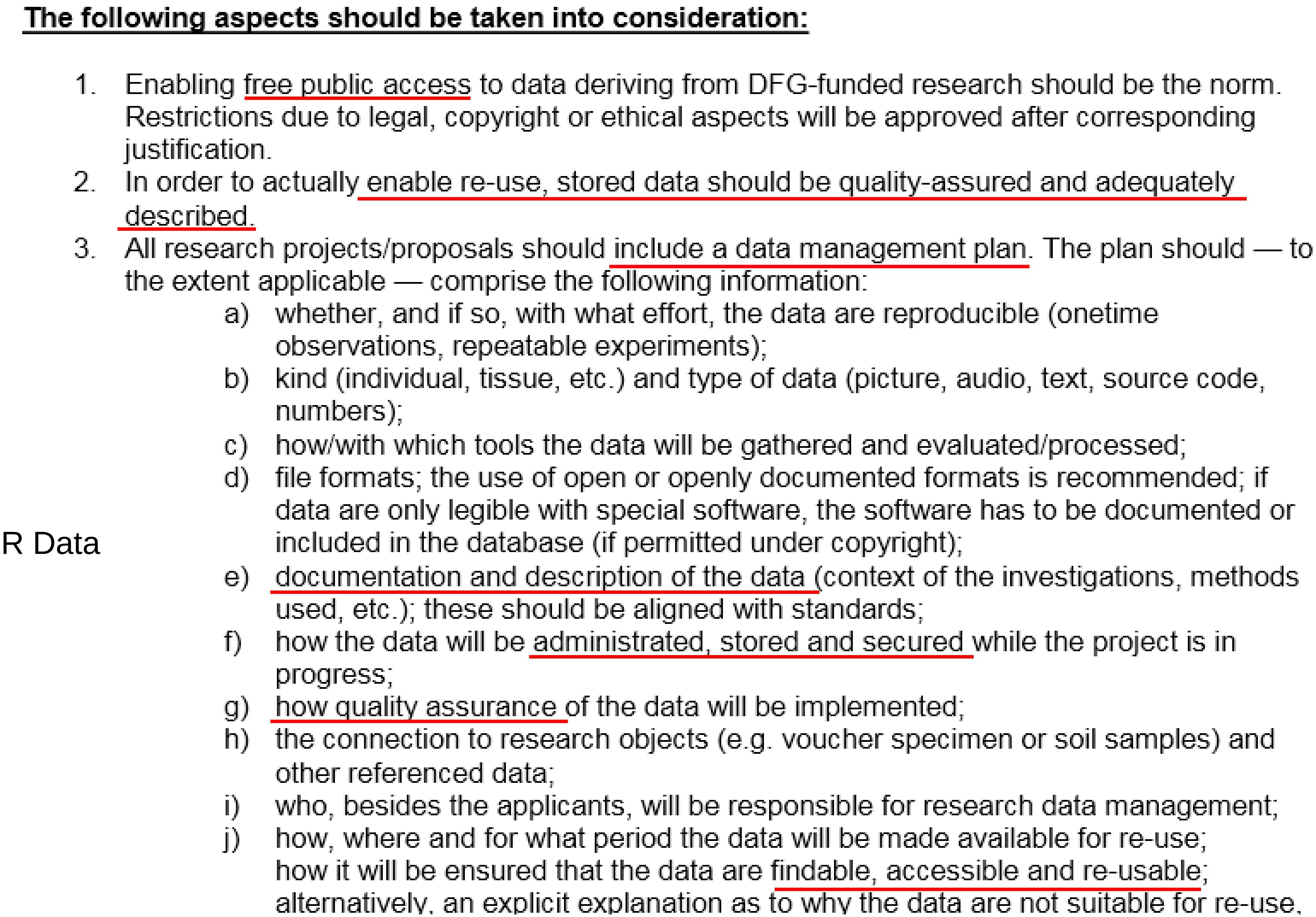
http://www.nwo.nl/en/policies/open+science/data+management

Example EU H2020


R
e
s
e
a
r
c
h
Dissemination plan
Research results
Data management plan
Decision to disseminate / share
Decision to exploit / protect
and/or
Publications
Depositing research data
Patenting (or other form of protection)
Gold Open Access
Green Open Access
Access and use free of charge
Restricted access and/or use

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

Example DFG – DMP



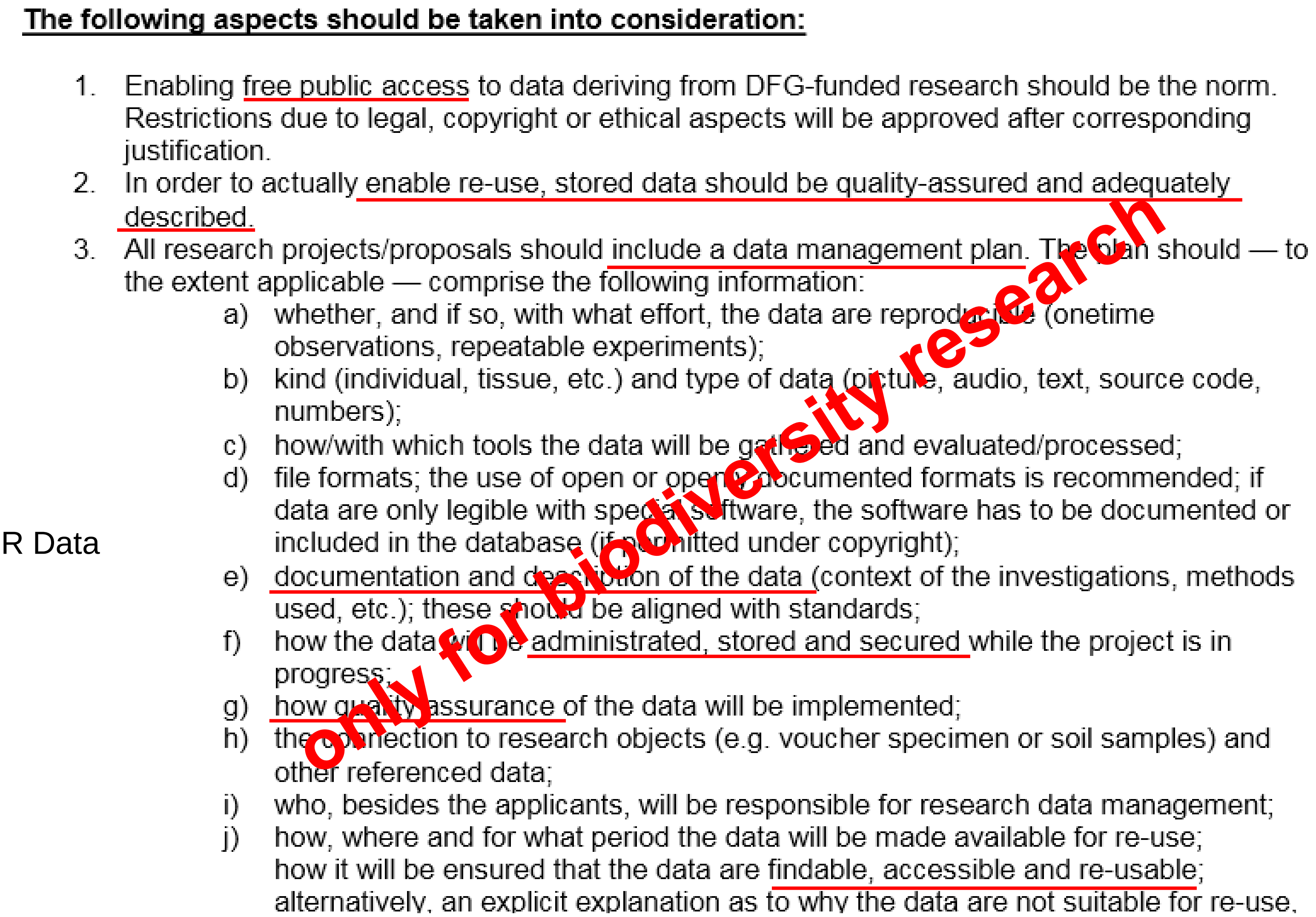
The following aspects should be taken into consideration:

1. Enabling free public access to data deriving from DFG-funded research should be the norm. Restrictions due to legal, copyright or ethical aspects will be approved after corresponding justification.
2. In order to actually enable re-use, stored data should be quality-assured and adequately described.
3. All research projects/proposals should include a data management plan. The plan should — to the extent applicable — comprise the following information:
   a) whether, and if so, with what effort, the data are reproducible (onetime observations, repeatable experiments);
   b) kind (individual, tissue, etc.) and type of data (picture, audio, text, source code, numbers);
   c) how/with which tools the data will be gathered and evaluated/processed;
   d) file formats; the use of open or openly documented formats is recommended; if data are only legible with special software, the software has to be documented or included in the database (if permitted under copyright);
   e) documentation and description of the data (context of the investigations, methods used, etc.); these should be aligned with standards;
   f) how the data will be administrated, stored and secured while the project is in progress;
   g) how quality assurance of the data will be implemented;
   h) the connection to research objects (e.g. voucher specimen or soil samples) and other referenced data;
   i) who, besides the applicants, will be responsible for research data management;
   j) how, where and for what period the data will be made available for re-use; how it will be ensured that the data are findable, accessible and re-usable; alternatively, an explicit explanation as to why the data are not suitable for re-use.

FAIR Data

http://www.dfg.de/download/pdf/foerderung/antragstellung/forschungsdaten/guidelines_biodiversity_research.pdf

Example DFG – DMP



The following aspects should be taken into consideration:

1. Enabling free public access to data deriving from DFG-funded research should be the norm. Restrictions due to legal, copyright or ethical aspects will be approved after corresponding justification.
2. In order to actually enable re-use, stored data should be quality-assured and adequately described.
3. All research projects/proposals should include a data management plan. The plan should — to the extent applicable — comprise the following information:
   a) whether, and if so, with what effort, the data are reproducible (onetime observations, repeatable experiments);
   b) kind (individual, tissue, etc.) and type of data (picture, audio, text, source code, numbers);
   c) how/with which tools the data will be gathered and evaluated/processed;
   d) file formats; the use of open or openly documented formats is recommended; if data are only legible with special software, the software has to be documented or included in the database (if permitted under copyright);
   e) documentation and description of the data (context of the investigations, methods used, etc.); these should be aligned with standards;
   f) how the data will be administrated, stored and secured while the project is in progress;
   g) how quality assurance of the data will be implemented;
   h) the connection to research objects (e.g. voucher specimen or soil samples) and other referenced data;
   i) who, besides the applicants, will be responsible for research data management;
   j) how, where and for what period the data will be made available for re-use; how it will be ensured that the data are findable, accessible and re-usable; alternatively, an explicit explanation as to why the data are not suitable for re-use.

only for biodiversity research

FAIR Data

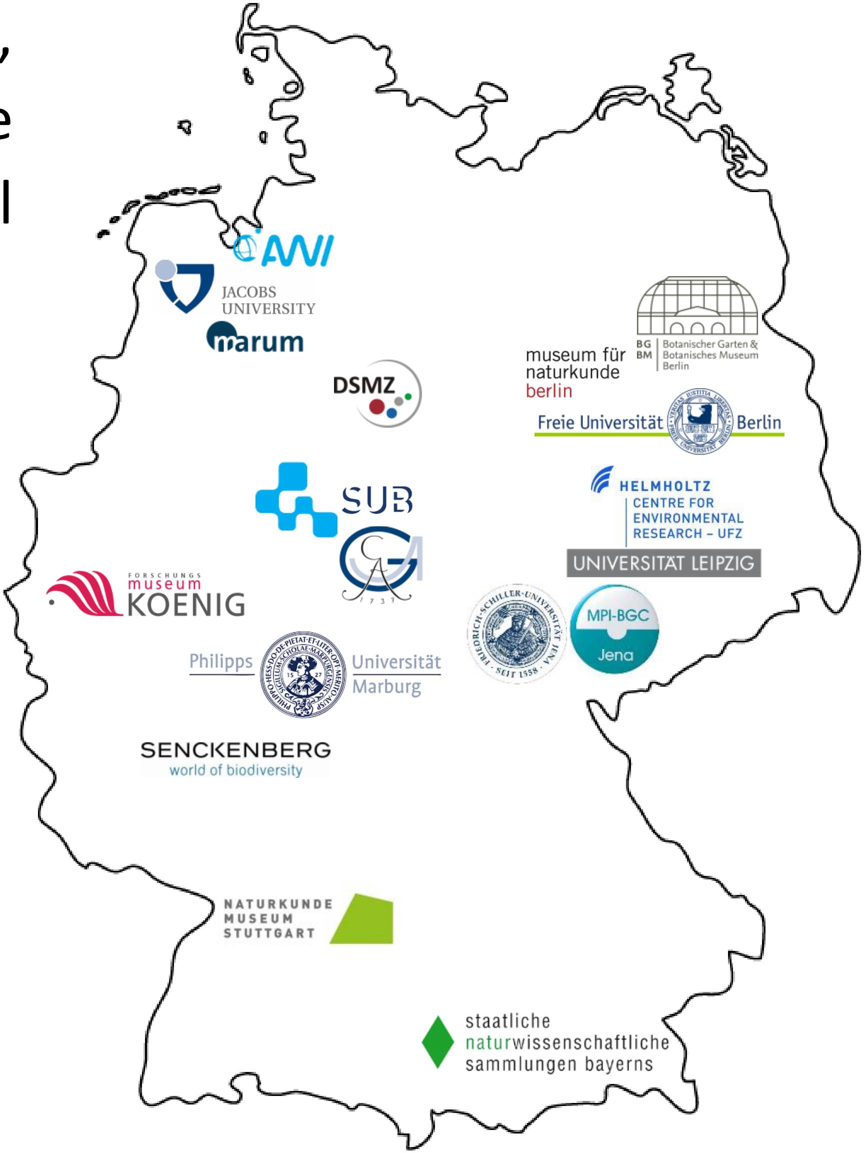
http://www.dfg.de/download/pdf/foerderung/antragstellung/forschungsdaten/guidelines_biodiversity_research.pdf

Incentives

- Making data available is an essential part of the research process
  - It must be in the culture – the norm
- Career
  - Visibility – more citations
  - Credibility – more credits
  - Exchange – improve accessibility
- Standards/SOPs
- Financial and legal framework
- Expectation "policy" by funders and publishers
- Adequate support and infrastructures

German Federation for Biological Data

Sustainable, service oriented, national data infrastructure facilitating data sharing for biological and environmental research.


AWI
JACOBS UNIVERSITY
marum
DSMZ
museum für naturkunde berlin
BG BM Botanischer Garten & Botanisches Museum Berlin
Freie Universität Berlin
SUB
HELMHOLTZ CENTRE FOR ENVIRONMENTAL RESEARCH – UFZ
UNIVERSITÄT LEIPZIG
FORSCHUNGS museum KOENIG
MPI-BGC Jena
Philipps Universität Marburg
SENCKENBERG world of biodiversity
NATURKUNDE MUSEUM STUTTGART
staatliche naturwissenschaftliche sammlungen bayerns

gfbio

Funded by

DFG Deutsche Forschungsgemeinschaft

www.gfbio.org

GFBio Services

- Single point of contact for:
  - – Data management
  - – Long-term data archival
  - – Integrated data discovery
  - – Visualization and analyses
- Helpdesk
- Support & Training

Data Management Plan

Should cover the following points:

- Data acquisition (size, type)
- Quality assurance, standards
- Intermediate handling and storage
- Long-term archiving (data centers)
- Analysis (tools)
- Publication (open-access)

Contact us info@gfbio.org

Long-term Data Archival

GFBio data centers and their services at a glance

- Collection data

staatliche naturwissenschaftliche sammlungen bayerns

BGBM

STAATLICHES MUSEUM FÜR NATURKUNDE STUTTGART
Forschungsmuseum
Am Löwentor und
Schloss Rosenstein

FORSCHUNGS museum KOENIG

SENCKENBERG
world of biodiversity

museum für naturkunde berlin

DSMZ

- Environmental data

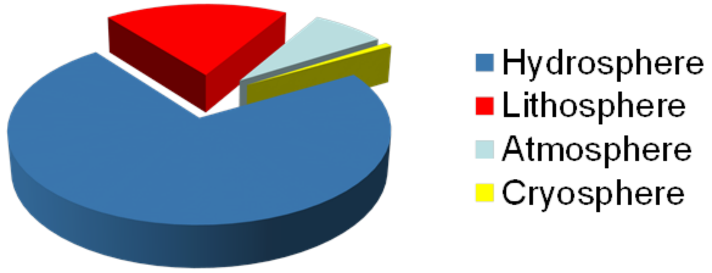
- Molecular data

ENA
European Nucleotide Archive

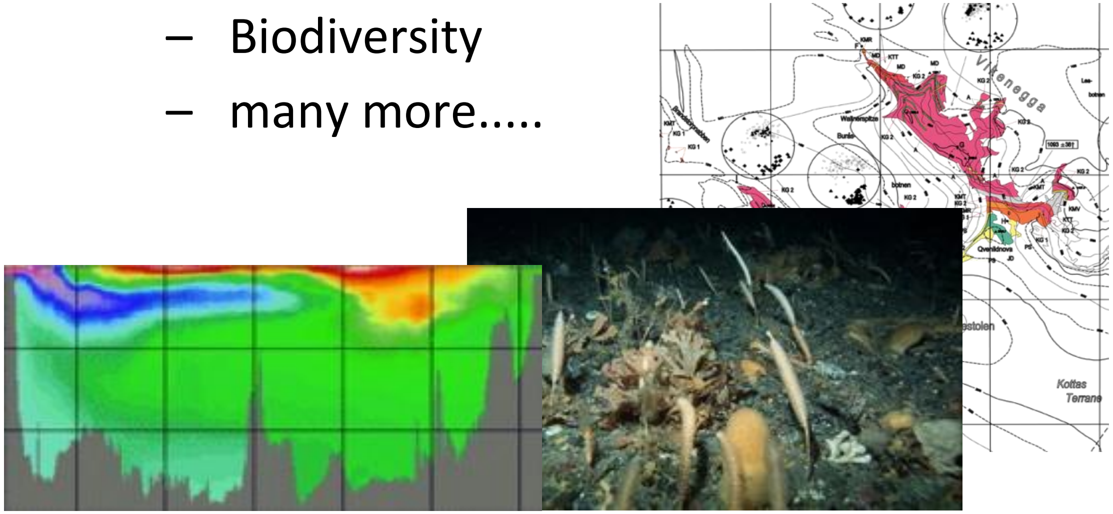
http://www.gfbio.org/data-centers

Environmental Data PANGAEA

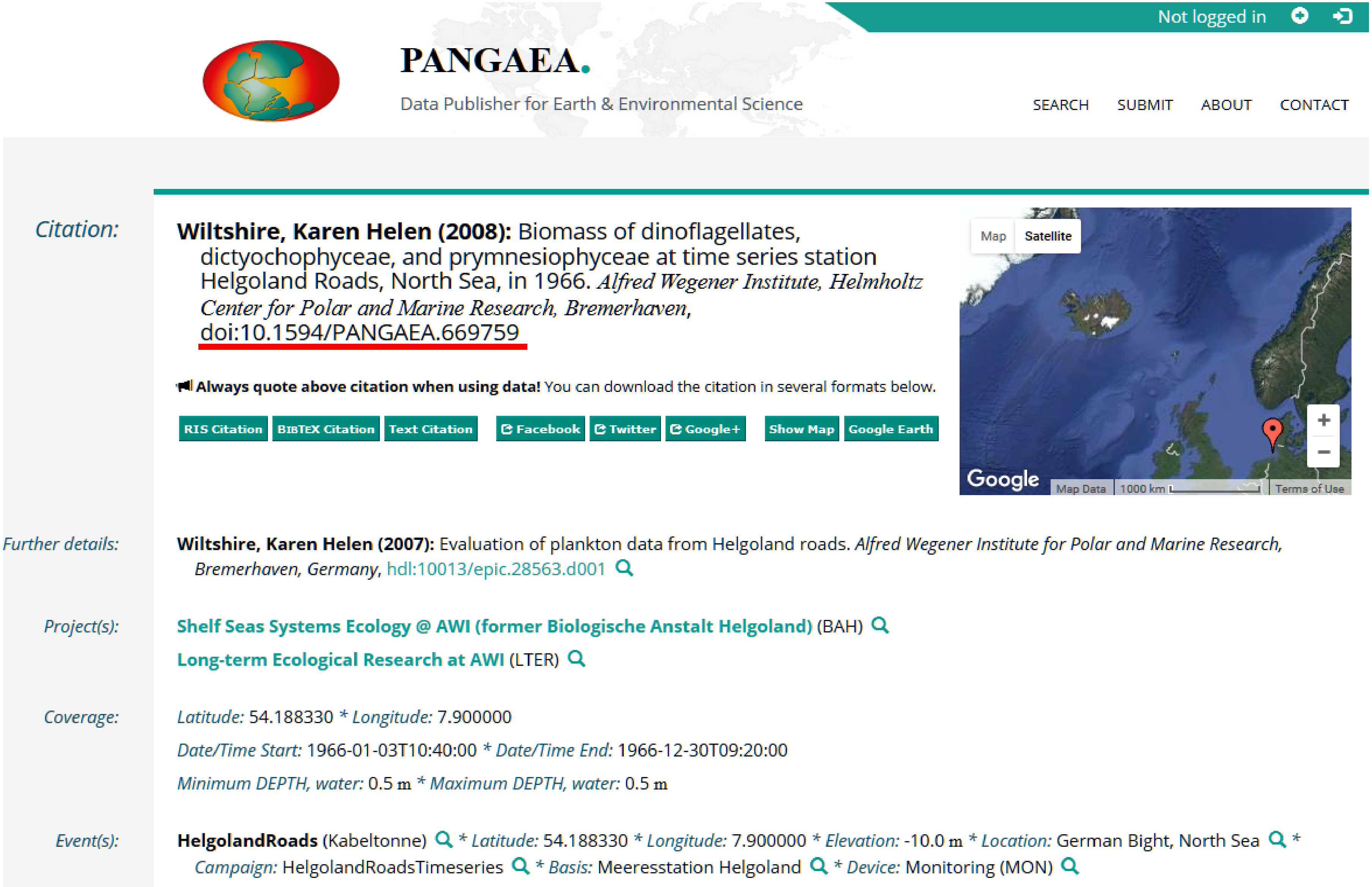
- Hosted by the MARUM - Center for Marine Environmental Sciences (Bremen) & Alfred Wegener Institute for Polar and Marine Research, Bremerhaven
- Since 1993 - Information system for long-term archiving and publication of data from earth & environmental science
- Large range of different environment related data e.g.
  - Environmental time series
  - Photos, movies
  - Sediment samples
  - Biodiversity
  - many more.....



Total number of data sets ~ 350.000
Data items ~ 10 billions



Environmental Data PANGAEA



Citation: **Wiltshire, Karen Helen (2008):** Biomass of dinoflagellates, dictyochophyceae, and prymnesiophyceae at time series station Helgoland Roads, North Sea, in 1966. *Alfred Wegener Institute, Helmholtz Center for Polar and Marine Research, Bremerhaven*, doi:10.1594/PANGAEA.669759

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

Further details: **Wiltshire, Karen Helen (2007):** Evaluation of plankton data from Helgoland roads. *Alfred Wegener Institute for Polar and Marine Research, Bremerhaven, Germany*, hdl:10013/epic.28563.d001

Project(s): **Shelf Seas Systems Ecology @ AWI (former Biologische Anstalt Helgoland)** (BAH)

Long-term Ecological Research at AWI (LTER)

Coverage: *Latitude:* 54.188330 * *Longitude:* 7.900000

Date/Time Start: 1966-01-03T10:40:00 * *Date/Time End:* 1966-12-30T09:20:00

Minimum DEPTH, water: 0.5 m * *Maximum DEPTH, water:* 0.5 m

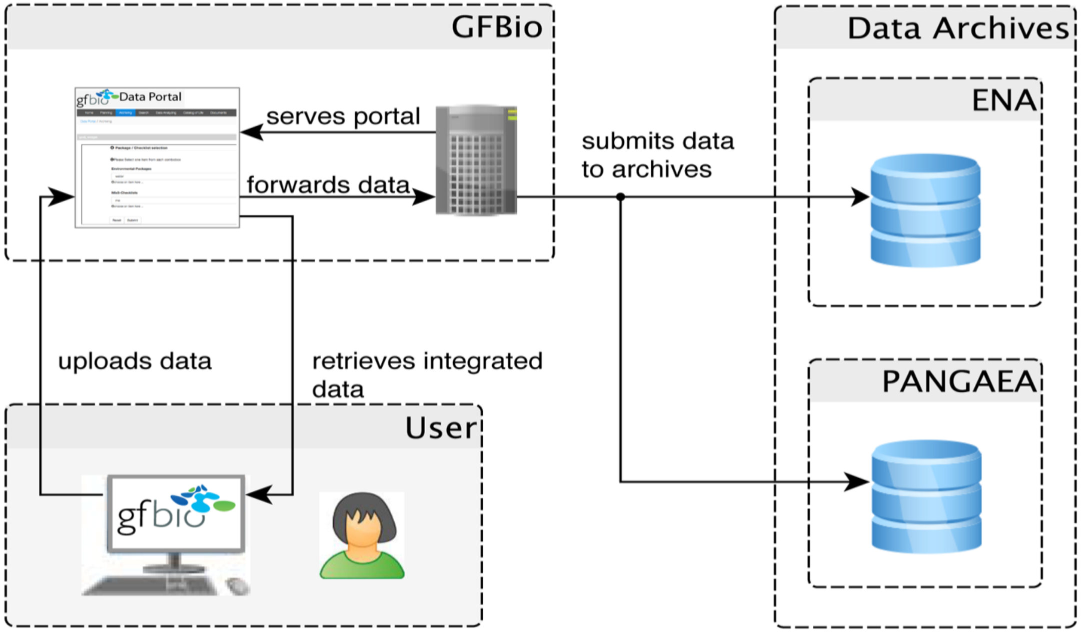
Event(s): **HelgolandRoads** (Kabeltonne) * *Latitude:* 54.188330 * *Longitude:* 7.900000 * *Elevation:* -10.0 m * *Location:* German Bight, North Sea * *Campaign:* HelgolandRoadsTimeseries * *Basis:* Meeresstation Helgoland * *Device:* Monitoring (MON)

Molecular Data Brokerage

What we offer:

- Standardization of molecular metadata according to the MIxS[1] standard
- Manual input and template download/upload
- Linking of persistent identifiers across data centers (ENA + PANGAEA)

Overview of Archiving Workflow for Molecular Data



[1] http://www.gensc.org/mixs

Sustainability

Basic operations/maintenance

Developments

User involvement

Transition

gfbio

“Research” project with 20 partners
project funding

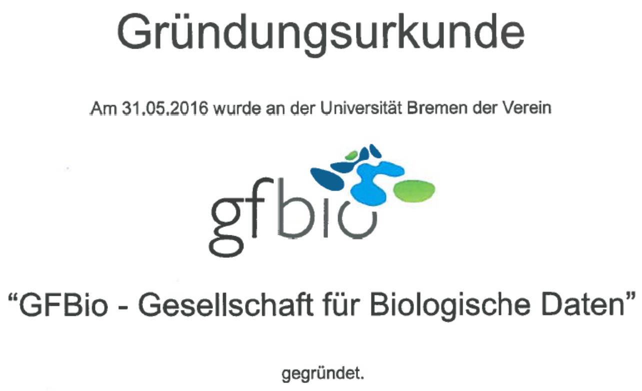
↓

Single legal entity
sustainable business model

gfbio e.V.

GFBio e.V.

- GFBio e.V. is the legal entity
- Founded on 31.05.2016
- 11 founding members (10 persons and GWDG)
  - 1. Chairman: Michael Diepenbroek
  - 2. Chairman: Birgitta König-Ries
  - Treasurer: Frank Oliver Glöckner
  - 1. Assessor: Dagmar Triebel
  - 2. Assessor: Anton Güntsch



Gründungsurkunde

Am 31.05.2016 wurde an der Universität Bremen der Verein

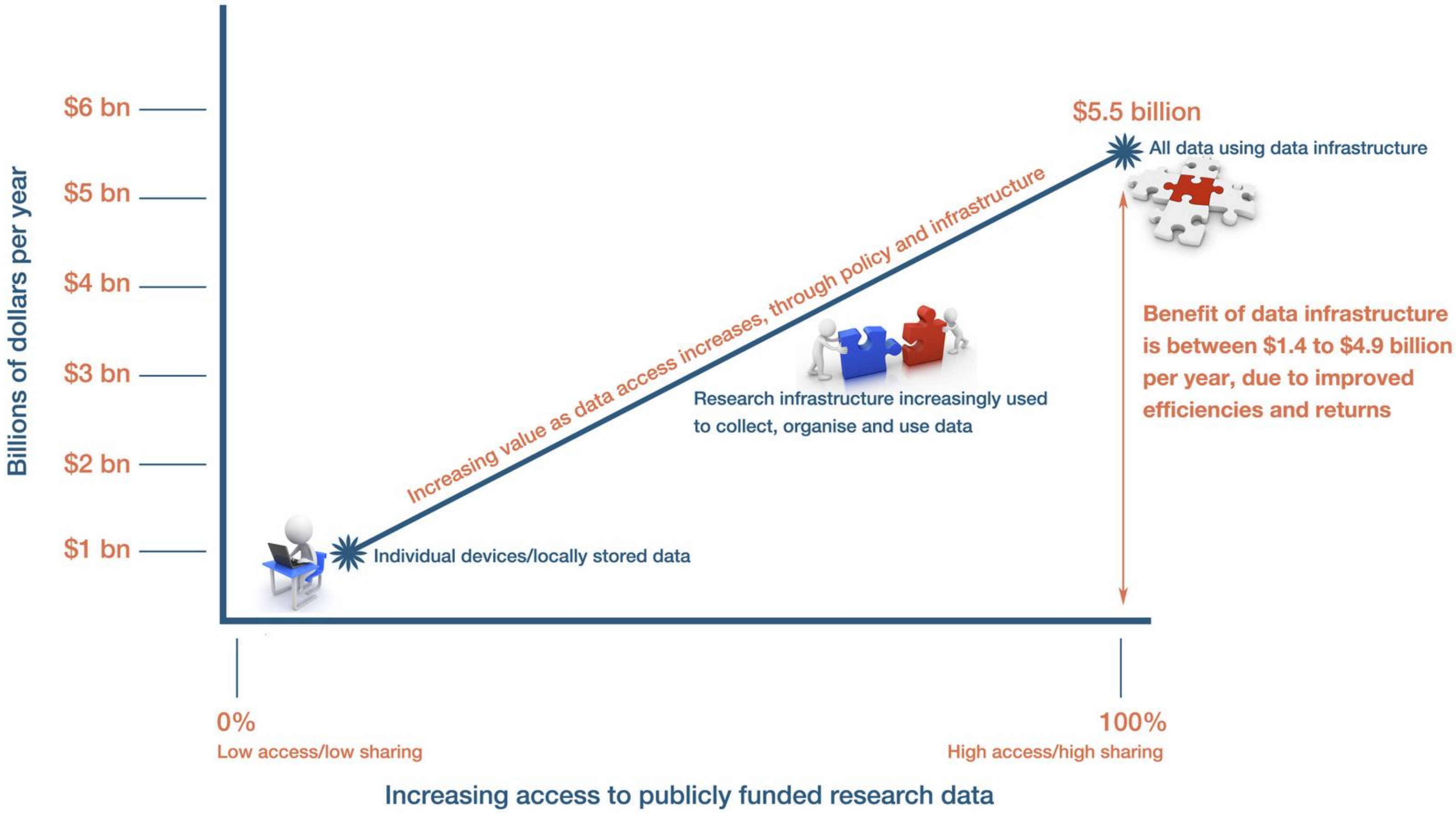
“GFBio - Gesellschaft für Biologische Daten”

gegründet.

The Costs?

Value of Access to Data

Benefits of increasing access to publicly funded research data and increasing use of data infrastructure in Australia



http://www.ands.org.au/working-with-data/articulating-the-value-of-open-data/open-research-data-report

Costs of Data Loss

Data stored on 76.2 million PCs (USA)

Type of loss	Average cost of each data loss incident
Technical service	$ 340
Loss of productivity	$ 217
Value of the lost data	$ 3400
Sub total	$ 3957
Episodes of data loss	4,607,100
Total US data loss costs	$ 18.2 billion = € 17.1 billion

David M. Smith, Graziado Business Review 2003 Volume 6 Issue 3

RDM Costs

6.76 Billion Euro third party funding in 2012

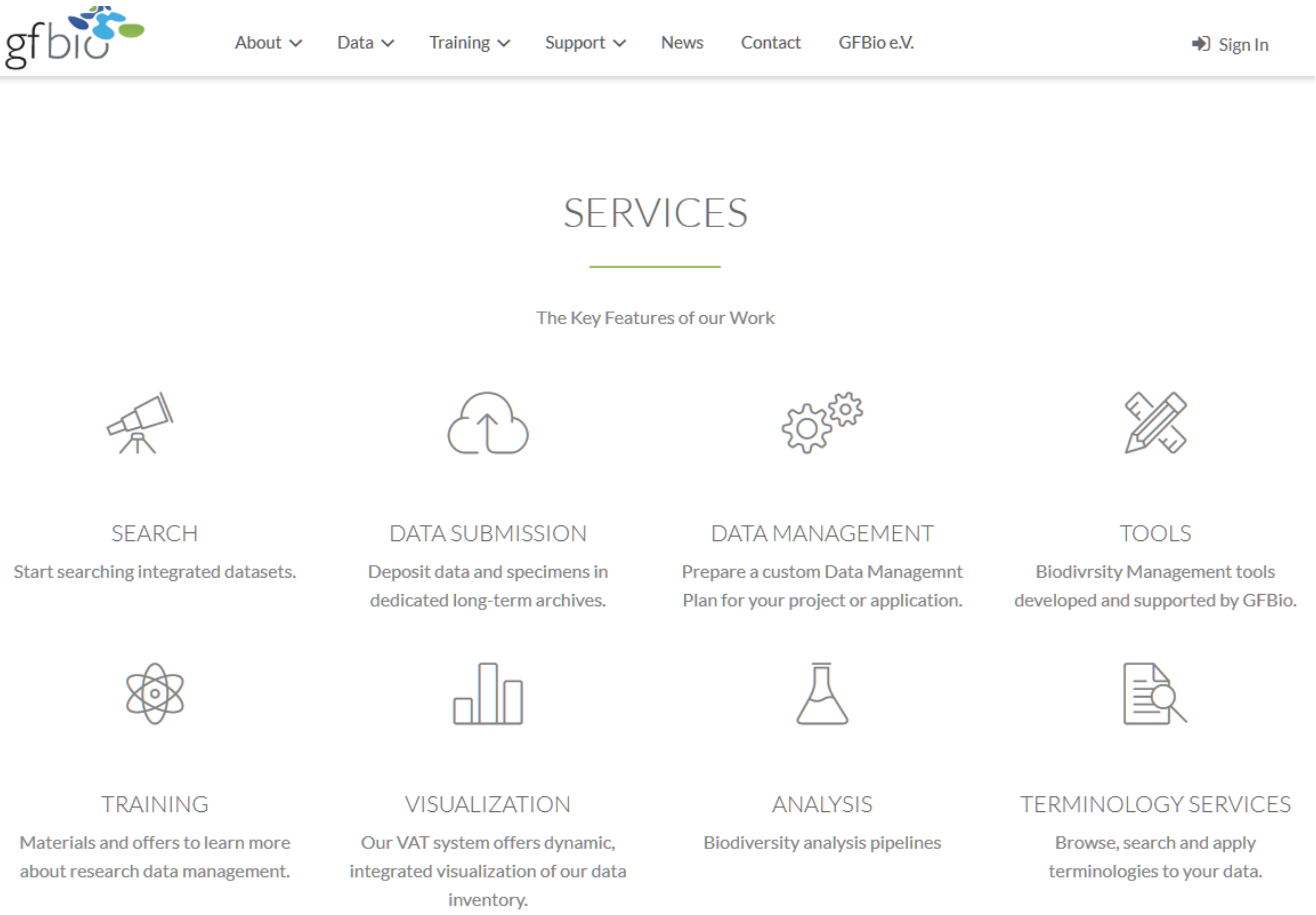
427 Universities in Germany

5-15% is needed for Research Data Management

338 – 1014 Million Euro

http://www.dfg.de/service/presse/pressemitteilungen/2015/pressemitteilung_nr_43

Contact & Services



SERVICES

The Key Features of our Work

SEARCH

Start searching integrated datasets.

DATA SUBMISSION

Deposit data and specimens in dedicated long-term archives.

DATA MANAGEMENT

Prepare a custom Data Managemnt Plan for your project or application.

TOOLS

Biodivrsity Management tools developed and supported by GFBio.

TRAINING

Materials and offers to learn more about research data management.

VISUALIZATION

Our VAT system offers dynamic, integrated visualization of our data inventory.

ANALYSIS

Biodiversity analysis pipelines

TERMINOLOGY SERVICES

Browse, search and apply terminologies to your data.

www.gfbio.org

info@gfbio.org

Thanks to...

GFBio

GFBio e.V.

DFG Deutsche Forschungsgemeinschaft

Thanks for your attention

gfbio

GERMAN FEDERATION
FOR BIOLOGICAL DATA