NFDI 4
BIODIVERSITY

Daten und Datenqualität als Basis der Nationalen Forschungs-dateninfrastruktur NFDI

Frank Oliver Glöckner, 16.11.2021

Research Data

Research data is data created in the course of scientific activity, e. g. through observations, experiments, simulations, surveys, interviews, the study of sources, records, digitisation, or evaluations

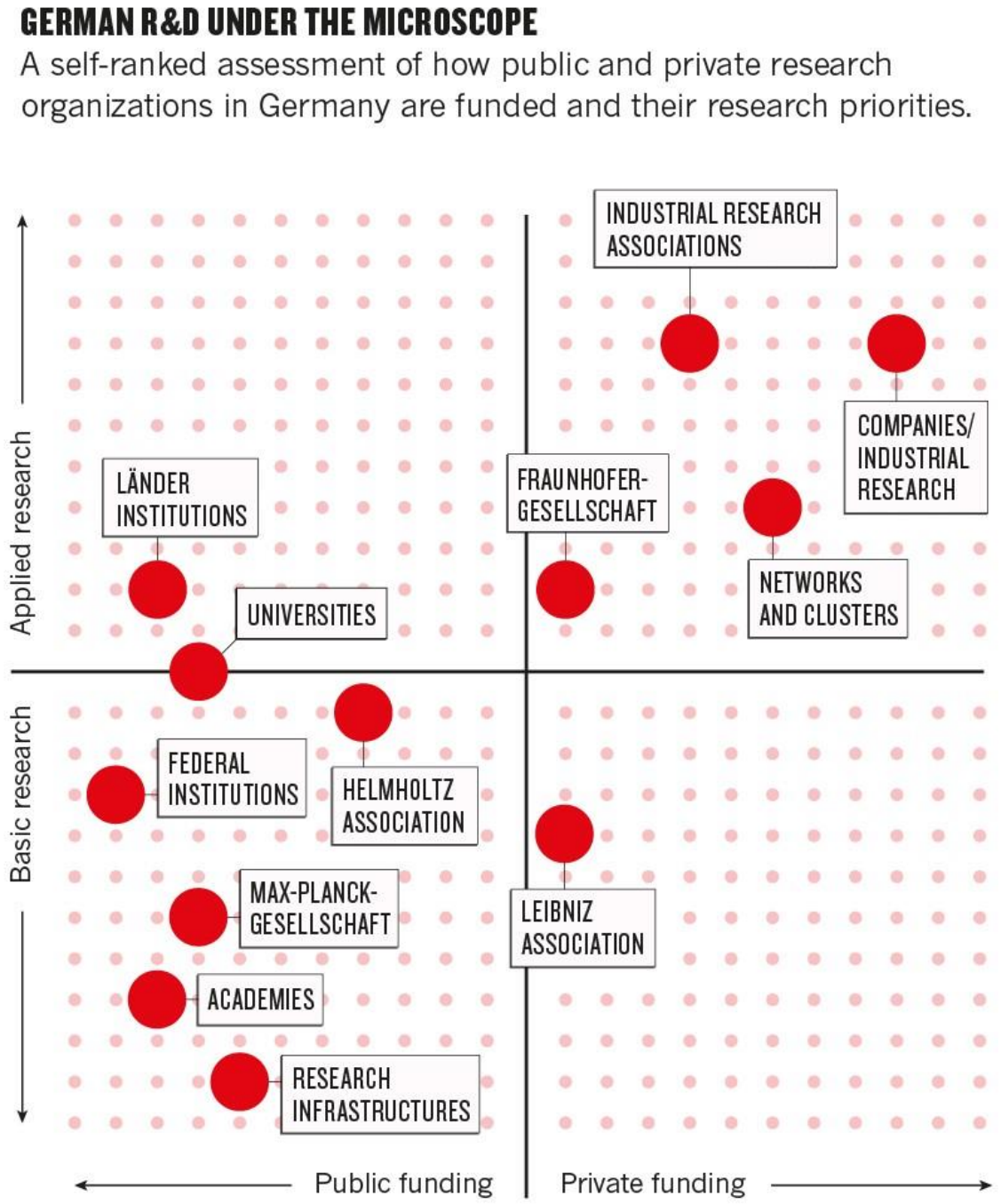
In actual research, one differentiates, although not always clearly, between primary research data and secondary research data, which documents and contextualises the process of creating primary data.

Rat für Informationsinfrastrukturen, 2016

Origin of data

- publicly funded for example …
  - … at over 400 universities and technical colleges
  - … at the non-university research institutions and the scientific academies
  - … in large, jointly funded research infrastructures

=> Area of activity of the Joint Science Conference of the federal government and federal states (Article 91b of the Basic Law)



https://www.nature.com/articles/d41586-019-00910-7

Source: Federal Ministry of Education and Research

NFDI

Nationale Forschungsdaten Infrastruktur

Accepted community-driven federated initiatives in DE lack sustainability

Some examples:

- Since 2003: The German Astrophysical Virtu… Observato… (GAVO) connects German research projects to the… O.
- Since 2004: The German Data Forum RatSWD… …ps a network of research data centres at public au…orities an… scientific institutions
- Since 2014: German Federation for Biological Da… …is a network to host data from research projects on biodiversit…
- Since 2015: The German Network for Bioinformatic… Infrastructure (de.NBI)

Funding ended 2017

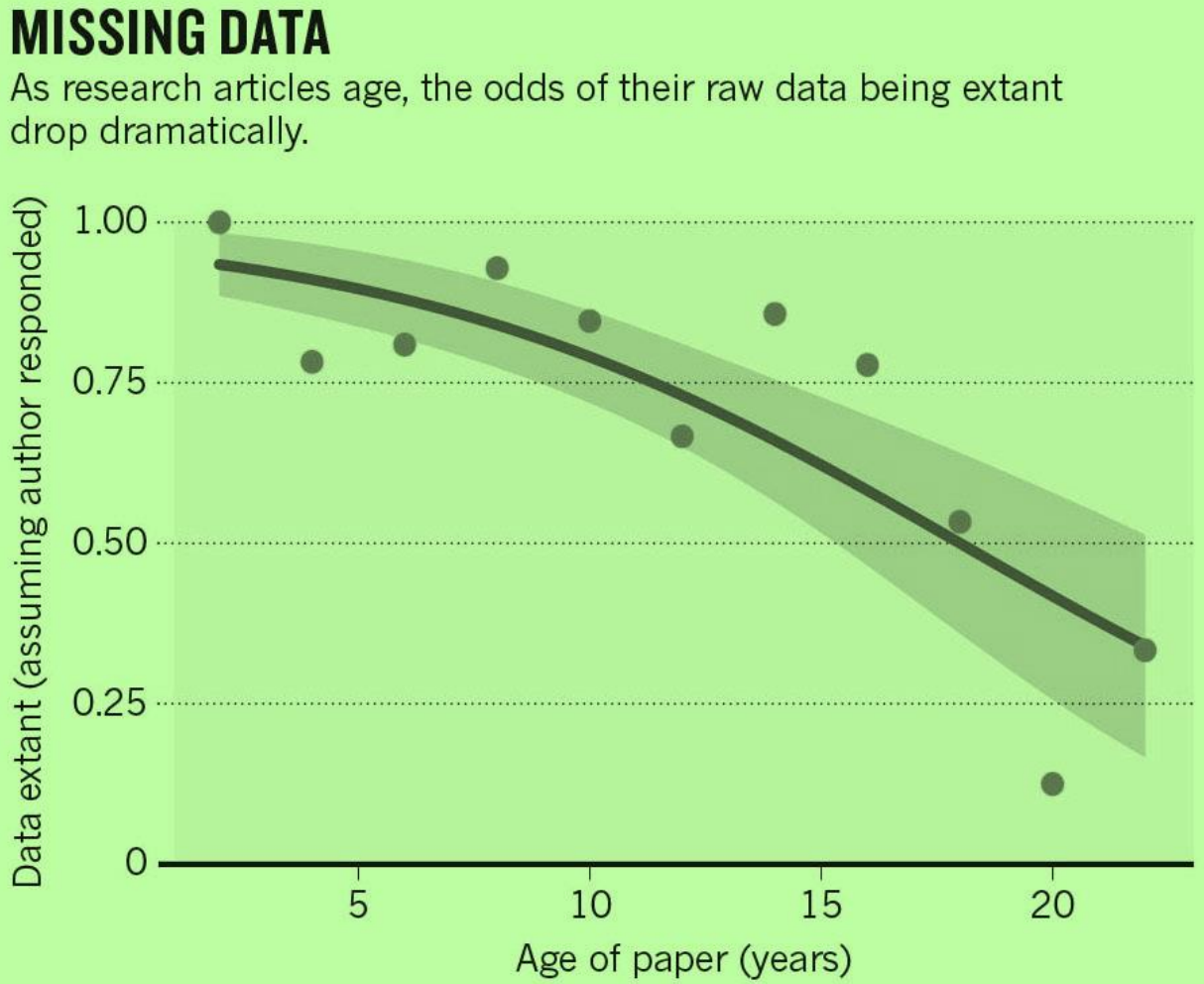
Funding ended 2020

Funding ended 2021

Funding ends 2021

Availability of Research Data with Time

- 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

Make or Buy?

- Data as a special form of scientific knowledge - several large scientific publishers are expanding their spectrum
  - but: the relationship of trust is currently damaged
- Data and scientific methods are closely linked.
  - there is a lot to favor internal scientific solutions

Acknowledging the need for coordinated action

The rising tide of data – nearly as many digital bits as there are stars in the universe.

Source: IDC's Digital Universe Study 2014
https://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm

The road towards NFDI

2011-2015: Structured exploration
Basic recommendations on how to adapt the organisation of scientific information infrastructures in the digital turn

KII 2011

WR 2012

RfII 2014

2016-2019 Realisation phase
Joint Science Minister Conference adapts RfII recommendation; agrees on investments of up to 90 Mio. € p.a.
DFG is tasked with the funding process

2016

GWK 2016-2018

DFG 2019

2020

2021

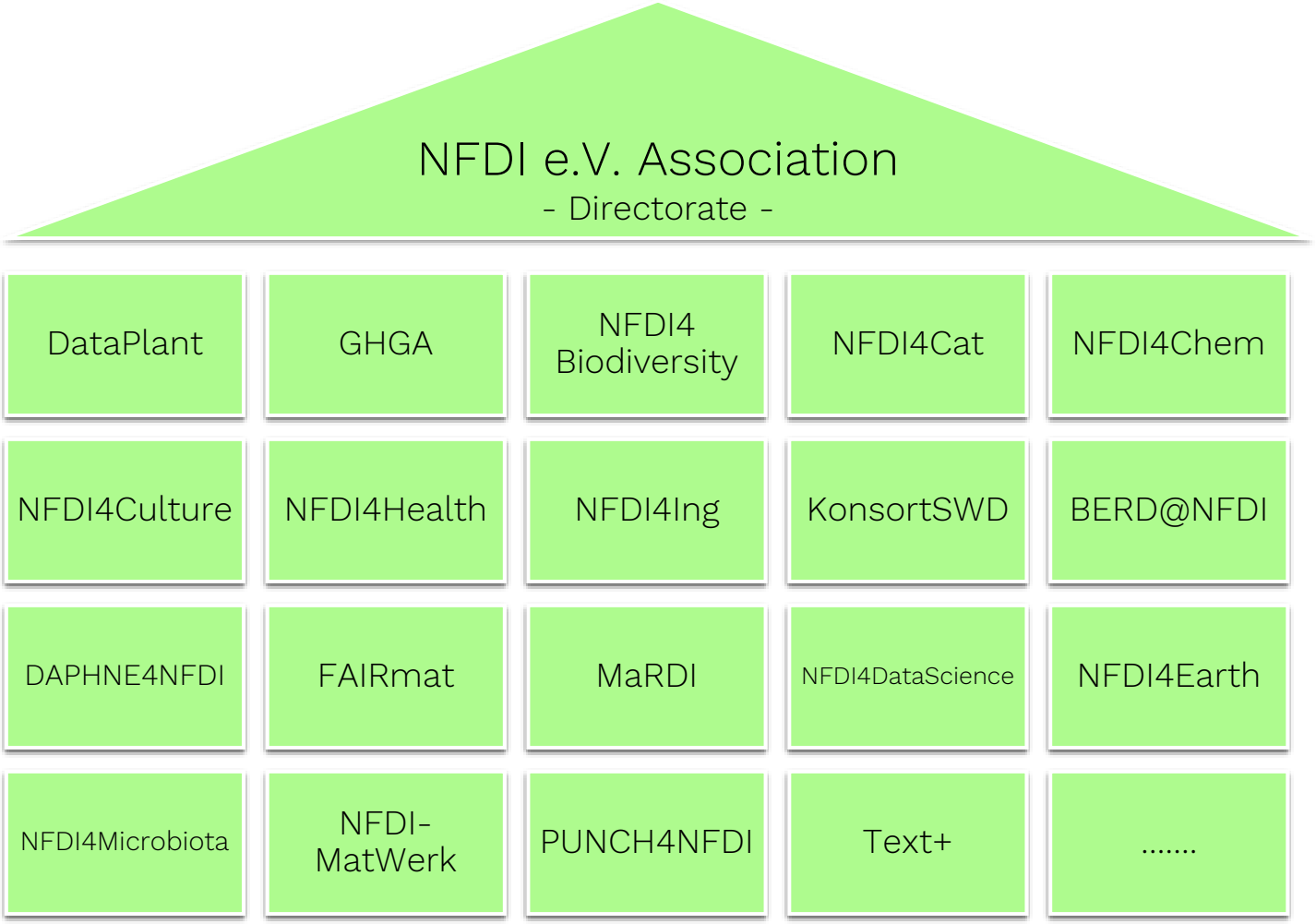
2022

2028ff.?

Operational phase of 10 years
3 rounds of funding decisions
Founding of the NFDI e.V.

Parallel initiatives on the international level

Structure of the NFDI



From January 2023 ff.

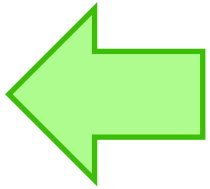
- Up to 30 Consortia, covering all research domains in Germany
- Several hundred organisations teaming up as suppliers and in decision processes
- > 1,000 professional staff

A Germany-wide alliance for community-oriented data services

Nationale Forschungsdaten Infrastruktur

- NFDI Association (e.V.)
  - Founded October 2020
  - November 2021: 193 member organisations have joined

- High-priority cross-cutting topics (started 2021)
  - Common infrastructure & interoperability
  - Metadata & terminologies & provenience
  - Training & education
  - Ethical & legal & social aspects



Association members join forces and create synergies in NFDI „sections“

Workshop report cross-cutting topics: https://doi.org/10.5281/zenodo.4593770

NFDI4Biodiversity

as an example

Our Motivation

“We are convinced that actors from science, politics and nature conservation need reliable data to make better contributions to the preservation of global biodiversity”

Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES)

Report published on May 6, 2019:

1 Million species threatened with extinction

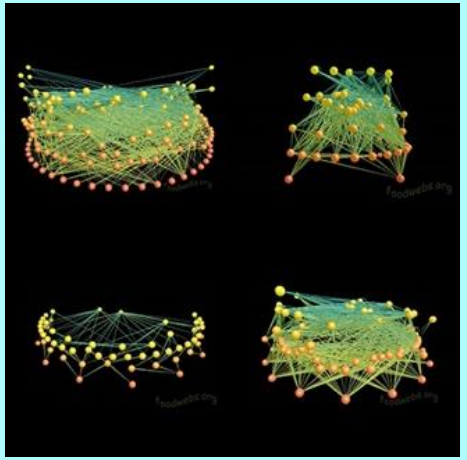
www.ipbes.net

Facets of Biodiversity

species

genes

functions



interactions

ecosystems

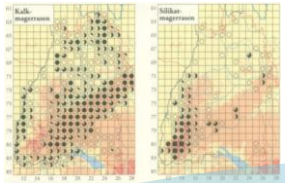
Very specific data types – often combined with geospatial data
Time series
Reference lists and taxonomies

The future - 2021ff: Data availability beyond research




occurrence data

research institutes, associations, foundations, citizen science



distributions

state & federal agencies

ecosystem data

national parks

experiments

research institutes

NFDI 4 BIODIVERSITY

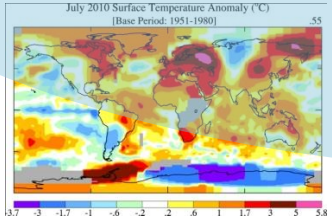
BIODIVERSITY, ECOLOGY & ENVIRONMENTAL DATA

functional traits

research institutes

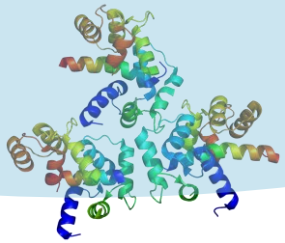
satellite data

research institutes & agencies



model output

research institutes



molecular data

research institutes

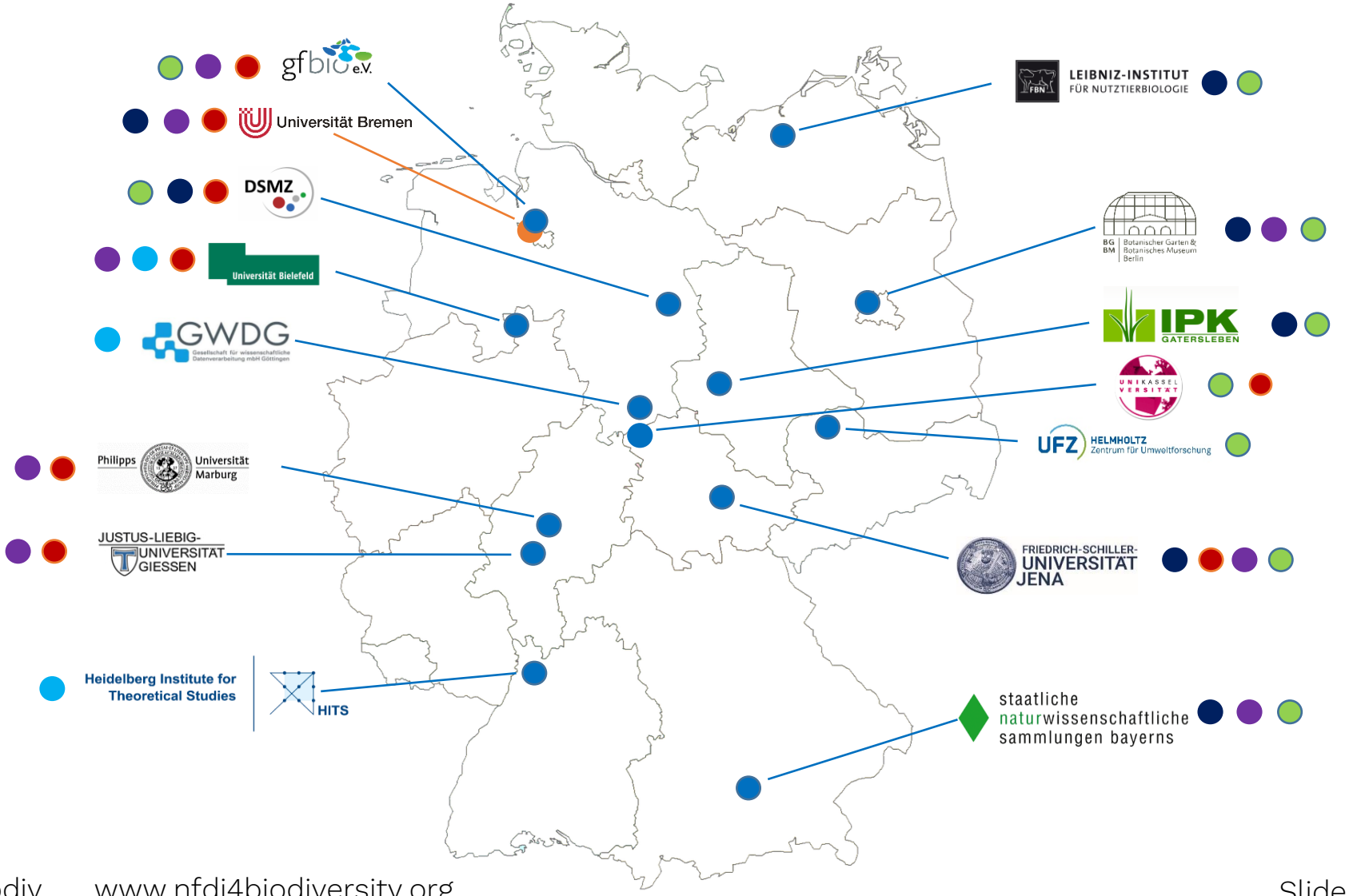
collection data

museums & collections

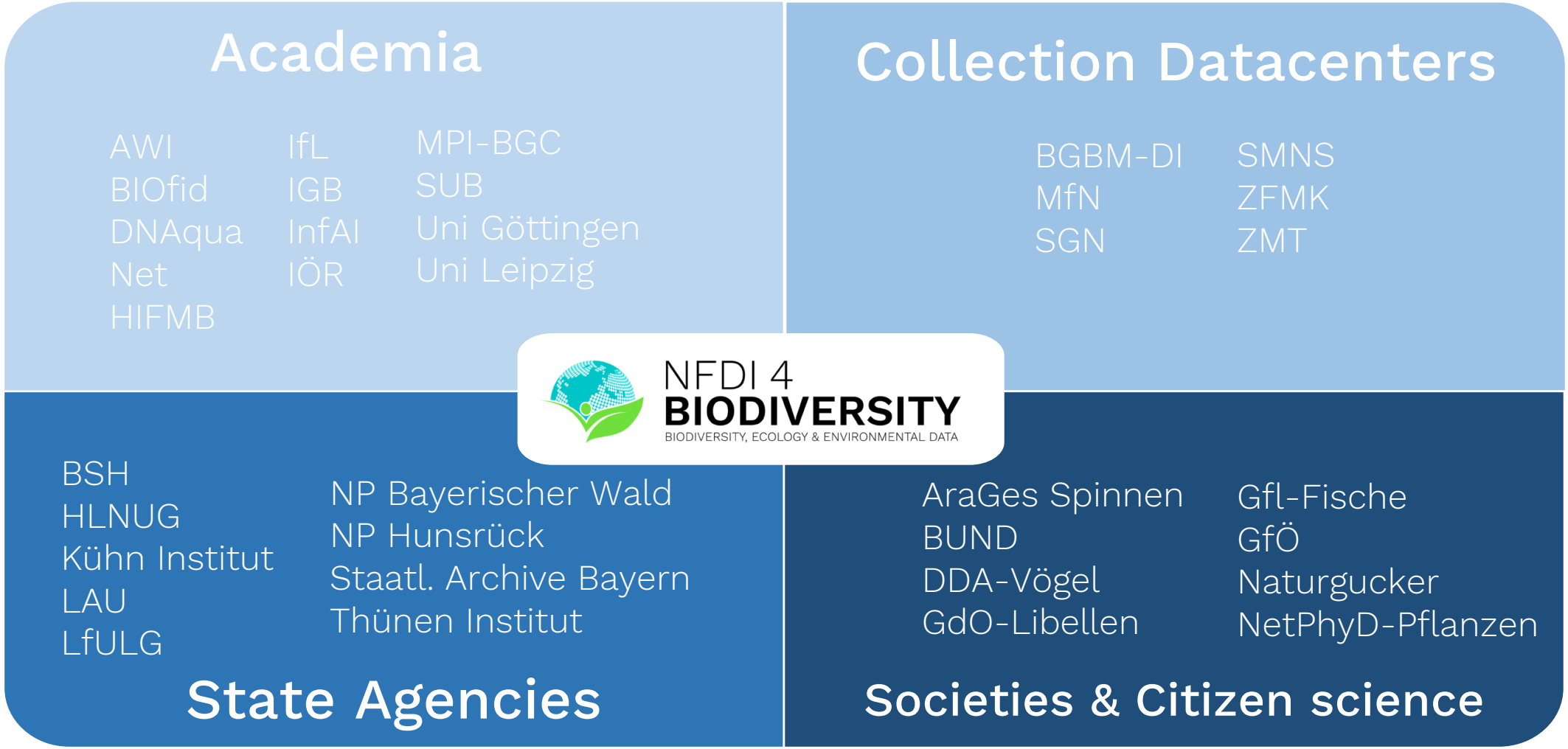
Adapted from Christian Wirth, iDiv

The NFDI4Biodiversity Consortium: 15 Co-Applicants

- Data Center
- Infrastructure Provider
- Computer Science
- Biology/Env. Sciences
- Teaching/Training



The NFDI4Biodiversity Consortium: >30 Participants



NFDI4Biodiversity Use Case Projects

NFDI 4 BIODIVERSITY
BIODIVERSITY, ECOLOGY & ENVIRONMENTAL DATA

01 eLTER	02 Plants - IPK	03 AMMOD Hub	04 DE Barcode of Life	05 Marine Life	06 Land use/cover
07 DNAquaNet	08 Naturgucker	09 Insekten Sachsen	10 NFDI4Agri	11 MultiBase, LA	12 NP Bayer. Wald
13 Multibase, LA (2)	14 AlgaTerra	15 GdO - Libellen	16 Ara-Ges - Spinnen	17 NetPhyD	18 GfI - Fische
19 Viz Tools	20 NP Hunsrück	21 MultiBase, LA (3)	22 iDiv PlantHub	23 ALA-DE	(...)

AMMOD use case

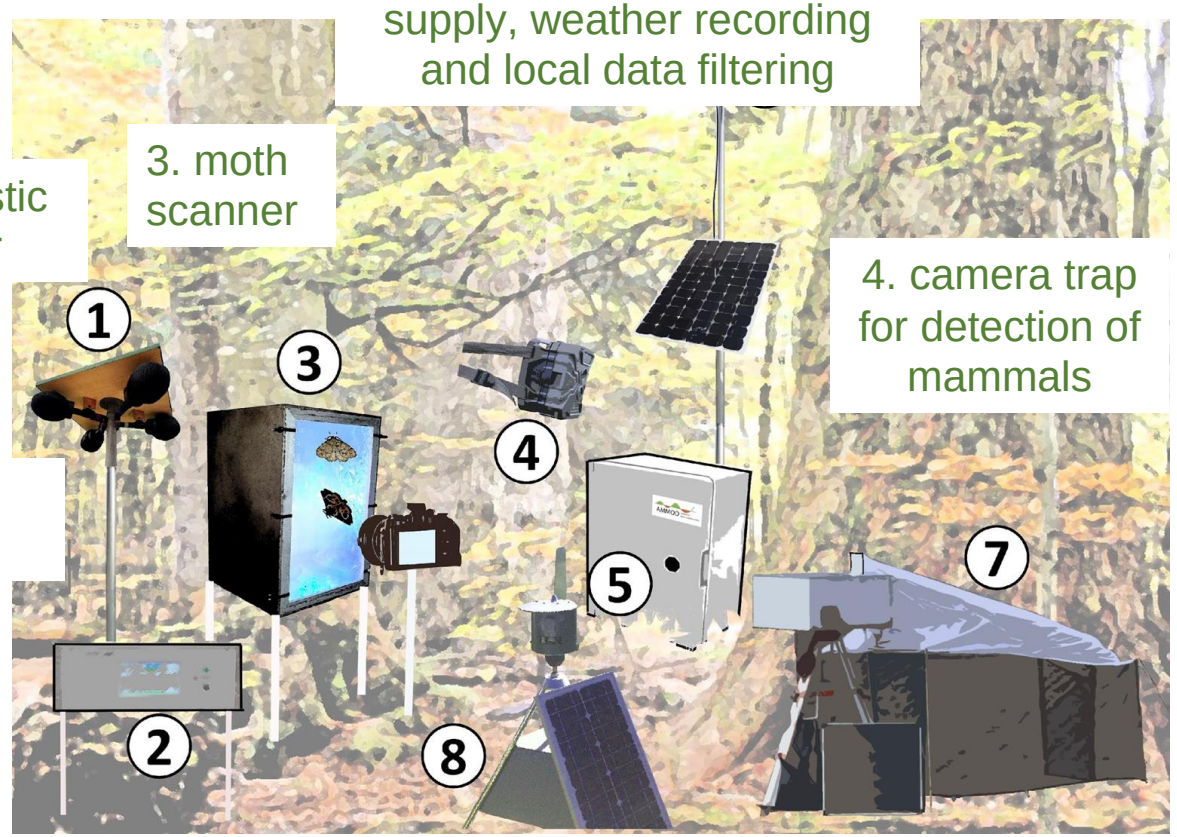
5. base station for energy supply, weather recording and local data filtering

1. acoustic recorder

3. moth scanner

4. camera trap for detection of mammals

2. detector for volatile organic compounds



8. autonomous pollen sampler

7. automated Malaise trap for monitoring of insects

AMMOD Automated Biodiversity Monitoring

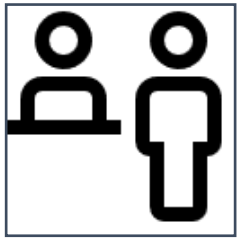
Services

Services in NFDI4Biodiversity

With our experienced partners, we offer access to tried-and-tested tools for handling biodiversity and environmental data.

Access to relevant data services from the domain

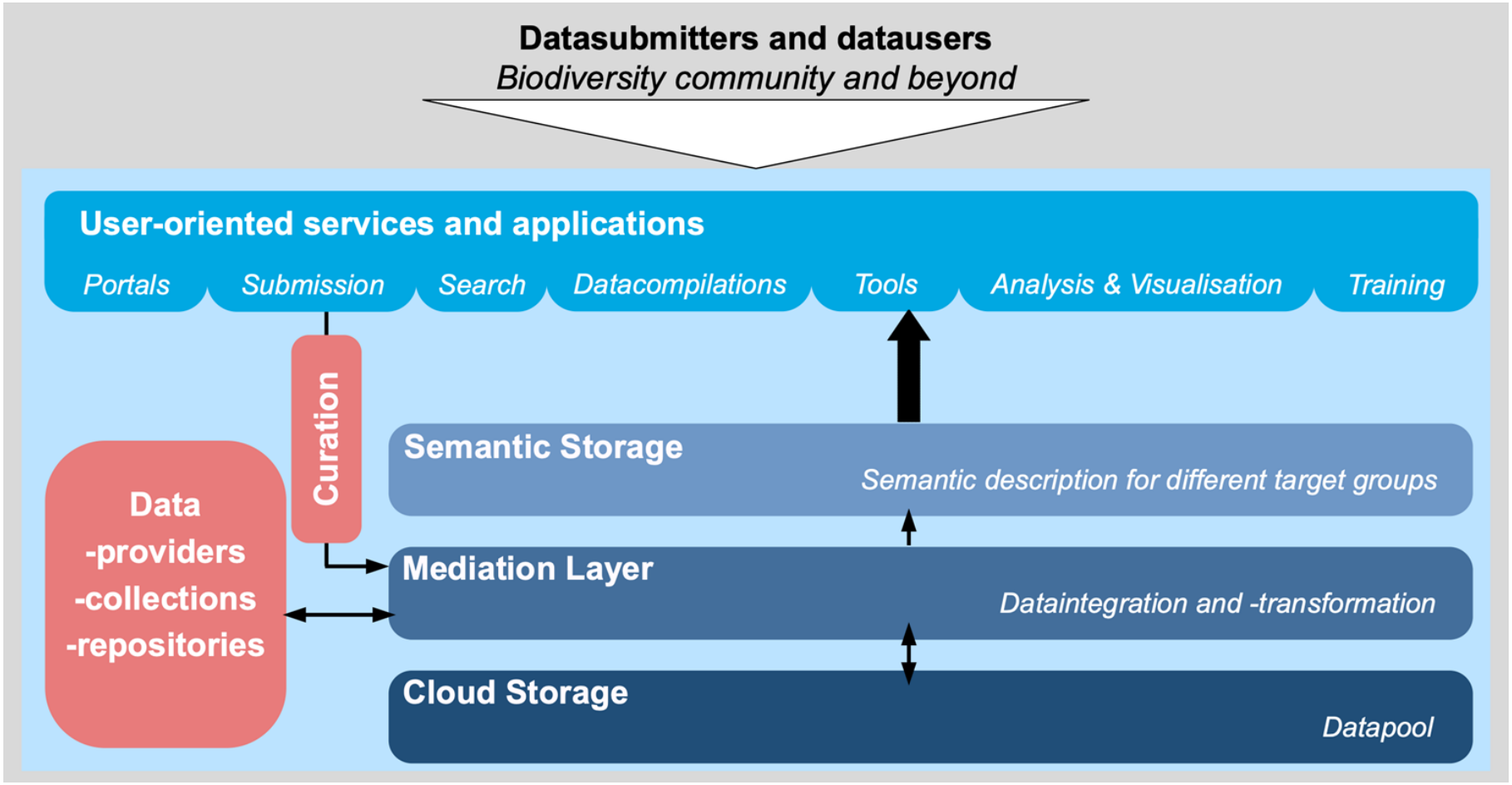
Front Office/Back Office Support for Research Data Management



Education and Training

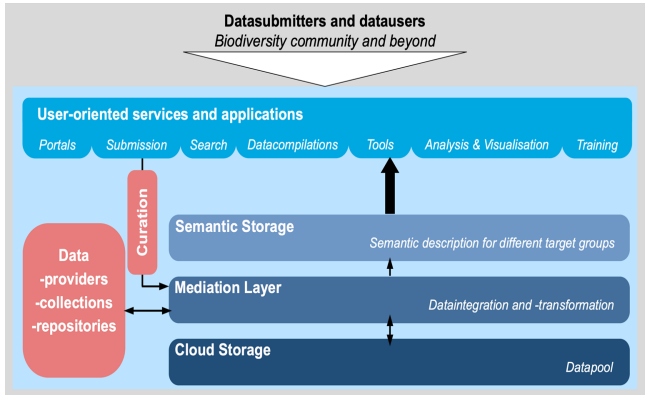
Common Infrastructure (Research Data Commons)

Next level technology: Moving services to the academic cloud



Added value from the perspective of a data user

Consolidated workflows of participating data providers



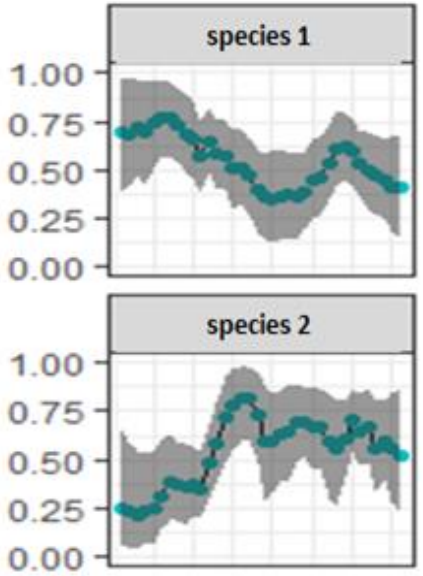
Novel statistical methods for heterogeneous data


Methods in Ecology and Evolution

Statistics for citizen science: extracting signals of change from noisy ecological data

Spatial data for environment and site characteristics

Infrastructure



Data Quality

Data Quality – My Definitions

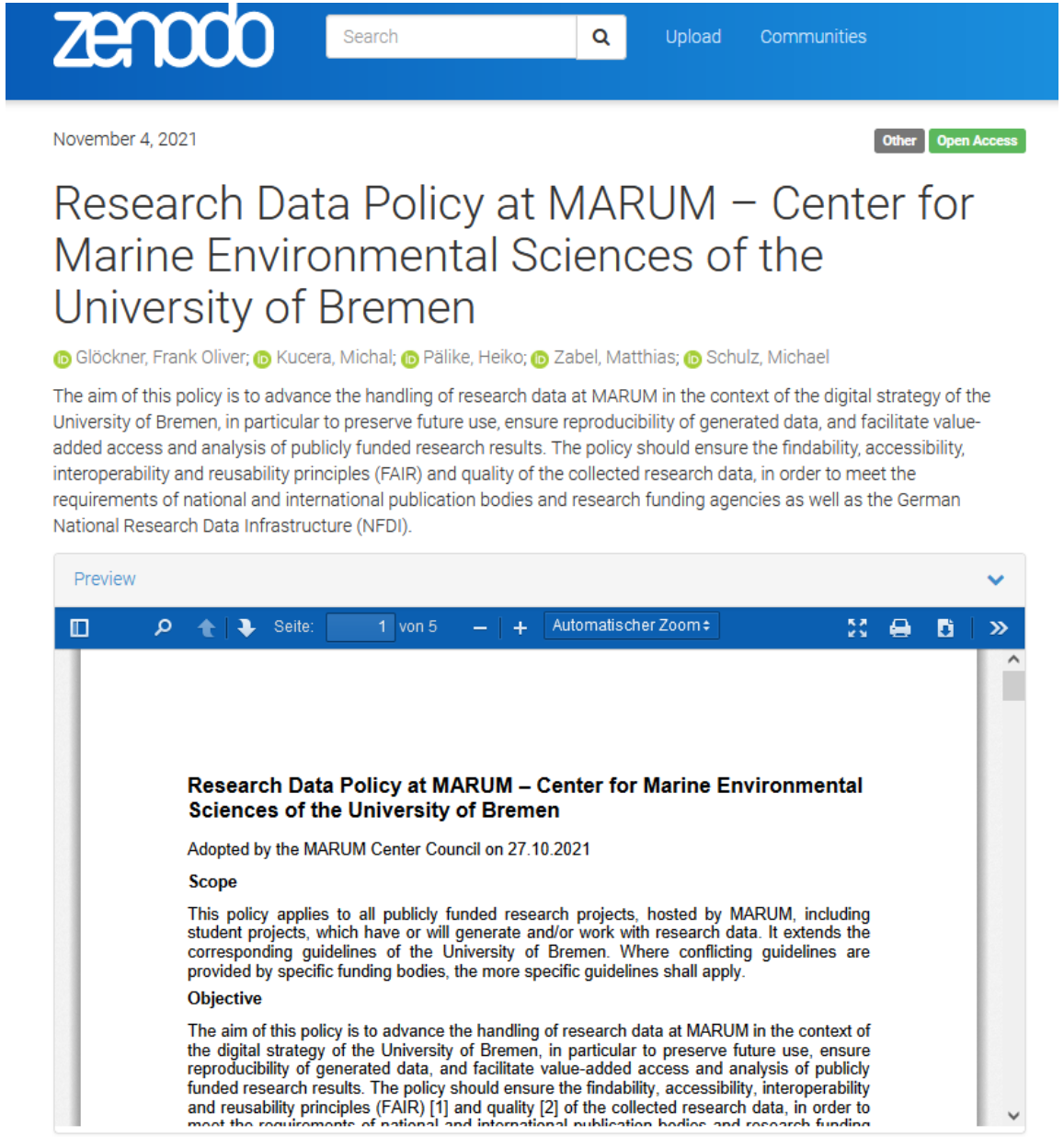
- Data Quality -> fit for use/fit for purpose
- Quality Standards
  - Metadata Standards
  - Methods Standards
  - Data Integrity (Storage/Archiving)
  - Data Curation
  - Provenance
  - Data Policy
  - Certification

FAIR Data Principles

The Data Quality Challenge, RfII 2020

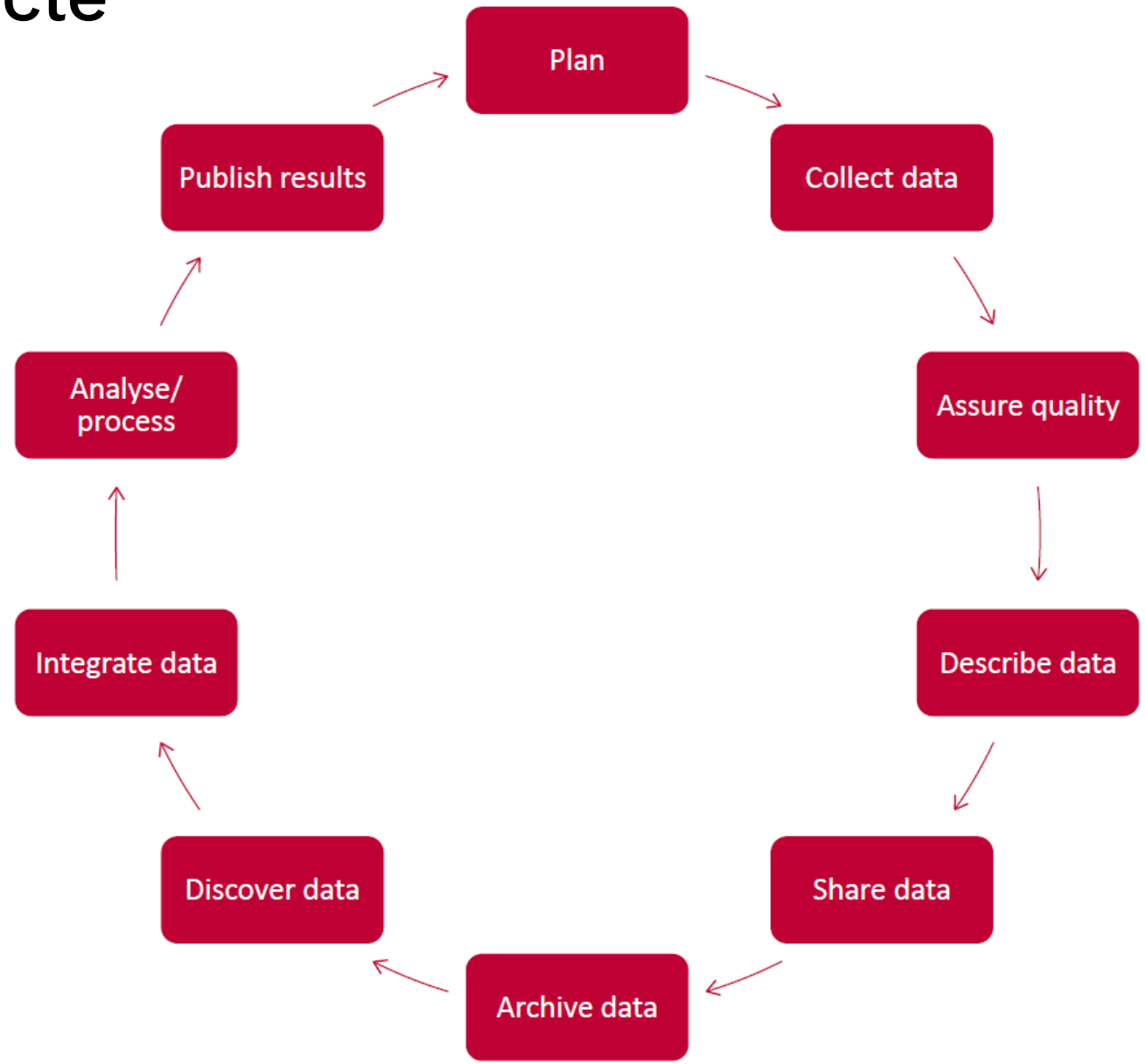
TO BE FINDABLE:	
F 1	(meta)data are assigned a globally unique and eternally persistent identifier.
F 2	data are described with rich metadata.
F 3	(meta)data are registered or indesed in a searchable resource.
F 4	metadata specify the data identifier.
TO BE ACCESSIBLE:	
A 1	(meta)data are retrievable by their identifier using a standardized communications protocol.
A 1.1	the protocol ist open, free and universally implementable.
A 1.2	the protocol allows for an authentication and authorization procedure, where necessary.
A 2	metadata are accessible, even when the data are no longer available.
TO BE INTEROPERABLE:	
I 1	(meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I 2	(meta)data use vocabularies that follow FAIR principles.
I 3	(meta)data include qualified references to other (meta)data.
TO BE RE-USABLE:	
R 1	meta(data) have a plurality of accurate and relevant attributes.
R 1.1	(meta)data are released with a clear and accessible data usage license.
R 1.2	(meta)data are associated with their provenance.
R 1.3	(meta)data meet domain-relevant standards.

Data Policy



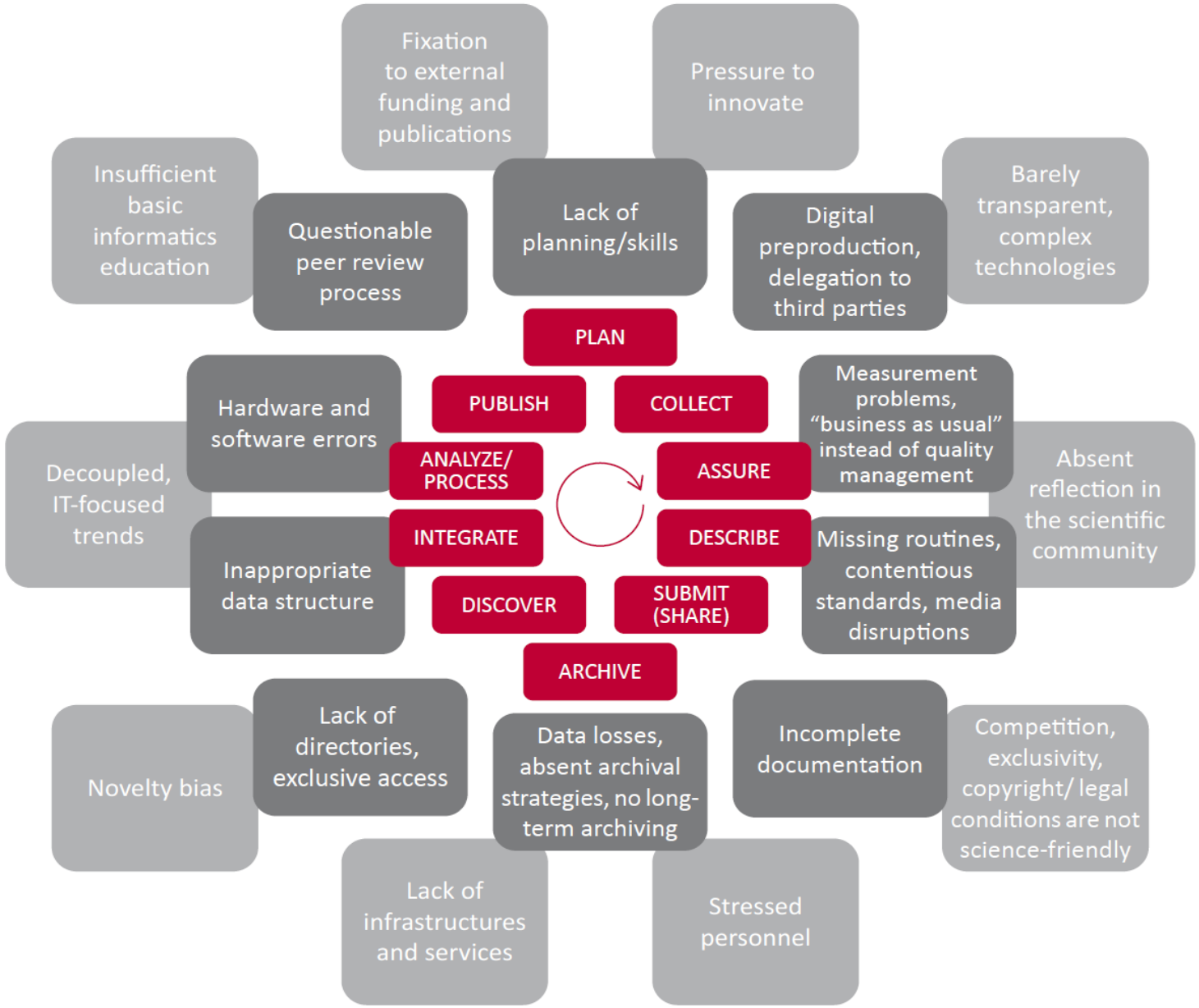
https://doi.org/10.5281/zenodo.5643724

Data Life Cycle



The Data Quality Challenge, RfII 2020

Constrained Data Life Cycle



The Data Quality Challenge, RfII 2020

Data goes Digital

The Data Quality Challenge, RfII 2020

- Disproportionately large impact of small inattentions, errors and failures
- Decisions on the usability of noisy mass data
- Decontextualized uses of individual data sequences
- Unclear or unrecognizable provenance of data (especially in the case of algorithm-generated outcomes and selections)
- Non-transparent computational processes
- Misdirecting algorithms (e.g. due to scaling problems)
- Lack of training sets for the programming of machine learning (AI)
- Complicated or impossible verification/validation of the practical value of oversized data batches
- Division of labour along by now only weakly integrated process chains (so-called “pipelines”) while working with scenarios or doing simulations
- Growing dependence of knowledge work on proprietary software
- Lack of archivability of the digital artefact
- Presentation problems for the result dimension of complex computations and data condensation (e.g. through “visualization”)
- Data protection and other legal issues
- Low-threshold manipulation possibilities
- Hacking and cyber espionage
- Targeted data sabotage
- … and more.

Digital Long Term Archiving / Publication

- Period?
- What?
- Documentation?
- Technology?
- Funding?
- Quality Management?
- Competences?

https://www.archive360.com/blog/ultra-long-term-archiving-and-the-cloud-its-a-good-thing


12
In less time

NFDI 4
BIODIVERSITY

Many thanks for your interest

@NFDI4Biodiv #NFDI4Biodiv

www.nfdi4biodiversity.org