Project NFDI4Earth

Erwartungen der Erdsystemwissenschaften an die Langzeitarchivierung und Wiedernutzbarmachung archivierter Daten

Wolfgang zu Castell
(Director Department Geoinformation & CIO)

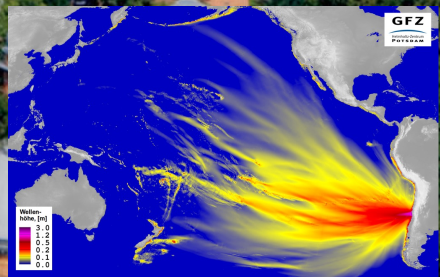
Introducing German Research Center for Geosciences

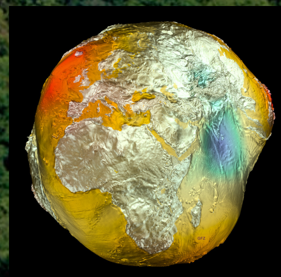
Helmholtz Association
- 43k employees in 18 centers
- addressing „big" questions
- interdisciplinary research

Changing Earth –
Sustaining our Future

(7 centers)

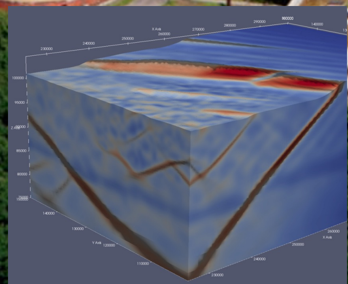
GFZ Potsdam
- 1.3k employees in 5 depts
- focussing on solid Earth
- geodesy, processes, risks

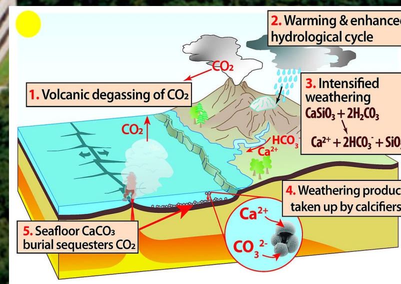




important components
- observatories, satellites, monitoring networks
- geoscience labs
- modelling

Data Science
- maths, physics, comp. sci.
- modelling & data analysis
- numerics, ML/AI, complexity





- Library & Information Services
- eScience Center
- ICT Services
- Data Science Center

Outline

- NFDI4Earth – one of the 27 consortia forming the National Research Data Infrastructure[1]
- Research in Gesosciences – Challenges towards letting data flow freely
- Research Data Management and Digital Preservation – Friends or Foes?

[1] slides by the NFDI4Earth Consortium

National Research Data Infrastructure for Earth System Sciences

NFDI4Earth Consortium 2022

https://www.nfdi4earth.de

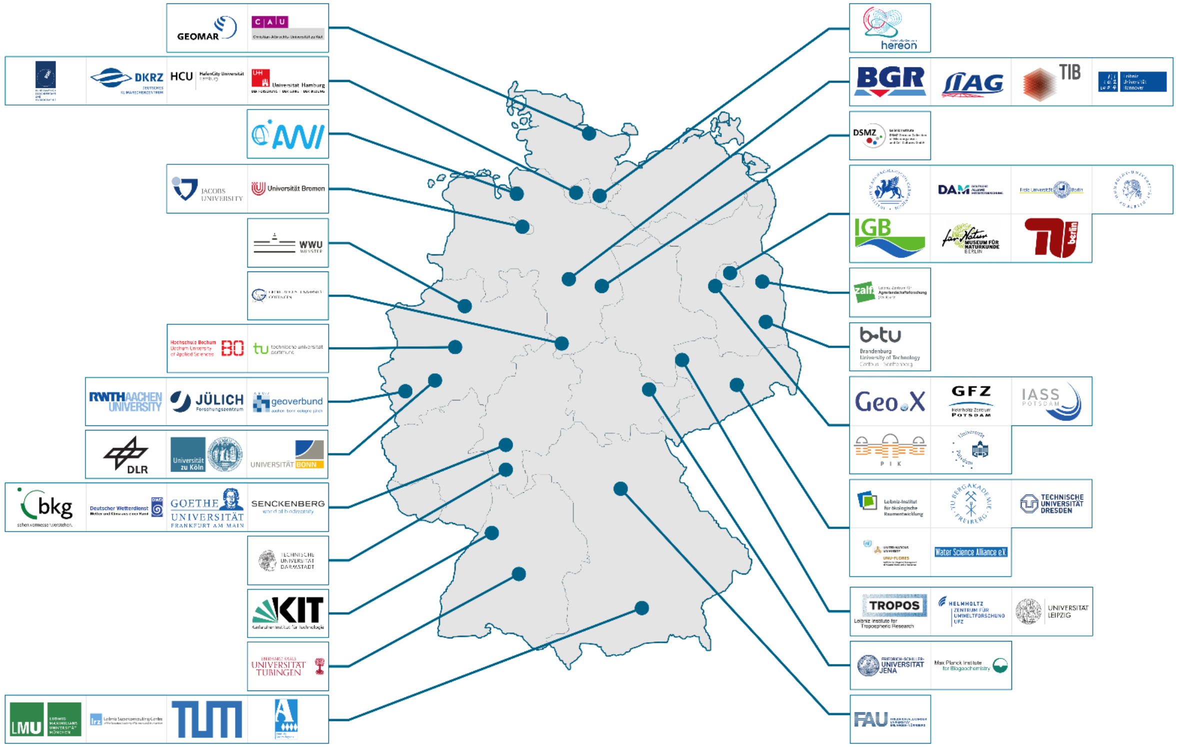
Earth System Sciences (ESS)

- Geosphere, Atmosphere, Biosphere, Hydrosphere, Cryosphere, Anthroposphere and all their interactions
- From Local Processes to Global Challenges
  e.g. earth systems dynamics, natural hazards, climate change, environmental pollution, water scarcity, land use change, scarcity of raw materials
- Observing, measuring, modelling, analyzing, predicting the Earth System
- In international and interdisciplinary settings using spatio-temporal data as the common reference

Who we are

NFDI4Earth

- 59 institutions
  covering the breadth of
  Earth System Sciences (ESS)
  - Universities
  - Research Organizations
  - Infrastructure Providers
  - Governmental Institutions
  - Scientific Associations
    & Networks



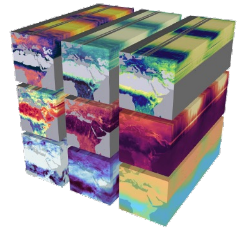
- Established 2018 as
  Open Consortium and the ESS branch in the NFDI
  DFG-Funded since 10/2021

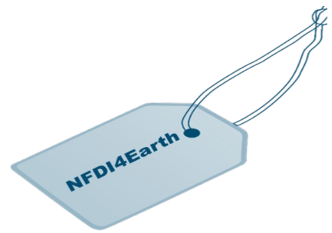
Our Issues

- Large, diverse number of data services (>> 100) and activities related to ESS...only a few sustainable
- Data heterogeneity
  quality, curation levels, licenses, semantics, sizes, repeatability, scale, spatio-temporal resolutions,...
- Different data cultures and velocities to FAIR
  from data streams to long tail data; RDM experts and novices
- Incomplete support along the data lifecycle
- Lack of support (platforms, tools) for joint and collaborative interpretation of heterogeneous and decentralized data

Key Goals

- *One Community* approach to sustainable, Open and FAIR Research Data Management in Earth System Sciences (ESS)
- Community driven agile development of innovative platforms for data integration and collaborative data analysis
- Qualification for people, data, tools and services as a basis for FAIR RDM and viability
- *OneStop4All* and *User Support Network* for ESS RDM as integral part of NFDI and International Infrastructures
- Key-driver of the build-up and operation of the NFDI





NFDI4Earth – Strategy 2021-26

- **2Participate** – Task Area 1
- **2Facilitate** – Task Area 2
- **2Interoperate** – Task Area 3
- **2Coordinate** – Task Area 4

Community Engagement

Piloting Innovations

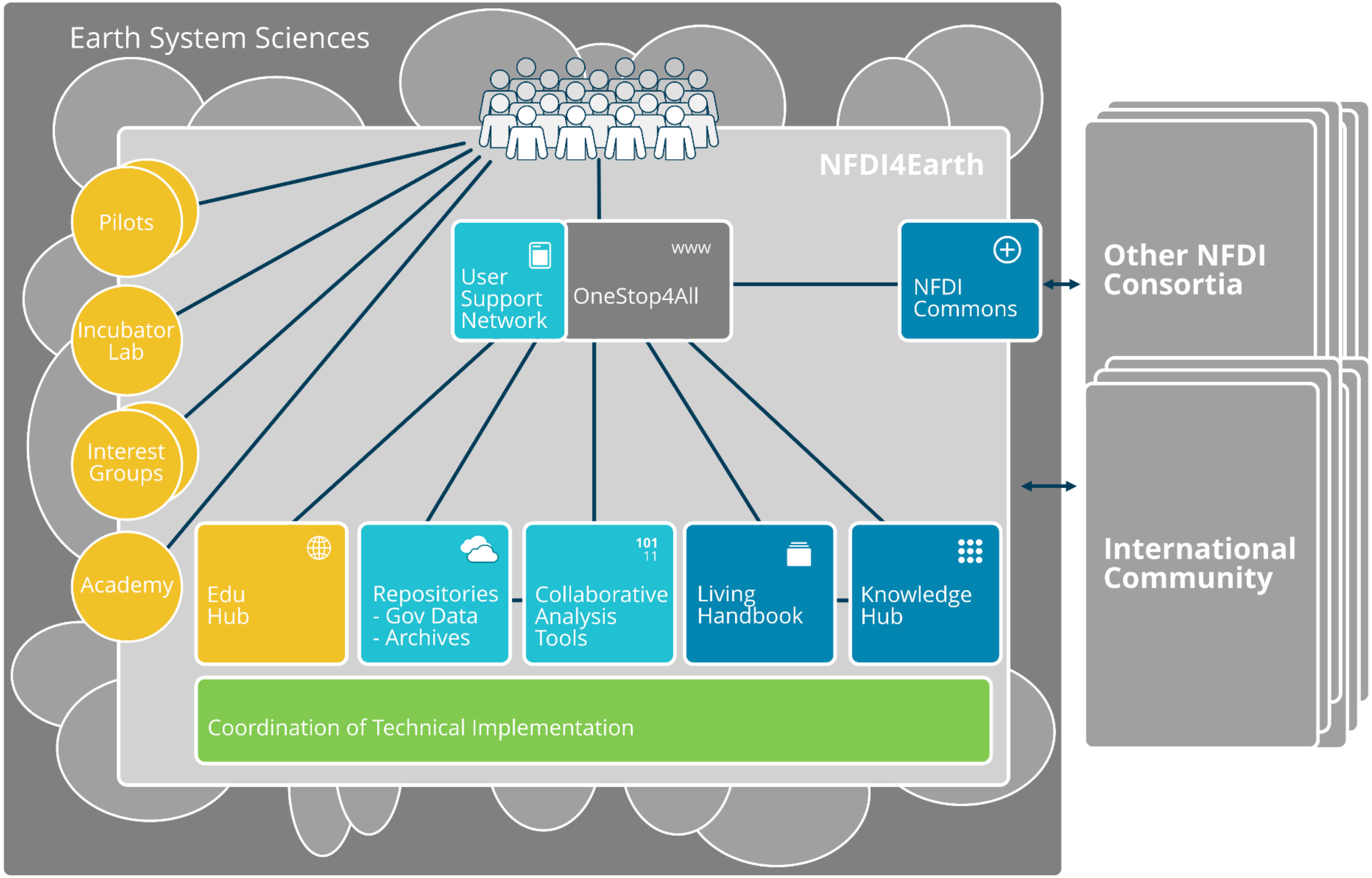
User Perspective

Infrastructure Perspective

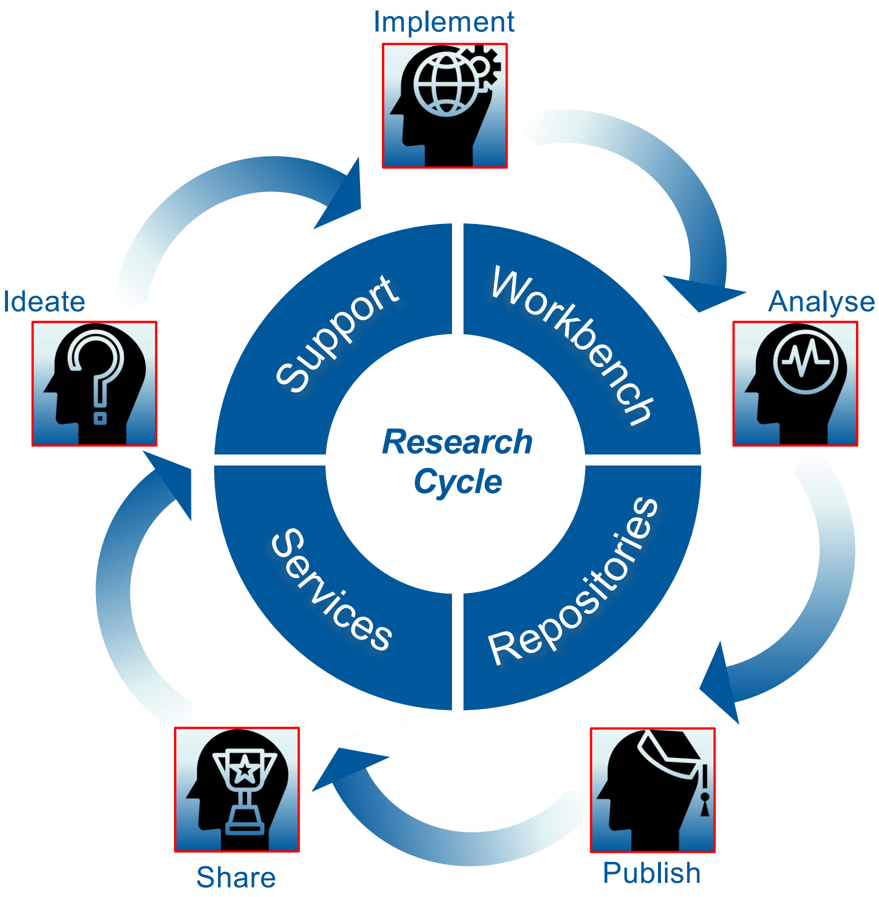
NFDI & International Integration

Community Support

Implementing the NFDI4Earth Strategy



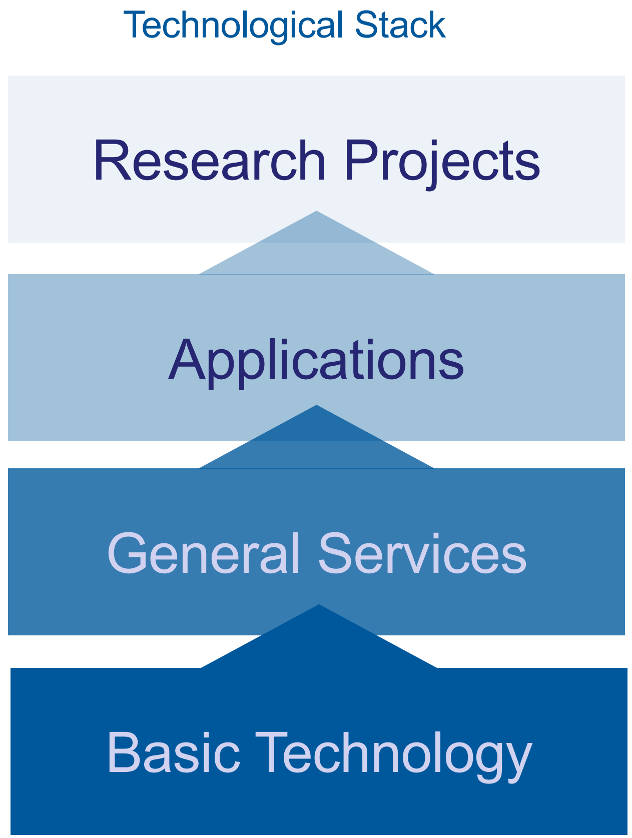
Starting point is the research cycle …



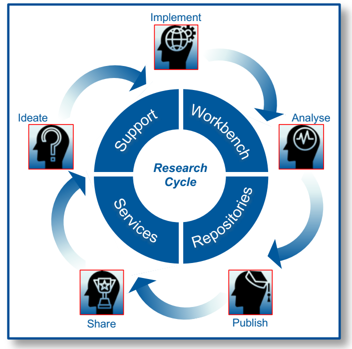
Almost all artefacts of scientific work today are digital.

Our goal is to build a digital ecosystem which allows to move through the steps of the cycle in arbitrary ways.

Therefore, we must integrate services on all levels of the technological stack.



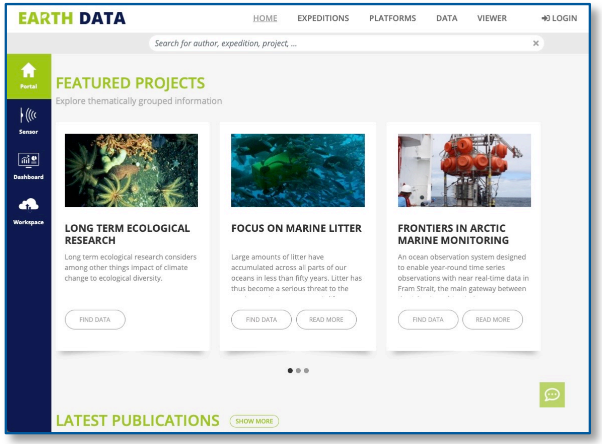
Our vision: a Digital Research Ecosystem …

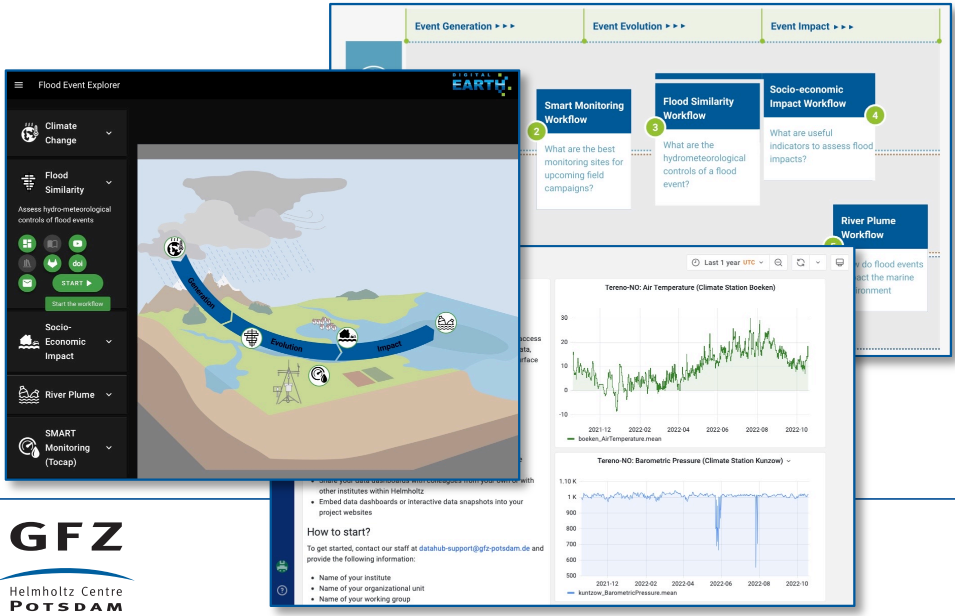


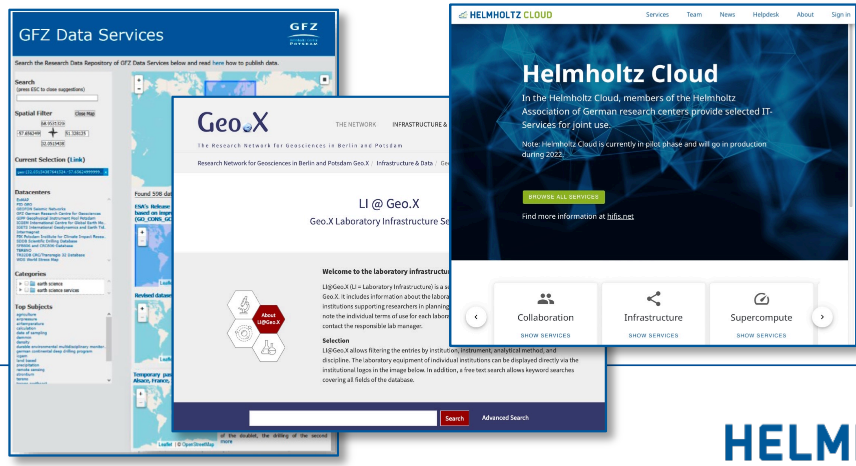
The **cloud** herby abstracts users from the explicit instantiation of technology.

Digitalization in the Research Field Earth & Environment

DataHub
Research Field Earth and Environment



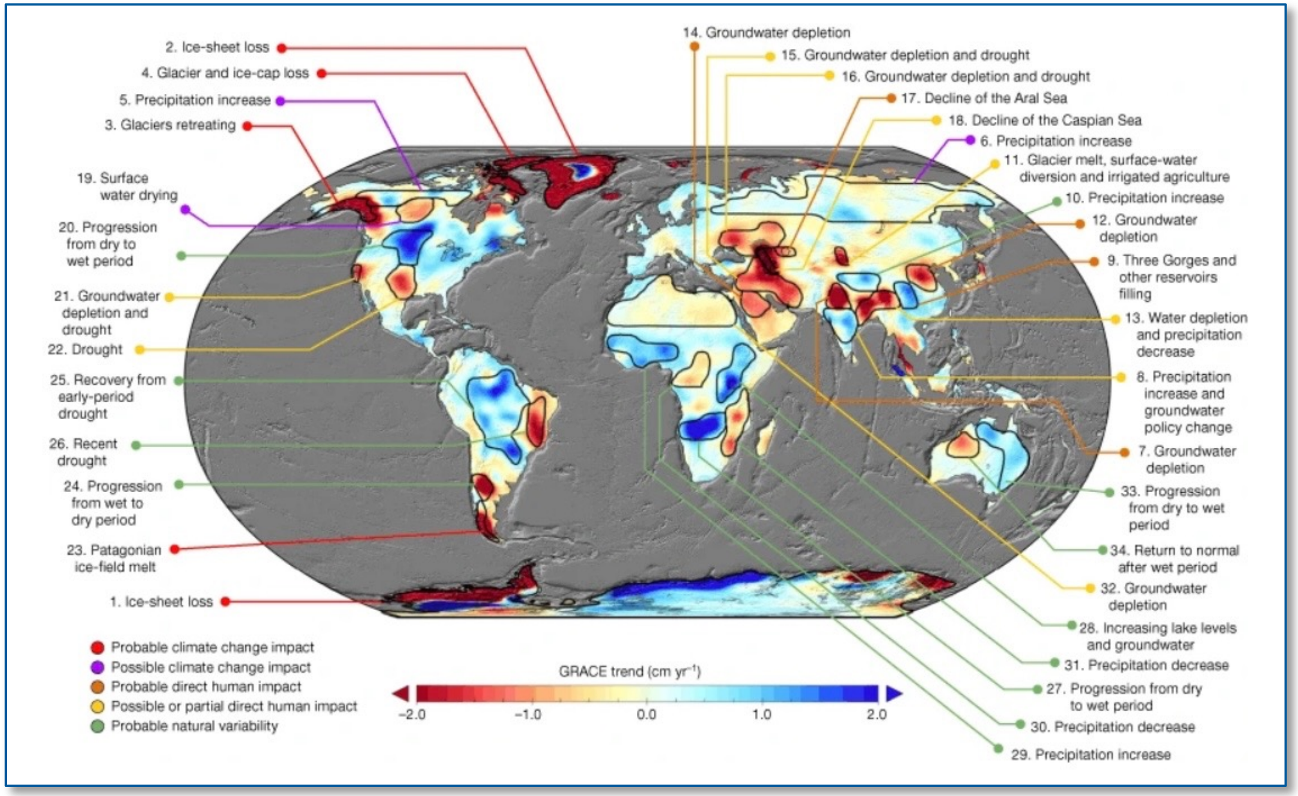




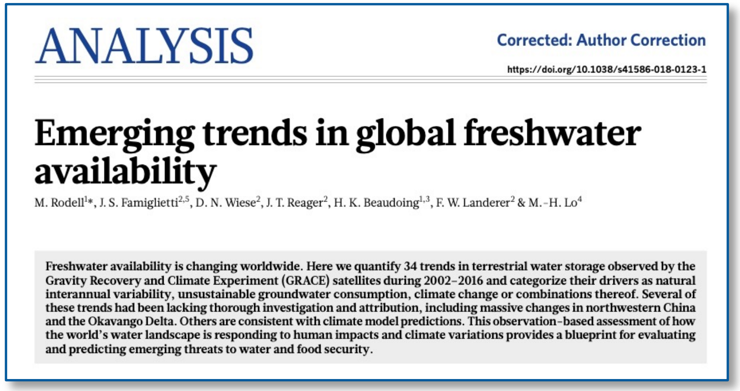
GFZ
Helmholtz Centre
POTSDAM

HELMHOLTZ

An example – analysis of terrestrial water storage



Rodell et al., Nat 2018



ANALYSIS

Corrected: Author Correction

https://doi.org/10.1038/s41586-018-0123-1

Emerging trends in global freshwater availability

M. Rodell[1*], J. S. Famiglietti[2,5], D. N. Wiese[2], J. T. Reager[2], H. K. Beaudoing[1,3], F. W. Landerer[2] & M.-H. Lo[4]

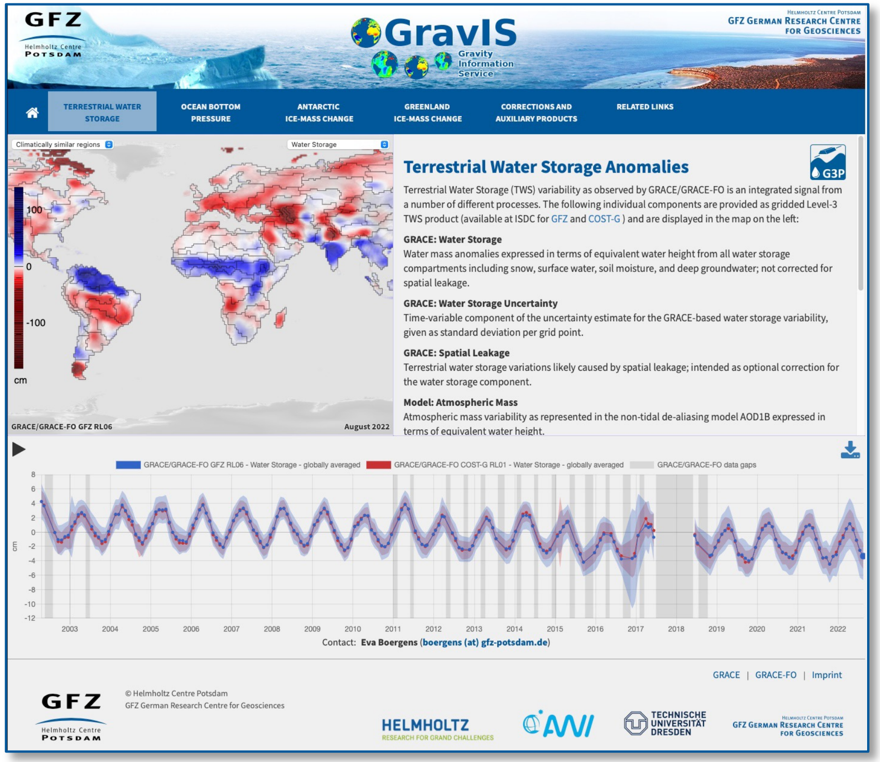
Freshwater availability is changing worldwide. Here we quantify 34 trends in terrestrial water storage observed by the Gravity Recovery and Climate Experiment (GRACE) satellites during 2002–2016 and categorize their drivers as natural interannual variability, unsustainable groundwater consumption, climate change or combinations thereof. Several of these trends had been lacking thorough investigation and attribution, including massive changes in northwestern China and the Okavango Delta. Others are consistent with climate model predictions. This observation-based assessment of how the world's water landscape is responding to human impacts and climate variations provides a blueprint for evaluating and predicting emerging threats to water and food security.

Analysis is build on using data of the GRACE satellite mission.

The example ctd´ – GRACE/GRACE-FO satellite mission

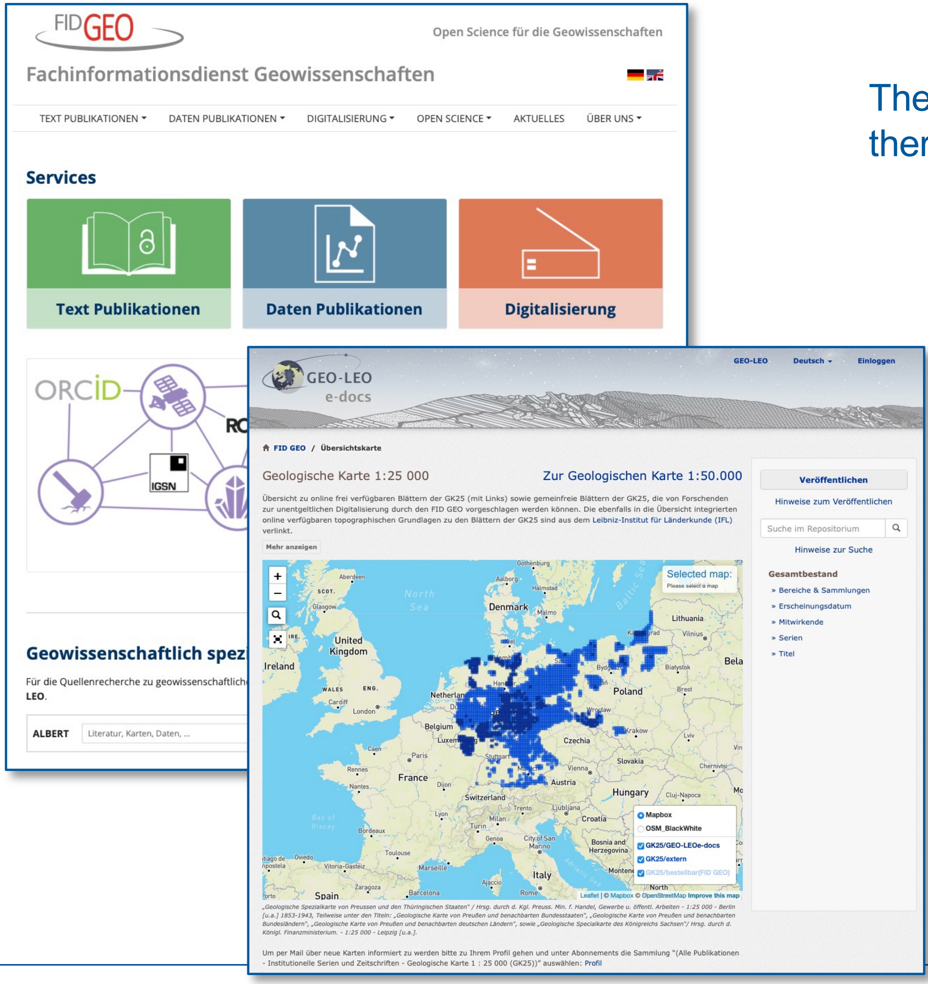
From data …

… to data products …



… to services!

First challenge – **finding data**



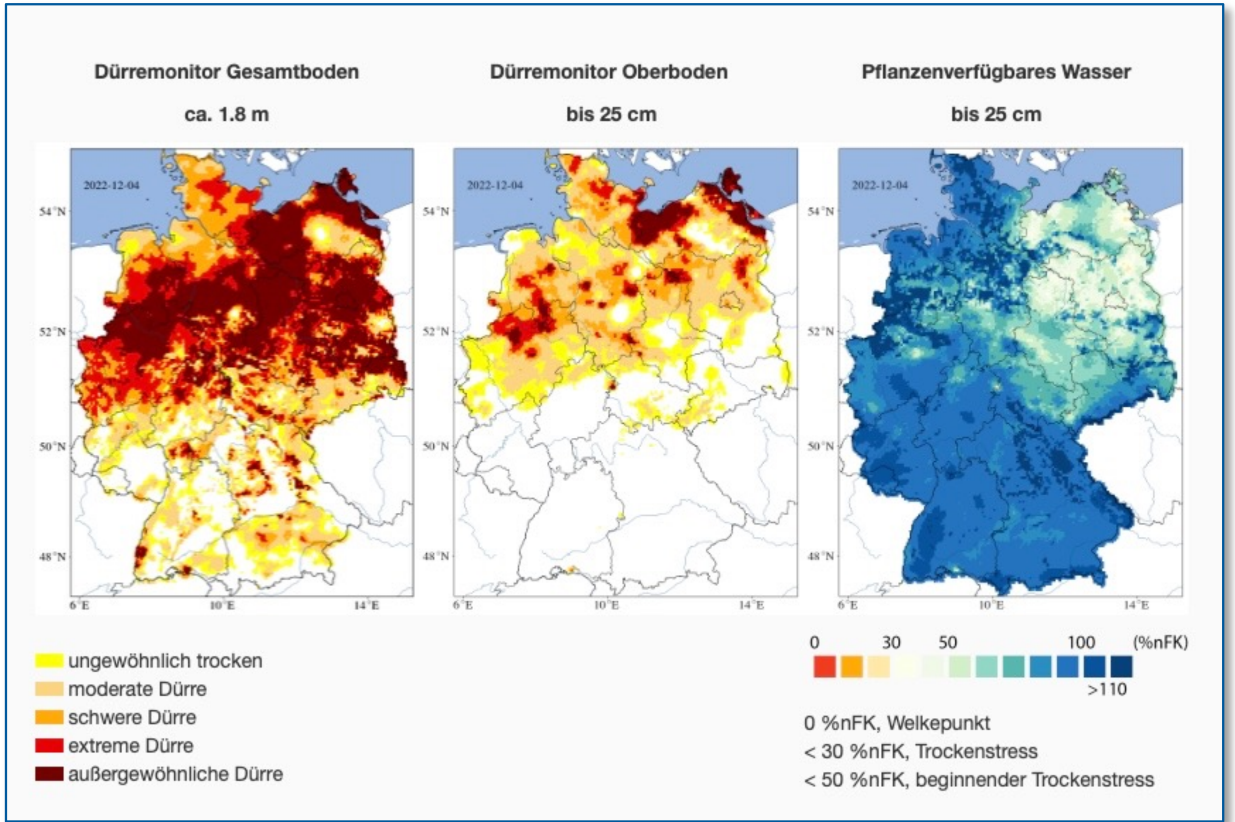
There is really a lot out there …

… organized by scientific domains, organizations, (inter-)national agents, public authorities etc.

Integrative services and metadata search engines do help!

Second challenge – showing vs **accessing data**

UFZ Dürremonitor Deutschland

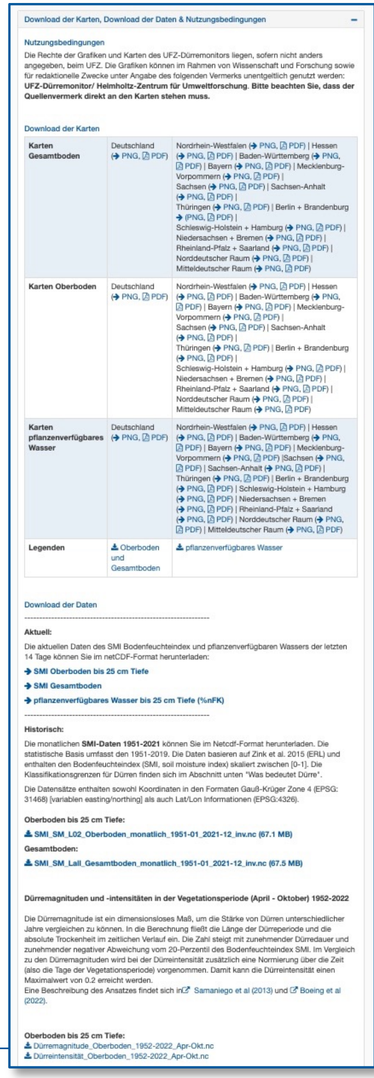


https://https://www.ufz.de/index.php?de=37937

Useful services are also a popular gateway to data.

Can the data actually be used … ?

… in the sense of FAIR?



Download der Karten, Download der Daten & Nutzungsbedingungen

Nutzungsbedingungen

Die Rechte der Grafiken und Karten des UFZ-Dürremonitors liegen, sofern nicht anders angegeben, beim UFZ. Die Grafiken können im Rahmen von Wissenschaft und Forschung sowie für redaktionelle Zwecke unter Angabe des folgenden Vermerks unentgeltlich genutzt werden: **UFZ-Dürremonitor/ Helmholtz-Zentrum für Umweltforschung. Bitte beachten Sie, dass der Quellenvermerk direkt an den Karten stehen muss.**

Download der Karten

Karten Gesamtboden	Deutschland (PNG, PDF)	Nordrhein-Westfalen (PNG, PDF) | Hessen (PNG, PDF) | Baden-Württemberg (PNG, PDF) | Bayern (PNG, PDF) | Mecklenburg-Vorpommern (PNG, PDF) | Sachsen (PNG, PDF) | Sachsen-Anhalt (PNG, PDF) | Thüringen (PNG, PDF) | Berlin + Brandenburg (PNG, PDF) | Schleswig-Holstein + Hamburg (PNG, PDF) | Niedersachsen + Bremen (PNG, PDF) | Rheinland-Pfalz + Saarland (PNG, PDF) | Norddeutscher Raum (PNG, PDF) | Mitteldeutscher Raum (PNG, PDF)
Karten Oberboden	Deutschland (PNG, PDF)	Nordrhein-Westfalen (PNG, PDF) | Hessen (PNG, PDF) | Baden-Württemberg (PNG, PDF) | Bayern (PNG, PDF) | Mecklenburg-Vorpommern (PNG, PDF) | Sachsen (PNG, PDF) | Sachsen-Anhalt (PNG, PDF) | Thüringen (PNG, PDF) | Berlin + Brandenburg (PNG, PDF) | Schleswig-Holstein + Hamburg (PNG, PDF) | Niedersachsen + Bremen (PNG, PDF) | Rheinland-Pfalz + Saarland (PNG, PDF) | Norddeutscher Raum (PNG, PDF) | Mitteldeutscher Raum (PNG, PDF)
Karten pflanzenverfügbares Wasser	Deutschland (PNG, PDF)	Nordrhein-Westfalen (PNG, PDF) | Hessen (PNG, PDF) | Baden-Württemberg (PNG, PDF) | Bayern (PNG, PDF) | Mecklenburg-Vorpommern (PNG, PDF) |Sachsen (PNG, PDF) | Sachsen-Anhalt (PNG, PDF) | Thüringen (PNG, PDF) | Berlin + Brandenburg (PNG, PDF) | Schleswig-Holstein + Hamburg (PNG, PDF) | Niedersachsen + Bremen (PNG, PDF) | Rheinland-Pfalz + Saarland (PNG, PDF) | Norddeutscher Raum (PNG, PDF) | Mitteldeutscher Raum (PNG, PDF)
Legenden	Oberboden und Gesamtboden	pflanzenverfügbares Wasser

Download der Daten

Aktuell:

Die aktuellen Daten des SMI Bodenfeuchteindex und pflanzenverfügbaren Wassers der letzten 14 Tage können Sie im netCDF-Format herunterladen:

SMI Oberboden bis 25 cm Tiefe

SMI Gesamtboden

pflanzenverfügbares Wasser bis 25 cm Tiefe (%nFK)

Historisch:

Die monatlichen **SMI-Daten 1951-2021** können Sie im Netcdf-Format herunterladen. Die statistische Basis umfasst den 1951-2019. Die Daten basieren auf Zink et al. 2015 (ERL) und enthalten den Bodenfeuchteindex (SMI, soil moisture index) skaliert zwischen [0-1]. Die Klassifikationsgrenzen für Dürren finden sich im Abschnitt unten "Was bedeutet Dürre".

Die Datensätze enthalten sowohl Koordinaten in den Formaten Gauß-Krüger Zone 4 (EPSG: 31468) [variablen easting/northing] als auch Lat/Lon Informationen (EPSG:4326).

Oberboden bis 25 cm Tiefe:

SMI_SM_L02_Oberboden_monatlich_1951-01_2021-12_inv.nc (67.1 MB)

Gesamtboden:

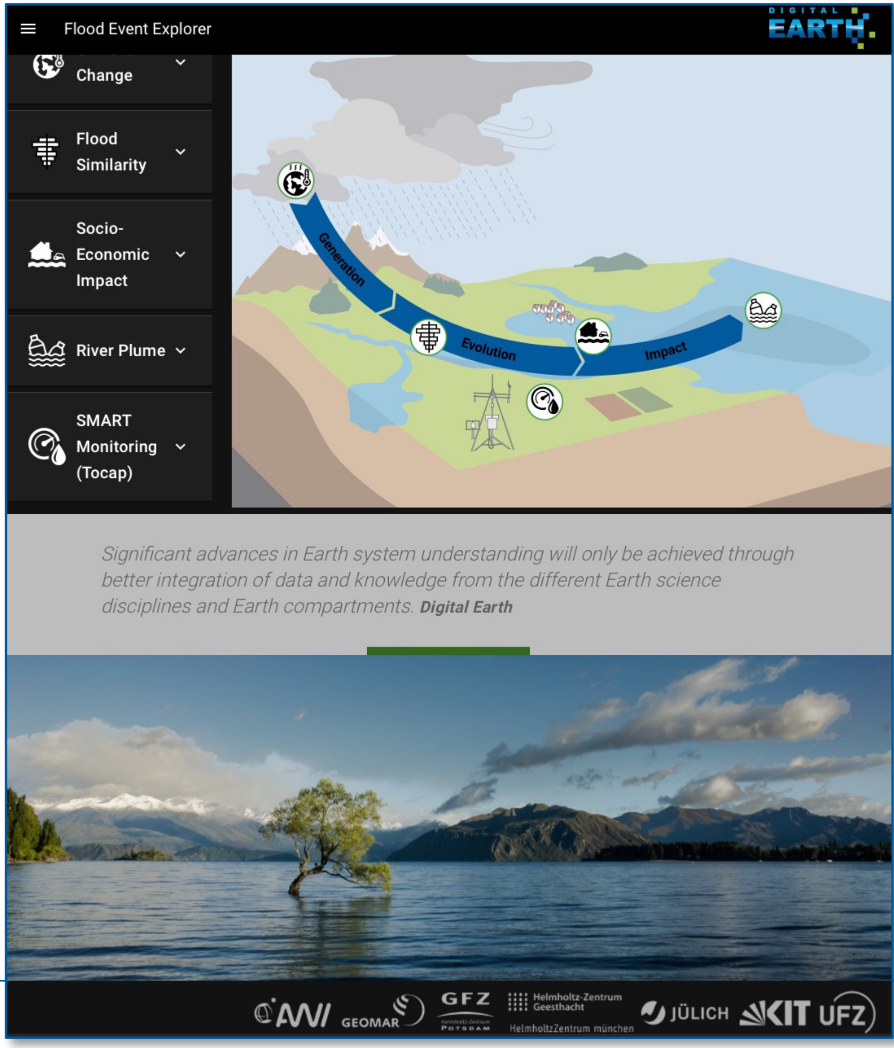
SMI_SM_Lall_Gesamtboden_monatlich_1951-01_2021-12_inv.nc (67.5 MB)

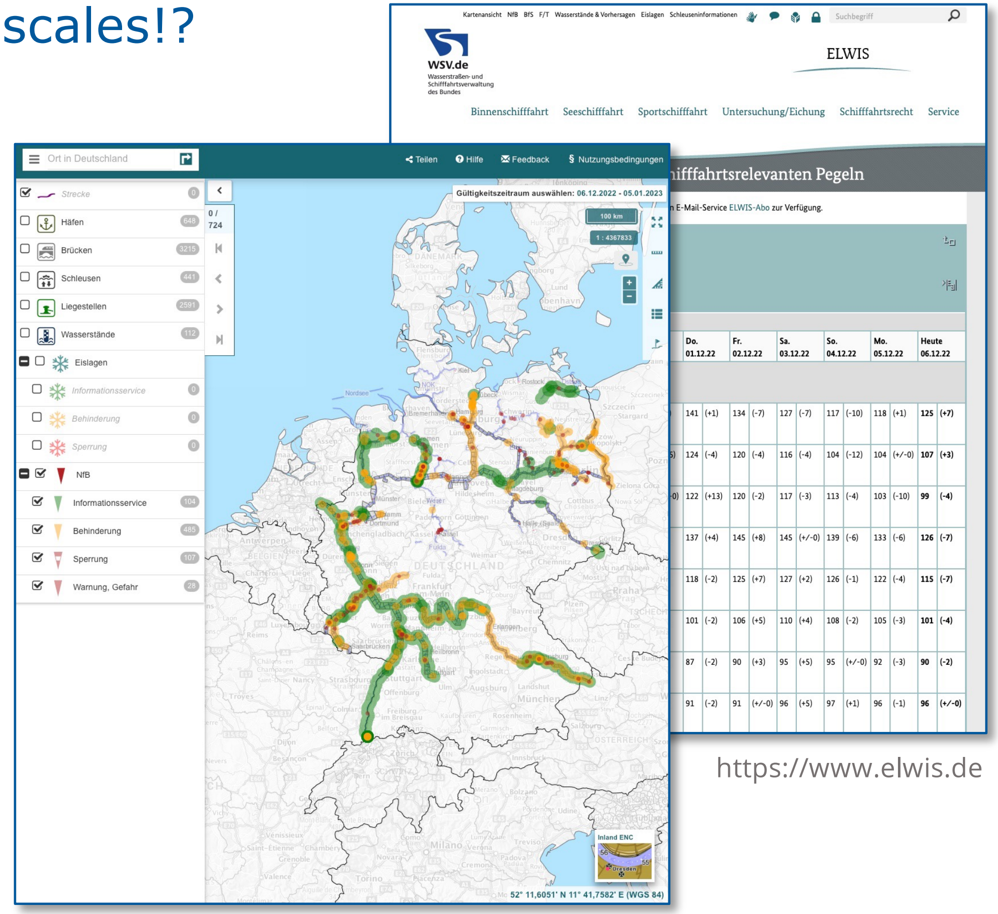
Dürremagnituden und -intensitäten in der Vegetationsperiode (April - Oktober) 1952-2022

Die Dürremagnitude ist ein dimensionsloses Maß, um die Stärke von Dürren unterschiedlicher Jahre vergleichen zu können. In die Berechnung fließt die Länge der Dürreperiode und die absolute Trockenheit im zeitlichen Verlauf ein. Die Zahl steigt mit zunehmender Dürredauer und zunehmender negativer Abweichung vom 20-Perzentil des Bodenfeuchteindex SMI. Im Vergleich zu den Dürremagnituden wird bei der Dürreintensität zusätzlich eine Normierung über die Zeit (also die Tage der Vegetationsperiode) vorgenommen. Damit kann die Dürreintensität einen Maximalwert von 0.2 erreicht werden.
Eine Beschreibung des Ansatzes findet sich in Samaniego et al (2013) und Boeing et al (2022).

Oberboden bis 25 cm Tiefe:
Dürremagnitude_Oberboden_1952-2022_Apr-Okt.nc
Dürreintensität_Oberboden_1952-2022_Apr-Okt.nc

Digital Earth Flood Event Explorer – bridging scales!?



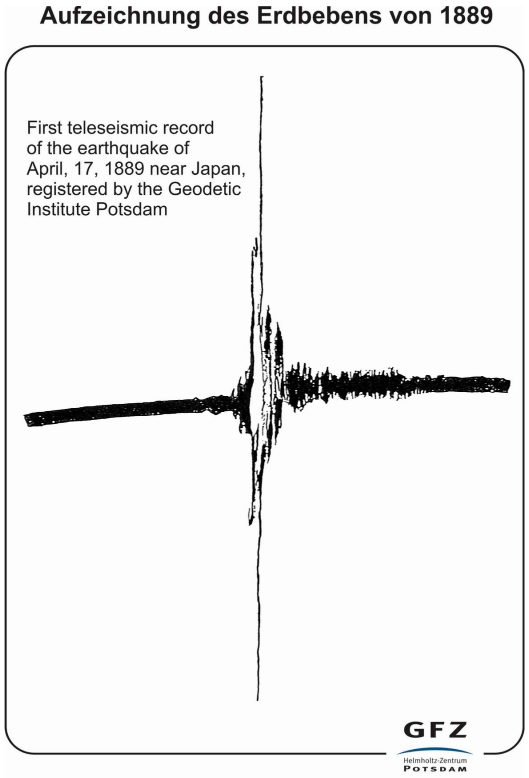


https://www.elwis.de

https://via.bund.de/wsv/elwis

Third challenge – **interoperable data**

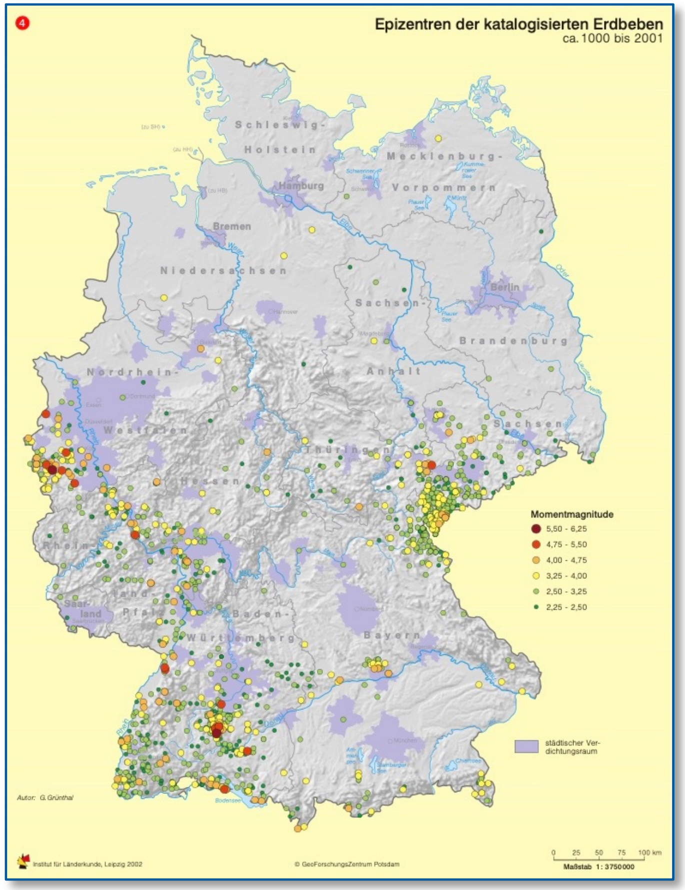
You never know who will do what with your data …



Working in geological timescales implies that mostly, proxies have to be used.

Historical recording of earthquake in the Bielefeld area 1612

Vogt & Grünthal, Geowissenschaften 1994



G. Grünthal, in: Nationalatlas Bundesrepublik Deutschland Band 2, 44-45

Fourth challenge – **reusable data**

Long-term preservation – scientists use what they can get …

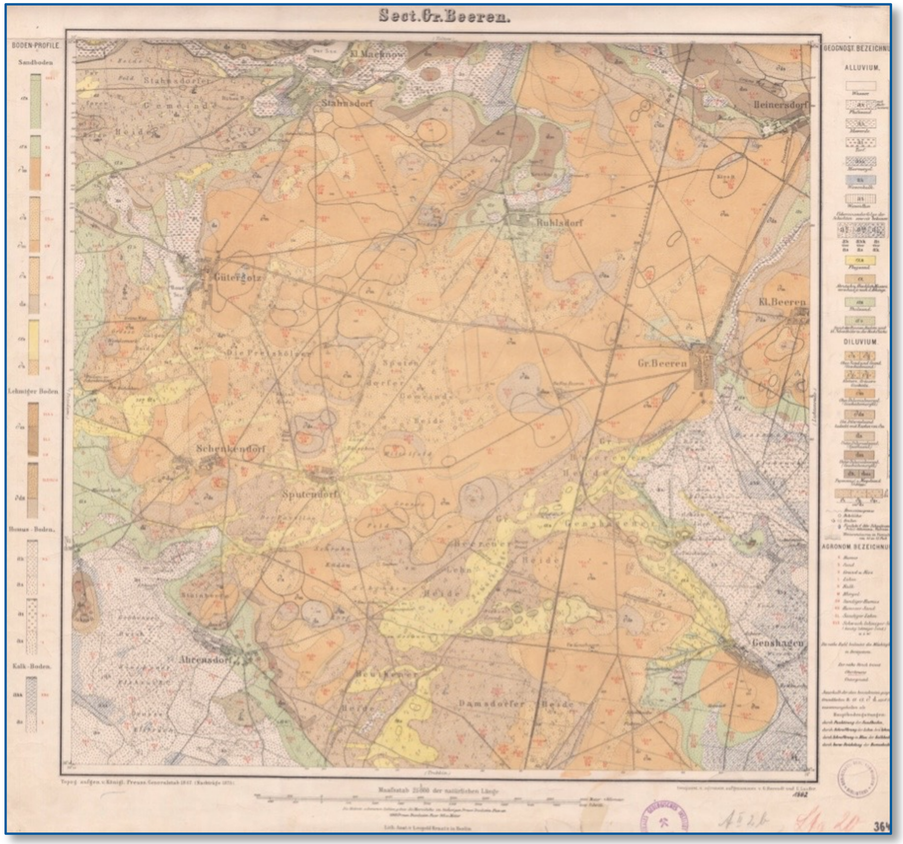
That‘s why they tend to keep everything!

Historical recording of earthquake in the Bielefeld area 1612

Vogt & Grünthal, Geowissenschaften 1994

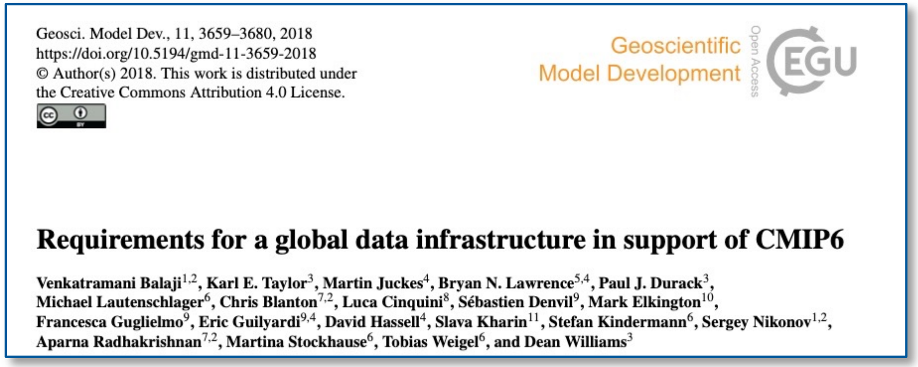
The clue of the data perspective is to use what is available rather than to wait for what can be created.

The treasure of time series!



Geologische Karte, G. Berendt 1875, in: Digitales Brandenburg

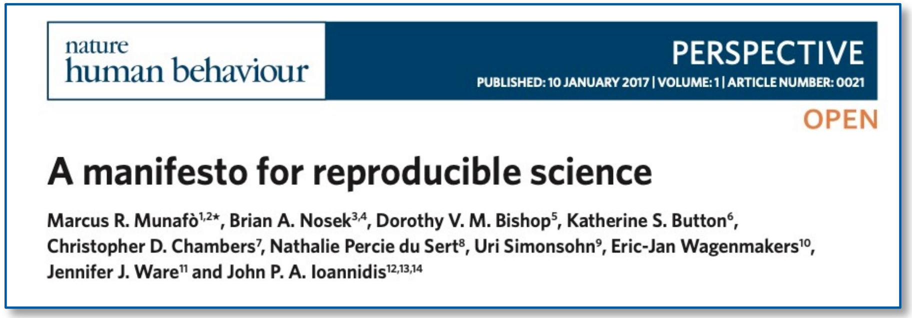
Preservation in science … more than 10 years?


Geosci. Model Dev., 11, 3659–3680, 2018
https://doi.org/10.5194/gmd-11-3659-2018

Geoscientific Model Development
Open Access
EGU

Requirements for a global data infrastructure in support of CMIP6

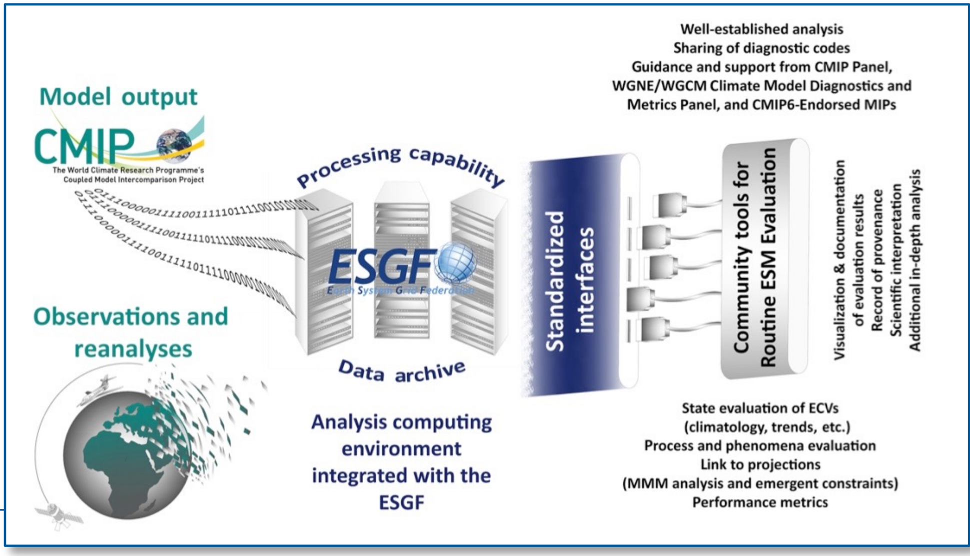
Venkatramani Balaji[1,2], Karl E. Taylor[3], Martin Juckes[4], Bryan N. Lawrence[5,4], Paul J. Durack[3], Michael Lautenschlager[6], Chris Blanton[7,2], Luca Cinquini[8], Sébastien Denvil[9], Mark Elkington[10], Francesca Guglielmo[9], Eric Guilyardi[9,4], David Hassell[4], Slava Kharin[11], Stefan Kindermann[6], Sergey Nikonov[1,2], Aparna Radhakrishnan[7,2], Martina Stockhause[6], Tobias Weigel[6], and Dean Williams[3]


nature human behaviour
PERSPECTIVE
PUBLISHED: 10 JANUARY 2017 | VOLUME: 1 | ARTICLE NUMBER: 0021
OPEN

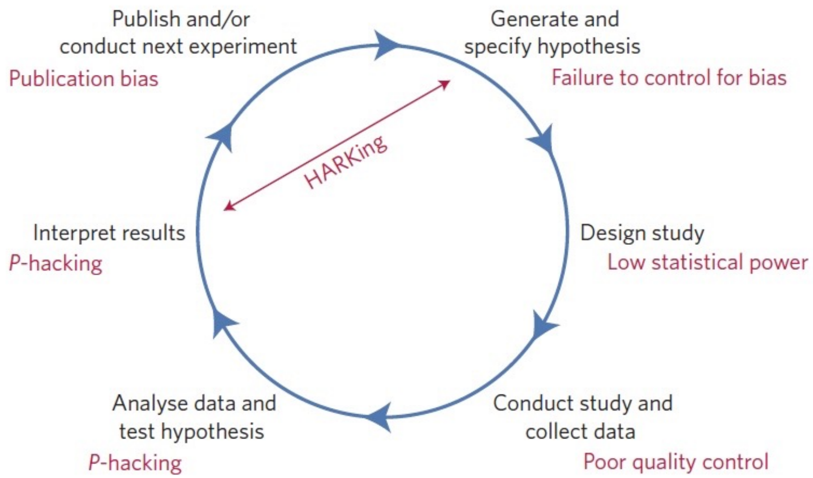
A manifesto for reproducible science

Marcus R. Munafò[1,2]*, Brian A. Nosek[3,4], Dorothy V. M. Bishop[5], Katherine S. Button[6], Christopher D. Chambers[7], Nathalie Percie du Sert[8], Uri Simonsohn[9], Eric-Jan Wagenmakers[10], Jennifer J. Ware[11] and John P. A. Ioannidis[12,13,14]

Coupled Model Intercomparison Project (CIMP)



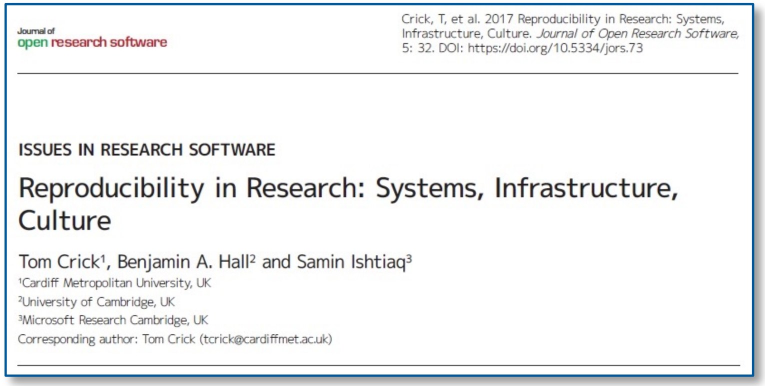
Eyring et al., Earth Syst Dynam 2016



Munafo et al., Nat Hum Beh 2017

„We need to talk about reproducibility“

Crick, Hall and Ishtiaq cite studies, finding that **50% of published studies** (including top-tier journals) **cannot be repeated** by an industrial lab.


Journal of open research software

Crick, T, et al. 2017 Reproducibility in Research: Systems, Infrastructure, Culture. *Journal of Open Research Software*, 5: 32. DOI: https://doi.org/10.5334/jors.73

ISSUES IN RESEARCH SOFTWARE

Reproducibility in Research: Systems, Infrastructure, Culture

Tom Crick[1], Benjamin A. Hall[2] and Samin Ishtiaq[3]

[1]Cardiff Metropolitan University, UK

[2]University of Cambridge, UK

[3]Microsoft Research Cambridge, UK

Corresponding author: Tom Crick (tcrick@cardiffmet.ac.uk)

conclusion

- just publishing linked data is not enough
- „set the code free“
- welcome Web 2.0 technologies (use and share workflows, provide web services, use cloud …)

To what **level of detail** do we need to keep data and what are appropriate **time scales**?

Scientific data … high diversity & low standardization

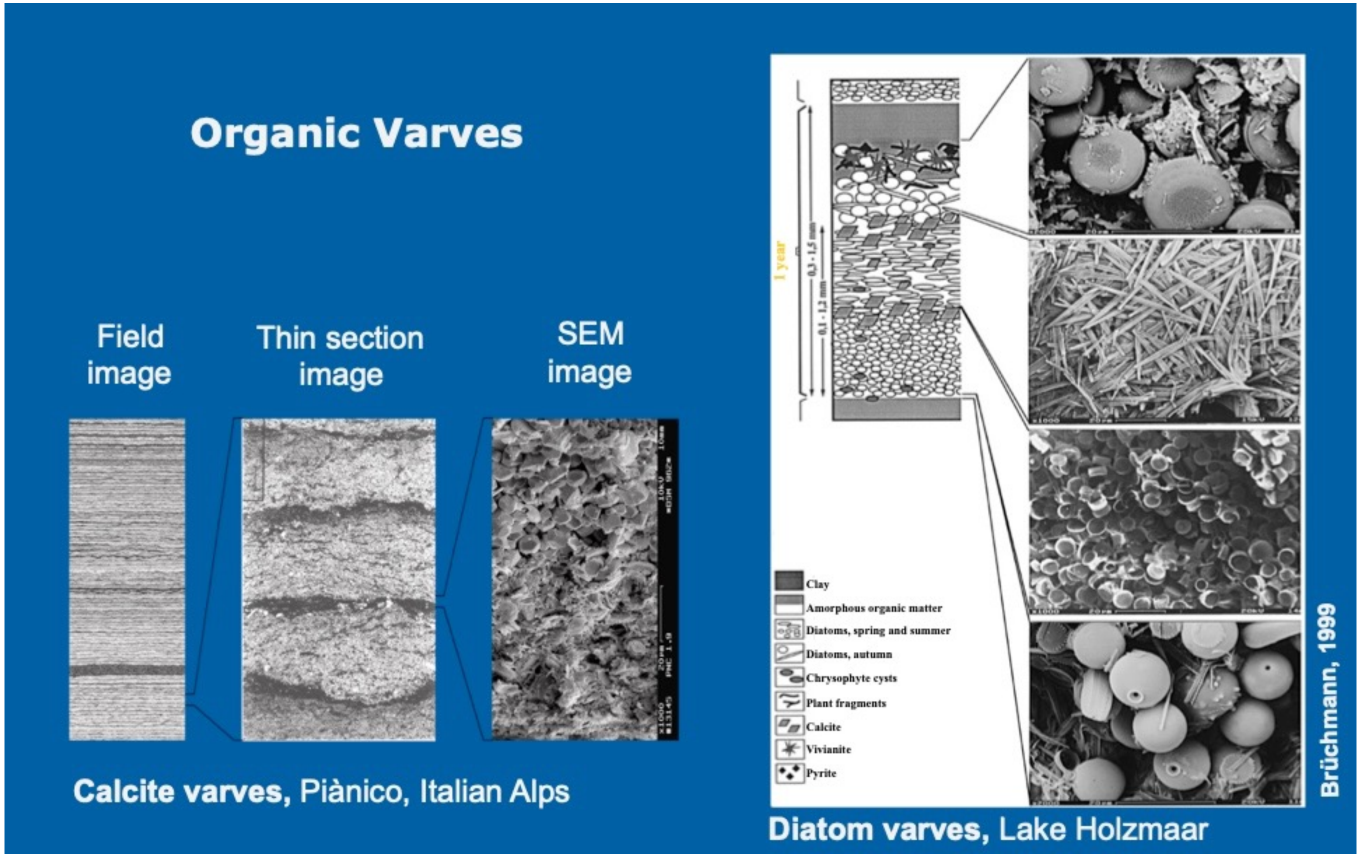
- need to deal with the „long-tail“ of research data
- lack of comprehensive criteria for long-term preservation
- joint issues: formats, accessibility, data migration

There are different challenges/requirements for

- closed projects
- real-time data products
- simulation data

Clouds above Potsdam
(by R. Süring)

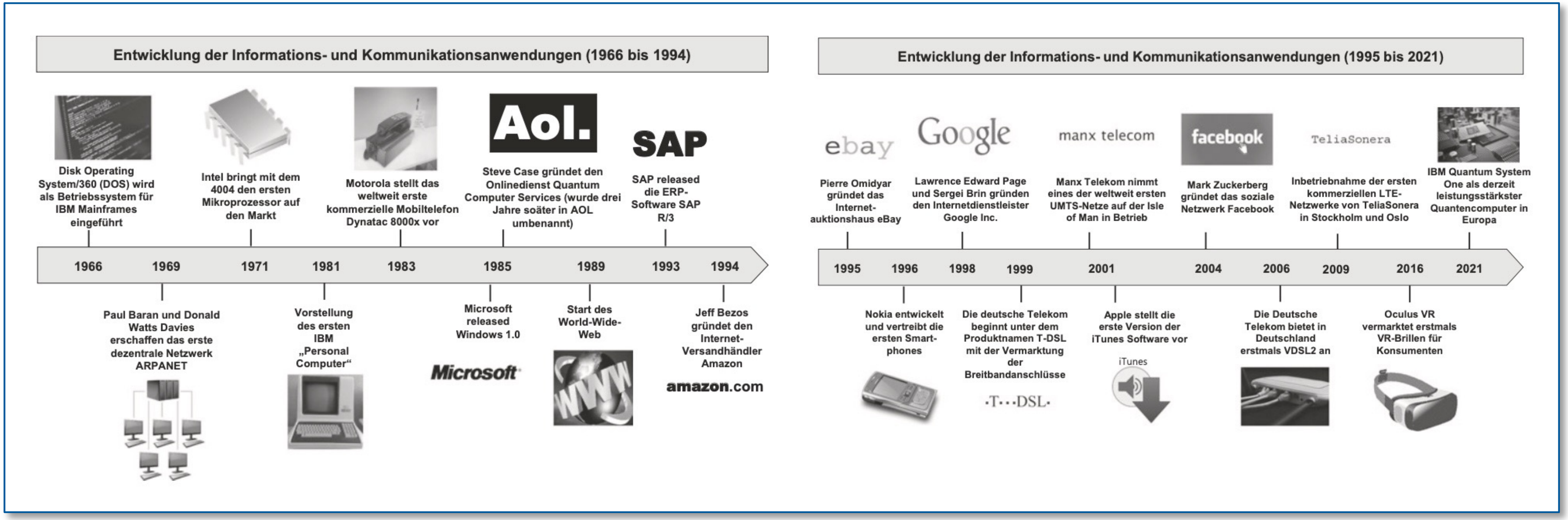
Digitalization – the challenge of formats and technology



A. Brauer, IRTG-STRATEGY Lecture 2016

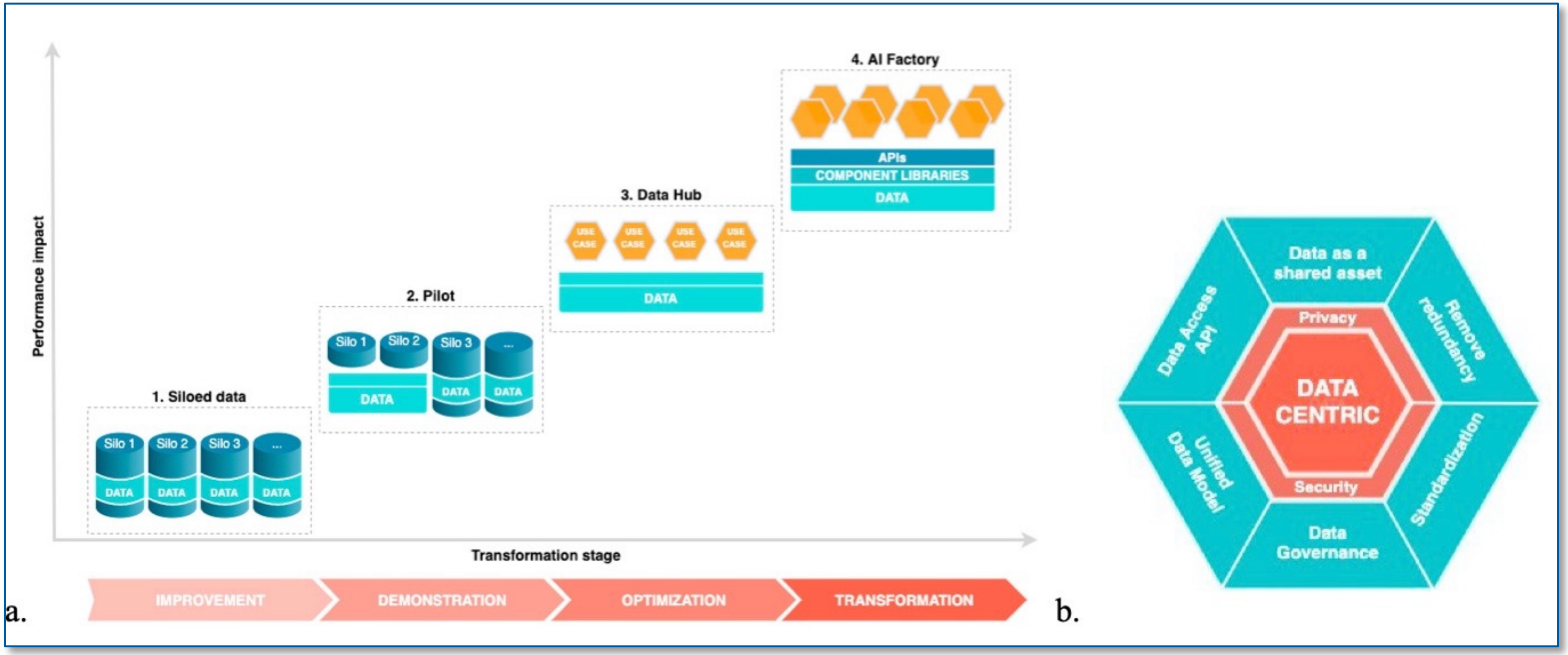
paleoclimate reconstruction
from lake sediments

Digitalization – the challenge of formats and technology



B.W. Wirtz, E-Government, Springer 2022

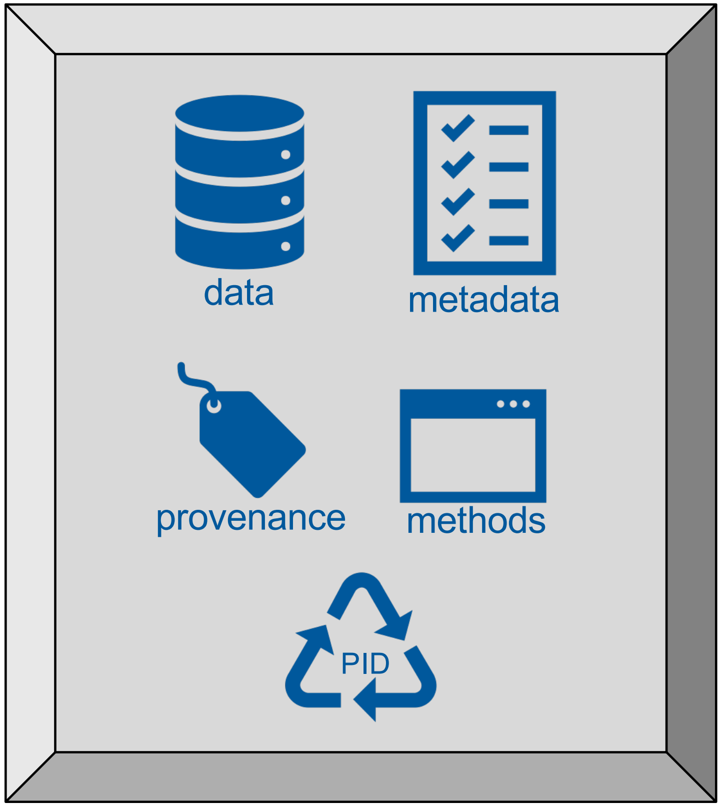
Towards a data centric architecture …



Alvarez-Coello et al., Procedia Comp Sci 2021

- overcoming silos is the first step, only
- organizing silos as we did data means just scaling the challenge
- keeping all necessary information together will become a bottleneck

From thinking in applications to thinking in data …



extended data containers

These are the key principles of the data centric manifesto:

1. Data is a key asset of any person, organization, and society.
2. Data is self-describing and does not rely on an application for interpretation and meaning.
3. Data is expressed in open, non-proprietary formats.
4. Access to and security of the data is a responsibility of the enterprise data layer or the personal data vault, and not managed by applications.
5. Applications are allowed to visit the data, perform their magic and express the results of their process back into the data layer.

The Data-Centric Manifesto

Thank you for your attention!