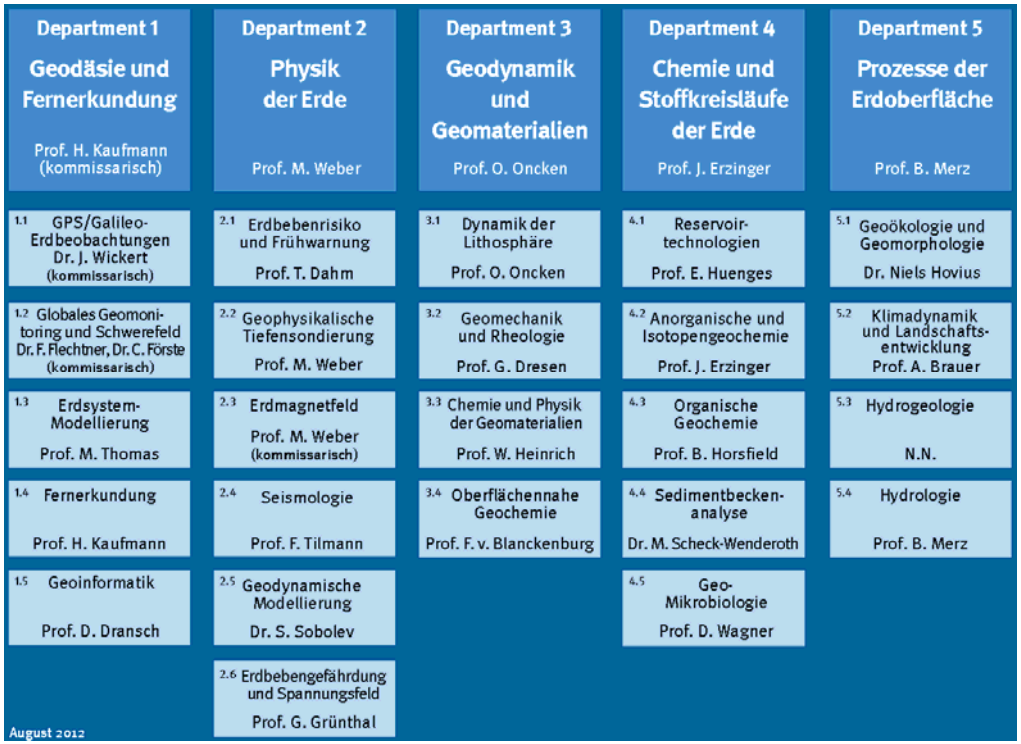
Infrastrukturen für Forschungsdaten
- Herausforderungen, Wünsche, Anreize -

Maik Thomas & Jens Klump

Helmholtz-Zentrum Potsdam
Deutsches GeoForschungsZentrum GFZ

Beispiel: GFZ Potsdam

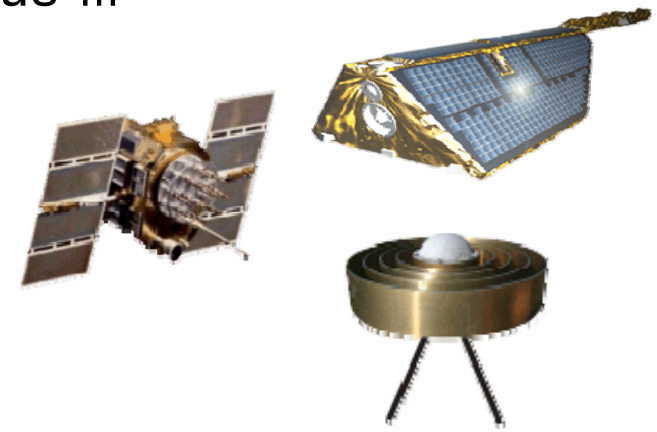


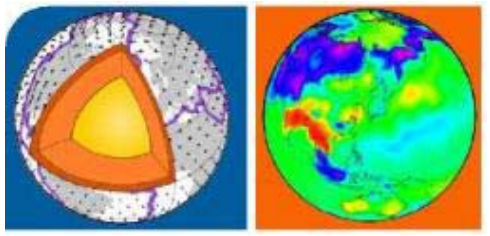
- **5 Forschungsbereiche (Departments) mit je 4-6 Sektionen und mehr als 1000 Mitarbeitern**
- **zahlreiche Projekte in allen Sektionen**
- **mehr als 100 verschiedene Datenquellen**

Beispiel: GFZ Potsdam

Eine Vielzahl von Datensätzen resultiert aus ...

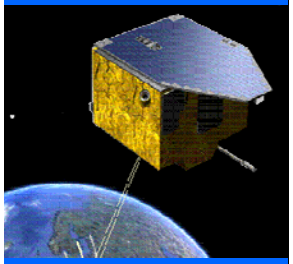
- **Erdbeobachtungssystemen**
  - erzeugen meist große Datenmengen
  - weitgehend homogene Strukturen
  - teilweise automatisierte Workflows
- **Erdsystemmodellierung**
  - erzeugt mäßig große Datenmengen
  - teilweise homogene Strukturen
  - aber meist ohne unterstützende Workflows
- **Labordaten und Feldbeobachtungen**
  - zumeist nur kleine Datenmengen
  - aber mit heterogenen Strukturen
  - resultierende Datensätze entstehen in ad-hoc formulierten Strukturen





Erdsystemdynamik: Monitoring, Analyse, Simulation

Hyperspectral mission

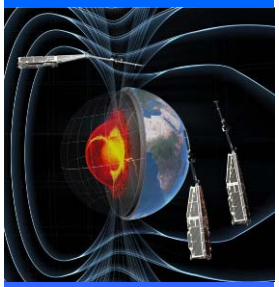


EnMAP

Terrestrial networks

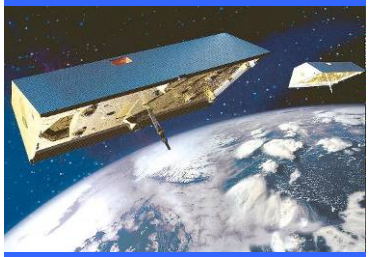
Tide gauge | GPS | SG | Geophone

Magnetic field



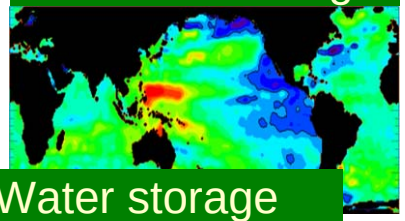
SWARM

Gravimetry missions



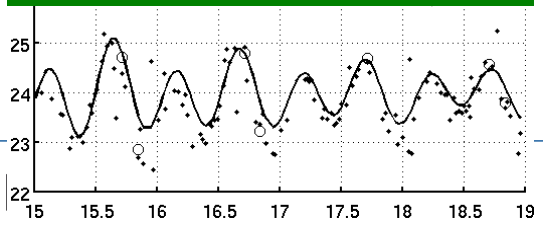
GRACE-FO

Sea level change



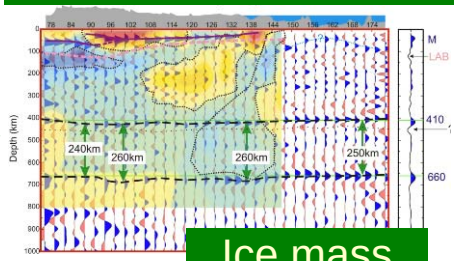
Water storage

Ocean tides

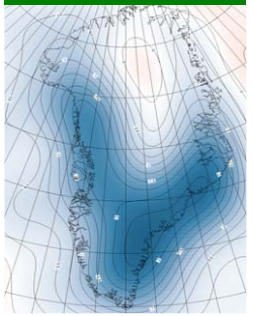


Integration, combination, assimilation

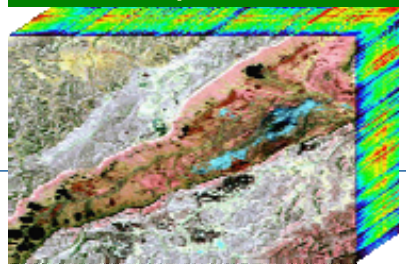
Lithospheric structure



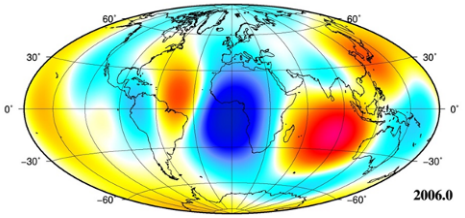
Ice mass change



Surface parameters



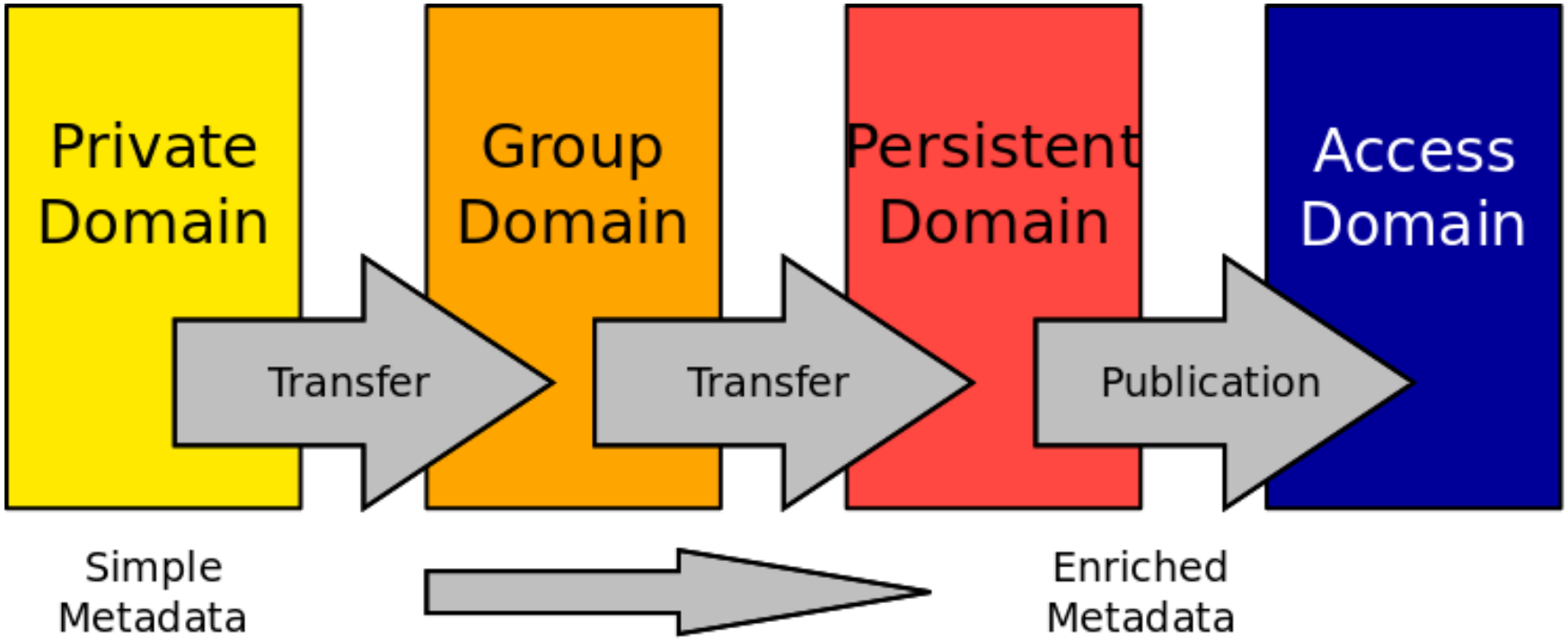
Magnetic field variations



Helmholtz-Zentrum Potsdam

HELMHOLTZ GEMEINSCHAFT

Data Life Cycle Model



↳ **Wie kommen Forschungsdaten zu externen Nutzern ?**

Eine Insel(-lösung?):
Information System and Data Center (ISDC)

Schwerefelddaten & -modelle

terrestrische Gravimetrie

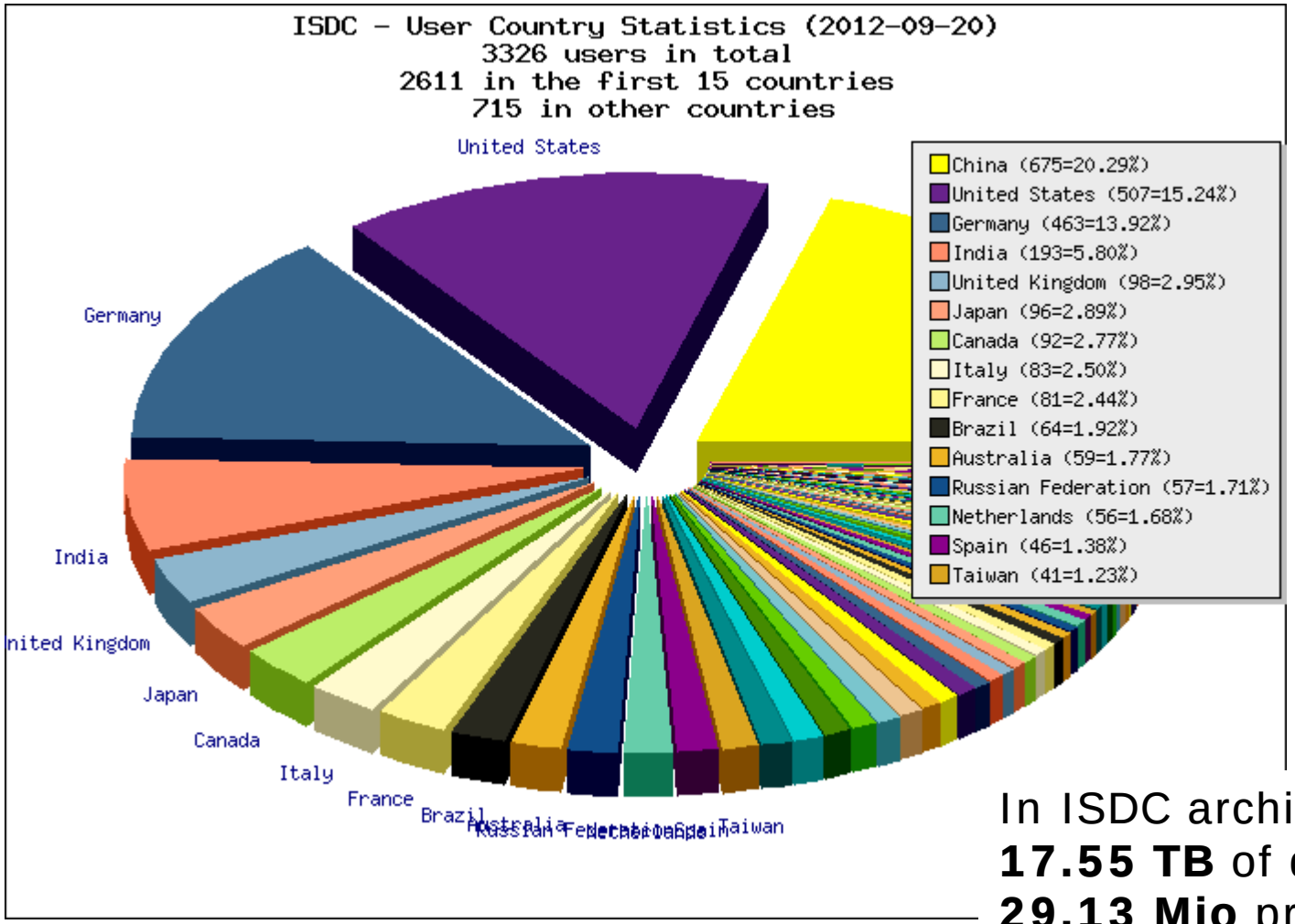
GPS-Daten, Galileo-Service

GPS-PDR

gsp

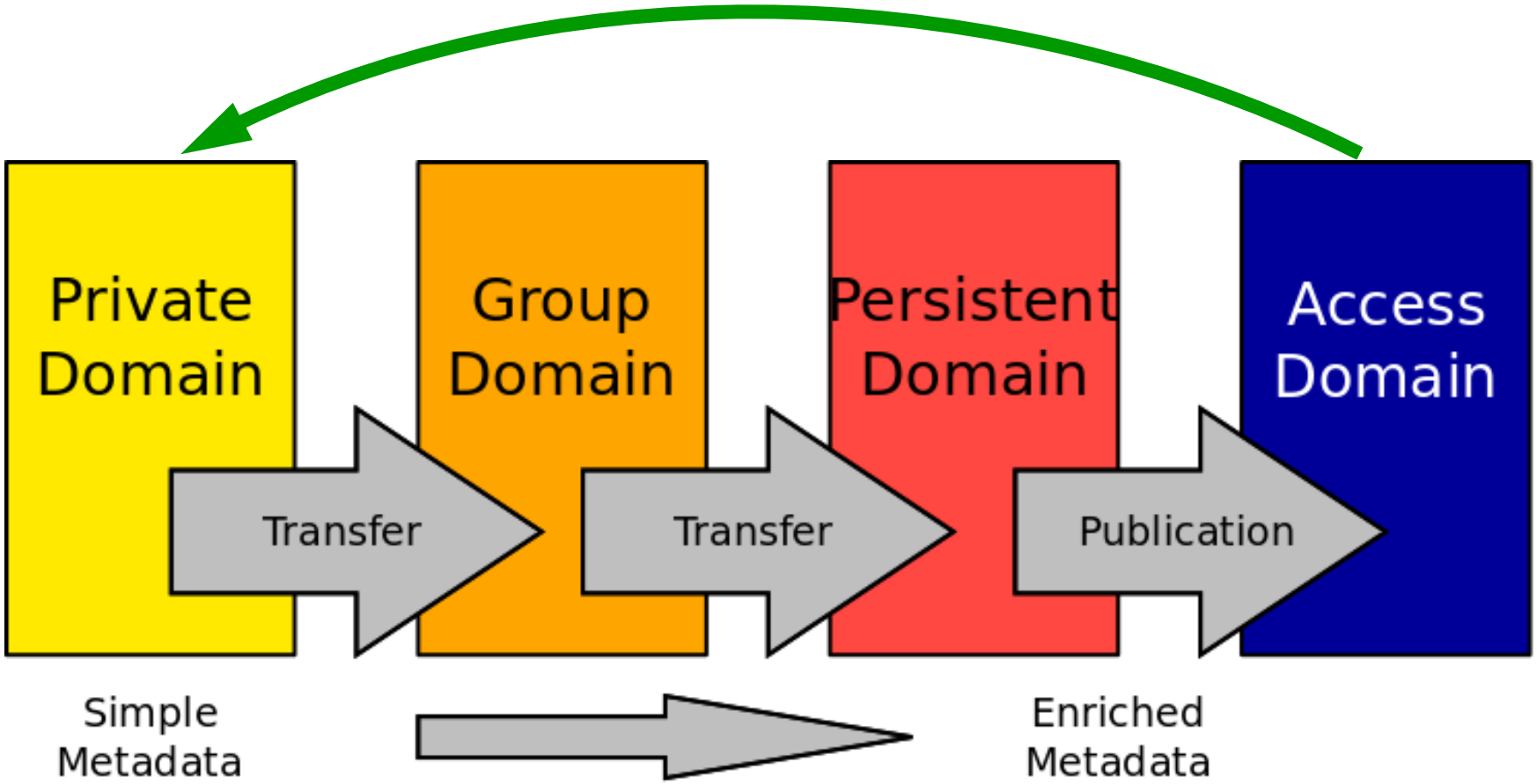
↳ **verschiedenartige Datenprodukte mit Metadaten, Genauigkeits-informationen, FAQs und User Support, sortiert nach Großprojekten**

ISDC: Statistics



In ISDC archive are stored:
17.55 TB of data
29.13 Mio products

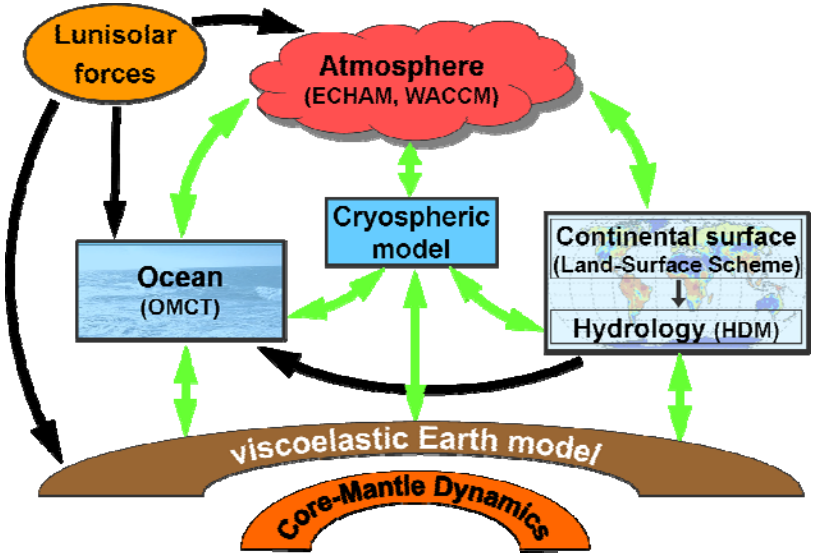
Data Life Cycle Model



↳ **Funktioniert die Integration von Daten benachbarter Forschungsbereiche ?**

Von Multi- über Inter- zu Transdisziplinarität

- viele Jahre der wissenschaftlichen Spezialisierung und damit **Differenzierung** in den Wissenschaften
- gegenwärtig erscheint wesentlicher **Fortschritt** besonders **durch Disziplinen übergreifende Forschung** erreichbar
- jüngere **IT-Entwicklungen** bieten grundsätzlich exzellente Voraussetzungen **für eine Wiederannäherung der Disziplinen** (und Forschungskulturen?)

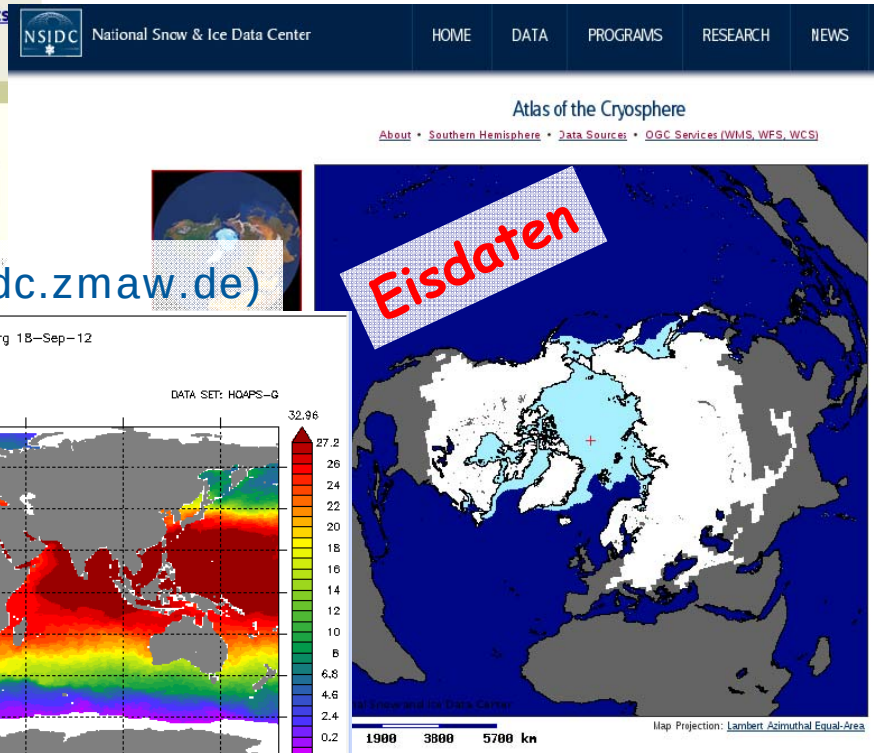


Beispiel:
Klimawandel und seine sozio-ökonomischen Auswirkungen

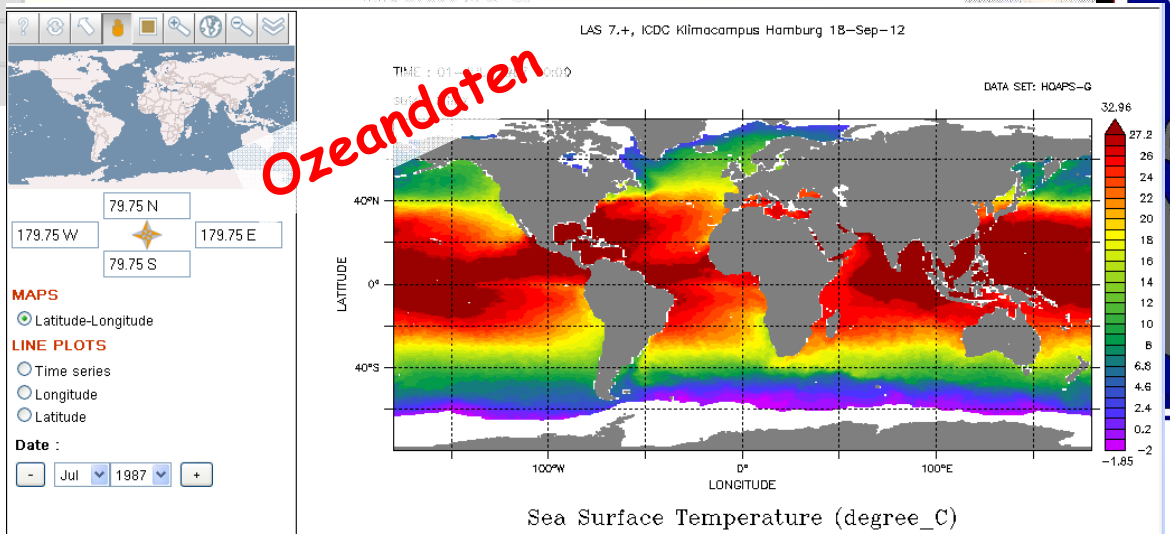
Der geowissenschaftliche Archipel

MARS @ ECMWF (www.ecmwf.int)

NSIDC (www.nsidc.org)



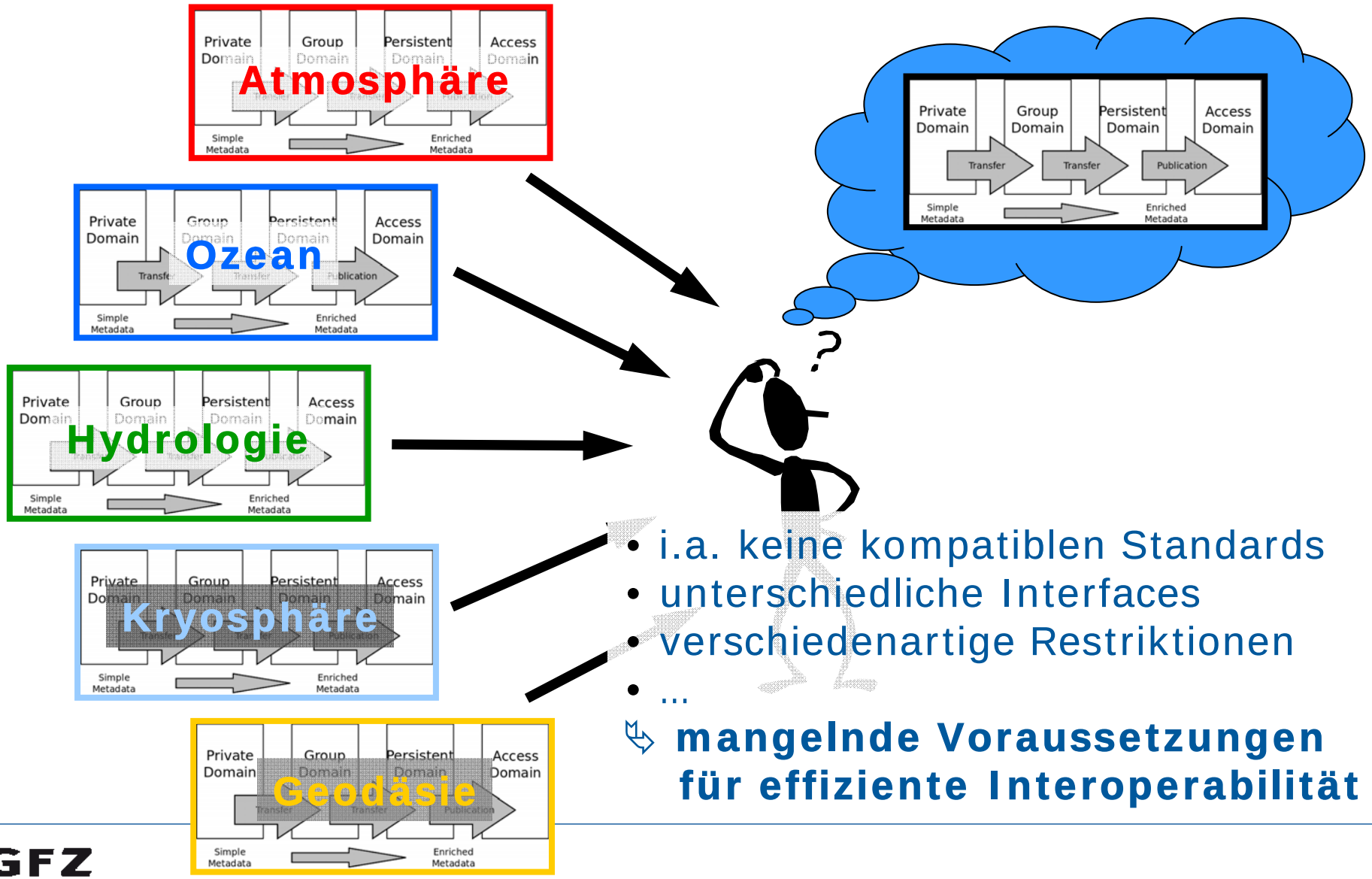
ICDC @ ZMAW Hamburg (icdc.zmaw.de)



GFZ Helmholtz-Zentrum Potsdam

HELMHOLTZ GEMEINSCHAFT

So viele Communities und noch mehr Inseln …



- i.a. keine kompatiblen Standards
- unterschiedliche Interfaces
- verschiedenartige Restriktionen
- …

↳ **mangelnde Voraussetzungen für effiziente Interoperabilität**

Expected Results?

EGS Abstract for
Nice, 2000

accepted !

A NEW WAY OF PRODUCING AND ANALYZING INCO[illegible] TENT, IRRELEVANT AND INCOMPREHENSIBLE DATA

[redacted] WWU Münster, D 48149 Münster, Germany.
[redacted]i-muenster.de/Fax: +49 251 83 39083

The ways by which data are produced are sometimes not fully understood. However, data may sometimes bear no significance at all, especially with respect to coffee making or egg frying ([redacted] *et al.*, 1974). In this paper, we present a new method to produce and analyze data that are inconsistent, irrelevant and incomprehensible.

To produce a set of irrelevant data we sampled the distribution of frozen rabbit droppings in front of our institute as cometary analogues. However, due to the climatic peculiarities of the sampling site, none of the samples collected during the period of July 2nd through August 3rd contained the necessary quantities of water ice. The droppings distribution was then analyzed as a function of space and time by means of the Hazelnut transform, with no understanding of the production process whatsoever. These data may have been inconsistent with the findings of other researchers we have probably never heard of. As a method for analyzing these irrelevant data points, [redacted] (1999) recently proposed a method applying the standard deviation of curvatures in clustered data points as a measure for quantifying an arbitrary line. We cunningly extended this method by inventing an adaptation technique for a semi–empirical line–drawing law for almost any distribution of data points on a double logarithmic scale.

Abstracts to be submitted on or before December 15, 1999 to

EGS Office
Max-Planck-Str. 13
37191 Katlenburg-Lindau
Germany

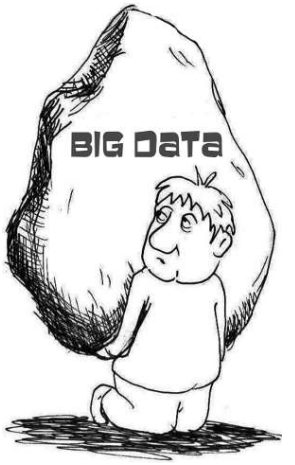
Tel.: [+49] 5556-1440
Fax.: [+49] 5556-4709
Email: EGS@Copernicus.org
http://www.copernicus.org/EGS/EGS.html

GFZ Helmholtz-Zentrum Potsdam

HELMHOLTZ GEMEINSCHAFT

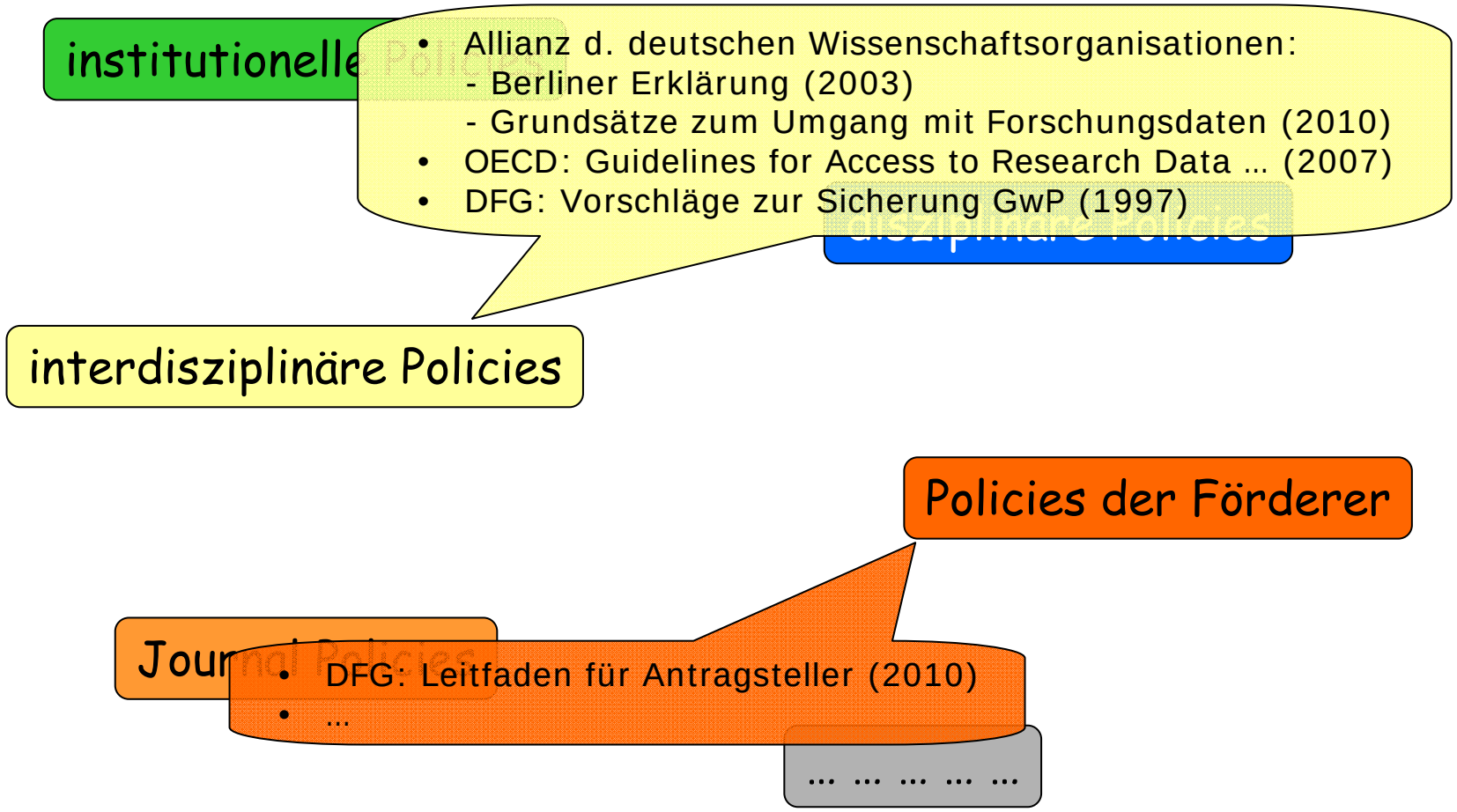
Data Science: From Big to Small

- **Bislang:** Fokus beim Umgang mit Forschungsdaten primär auf große Bestände, sog. „*Big Data Science*“. Wegen der großen Datenmengen gibt es hier (Projekt bedingt) immer ein Datenmanagement.
- „**Big Data Science**“ zeichnet sich durch große Datenmengen mit relativ homogenen Strukturen aus.
- **aber:** Ein sehr großer Aufwand an Zeit und intellektueller Leistung geht in eine Unzahl von kleinen Datensätzen, sog. „*Small Science Data*“.
  - ↳ Hier fehlt zumeist ein systematisches Datenmanagement!
- „**Small Data Science**“ zeichnet sich durch kleine Datenmengen mit heterogenen Strukturen aus.



(P. Vettiyattil, 2012)

Policies


institutionelle Policies
• Allianz d. deutschen Wissenschaftsorganisationen:
- Berliner Erklärung (2003)
- Grundsätze zum Umgang mit Forschungsdaten (2010)
• OECD: Guidelines for Access to Research Data … (2007)
• DFG: Vorschläge zur Sicherung GwP (1997)
disziplinäre Policies
interdisziplinäre Policies
Policies der Förderer
Journal Policies
• DFG: Leitfaden für Antragsteller (2010)
• …
… … … … …

Policy am GFZ

- DFG-Regeln für eine gute wissenschaftliche Praxis (GwP) wurden in eine Dienstanweisung überführt.
- Helmholtz-Gemeinschaft unterstützt den Offenen Zugang (Erstunterzeichner der Berliner Erklärung).
- GFZ-Publikationsordnung schließt die Regeln GwP und die Berliner Erklärung mit ein und ist Bestandteil des Arbeitsvertrages.

↳ **Daten müssten verfügbar sein, denn GwP und OA sind Bestandteile des Arbeitsvertrags, aber ...**

Der Geist ist willig, …

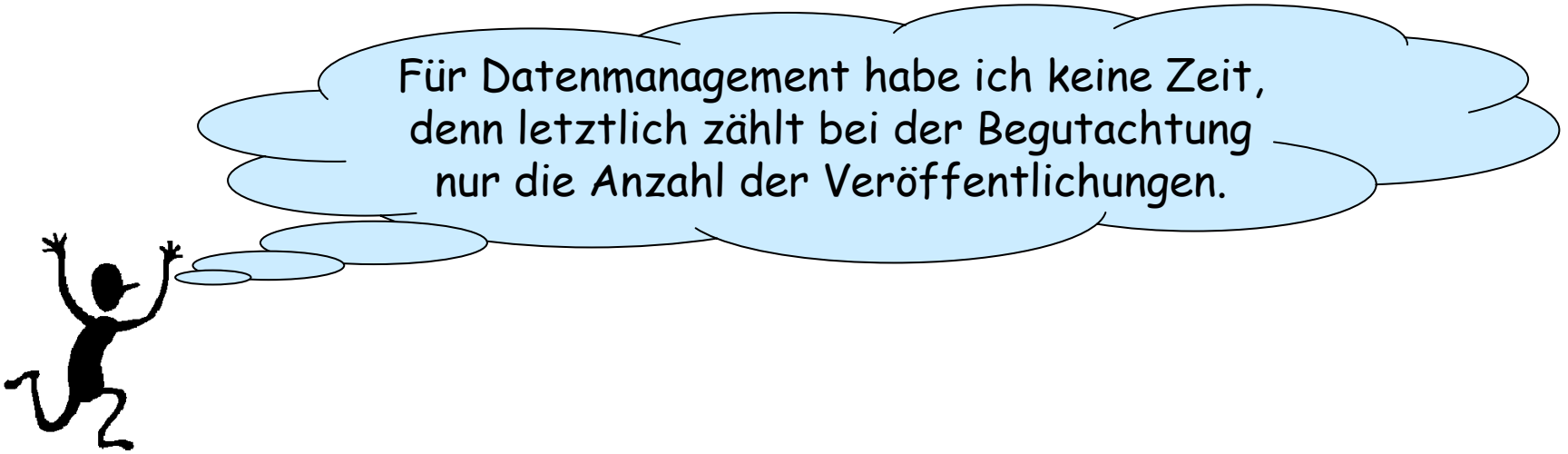
Wie lässt sich ein besserer Umgang mit Daten etablieren?

- Sanktionen greifen in diesem Umfeld nicht.
  ↳ Ohne Anreize wenig Hoffnung auf Umsetzung!

- Für die Umsetzung der Policies fehlen wichtige Bereiche:
  - Organisation und Integration in die wissenschaftlichen Arbeitsabläufe
  - Werkzeuge mit Integration in die wissenschaftlichen Arbeitsabläufe

- Heute ist der Umgang mit Daten oft nur Zweit- oder Drittaufgabe. Für den Einzelnen lohnt die investierte Zeit zumeist nicht!

Von Böcken und Gärtnern



Andererseits:

In eBay beschreiben jeden Tag tausende von Menschen Objekte mit Metadaten – und machen es (weitgehend) richtig.

↳ **Sind die Rollen der Akteure im Lebenszyklus von Forschungsdaten richtig verteilt?**

Umgang mit Forschungsdaten ...

... ist wie eine Herde Katzen
über die Prärie zu treiben.

Beispiel GFZ:

- ca. 150 Projekte gleichzeitig
- Fluktuation von ca. 30 Projekten pro Jahr
- Fluktuation von ca. 200 Forschern pro Jahr

↳ **Wir brauchen generische Werkzeuge!**

Beschreibung der Daten

- Ohne Beschreibung ist der Inhalt unbrauchbar.
- Wie kriege ich Metadaten ohne „Datenbürokratie“?

- Beschreibungen der Daten sollten als Teil des Arbeits-ablaufs (teil-)automatisch erfasst werden.

Warum funktioniert es außerhalb der Wissenschaft?

- Ebay:
  - Unter hohem Zeitaufwand werden Metadaten beschrieben, um Waren zu verkaufen.
  - Motivation: Geld

- Facebook:
  - Bereitwillig veröffentlichen Menschen ungeheure Mengen an persönlichen Daten.
  - Motivation: Selbstdarstellung, Vernetzung

- Wikipedia:
  - Unter Verzicht auf Autorennennung werden Inhalte publiziert und bearbeitet.
  - Motivation: Aussicht, Wikipedia-Administrator zu werden

… aber manchmal funktioniert es auch in der Wissenschaft:

Institutionelle Workflows

Beispiel ZALF:

- Anforderungen für Laborleistungen werden online eingegeben. Damit sind die Metadaten des Experiments bereits erfasst.

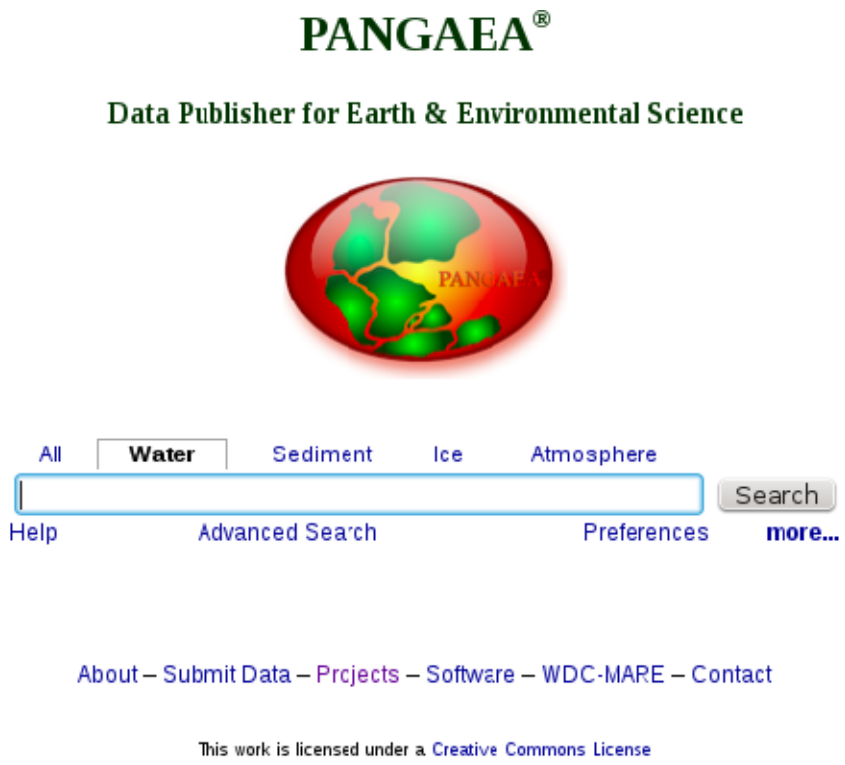
Beispiel Rutherford Appleton Laboratory:

- Anträge für Instrumentenzeit werden online gestellt. Damit sind die Metadaten des Experiments bereits erfasst.

Beispiel Uni Bremen:

- Absolventen der Geologie müssen ihre Daten in PANGAEA ablegen, um ihre Urkunde zu bekommen.

Beispiel: PANGAEA (www.pangaea.de)



- Daten werden nicht sortiert, sondern gesucht.
- Daten werden bei Aufnahme in den Suchindex mit einem DOI versehen.
- Derzeit keine Aufnahme operationell erweiterter, d.h. laufend wachsender Datensätze.

PANGAEA:
Vernetzung von Literatur und Daten

Vision und Praxis

- **Vision** der Wissenschaftsorganisationen:
  „Forschungsdaten sollten frei und überregional zugänglich und langfristig verfügbar sein.“

- **Zweck:** Offener und dauerhafter Zugang …
  - begünstigt neue Erkenntnisse und Forschungsansätze,
  - erlaubt neue Interpretationen,
  - ermöglicht Synthese von Forschungsergebnissen unterschiedlicher Bereiche,

  ↳ beschleunigt damit den Erkenntnisgewinn.

- **aber:**
  - Datenmanagement kostet Geld (, die der Forschung verlorengeht?).
  - Daten in einen neuen Kontext zu setzen kostet Zeit.
  - Datenbereitstellung in Repositorien ist oftmals noch zu kompliziert.
  - *„Meine Daten gehören mir!“*

Herausforderungen

- "Temporäre Wissenschaftler generieren einen großen Teil der Forschungsdaten." (und sind dann weg …)
- Integration des Datenmanagements in Arbeitsabläufe
- Spannungsfeld „Big Data Science" vs. „Small Data Science"
- Interoperabilität disziplinspezifischer, internationaler und Disziplinen übergreifender Strukturen
- Datenmanagement als allg. Ausbildungskomponente?
- Schaffung eines Anreizsystems
- kultureller Wandel: Verfügbarkeit von Forschungsdaten als Selbstverständlichkeit

123RF Limited

Wünsche

- Sichtbarmachung des wechselseitigen „*Benefits*“
- Datenmanagement, Werkzeuge und Dienste an typische Arbeitsabläufe der Disziplinen anpassen
- generische Werkzeuge, die leicht übertragbar sind (z.B. innerhalb temporärer Projekte)
- flexible Interfaces zur Anpassung an Bedürfnisse und Vorkenntnisse des Nutzers
- Transparenz: *Was geschieht mit meinen Daten?*
- Balance von Informationswissenschaft und Service

Anreize:
Von Ehre, Ruhm und Reichtum

- **Reputation**:
  - Schaffung einer Datenpublikationskultur
  - Integration in Reputationssystem (h-Index, …)
- **neue Professionen**:
  - Überbrückung der Diskontinuität zwischen Wissenschaft und Informationsmanagement
- **Bereitstellung wissenschaftsnaher Werkzeuge**:
  - generische (und verständliche) Interfaces, virtuelle Forschungsumgebungen, …
- **Service-orientierte Infrastrukturen**

↳ Anerkennung der Datenbereitstellung zur Nachnutzung als selbstverständliche Komponente im Wissenschaftsbetrieb

GFZ
Helmholtz-Zentrum
POTSDAM

HELMHOLTZ
GEMEINSCHAFT