Research Data Infrastructures
- Challenges, Desires, Incentives -

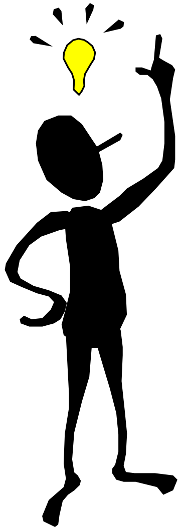
Maik Thomas & Jens Klump

Helmholtz Centre Potsdam
GFZ German Research Centre for Geosciences
Potsdam, Germany

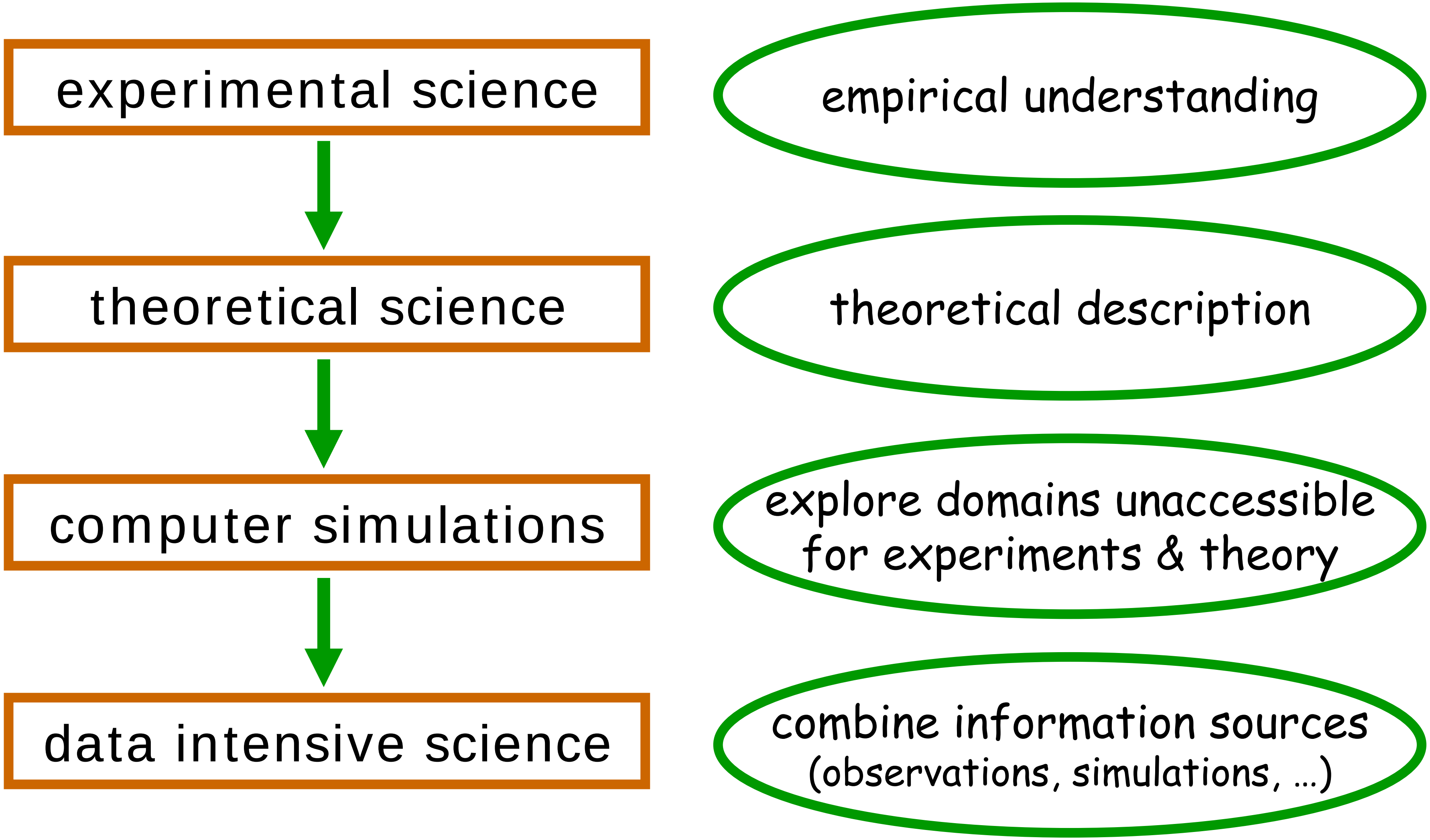
Wikipedia‘s definition

- **Data infrastructure** is a digital infrastructure promoting data sharing and consumption.

  Similarly to other infrastructures, it is a structure needed for the operation of a society as well as the services and facilities necessary for an economy to function, the data economy in this case.



Paradigms of science



experimental science

theoretical science

computer simulations

data intensive science

empirical understanding

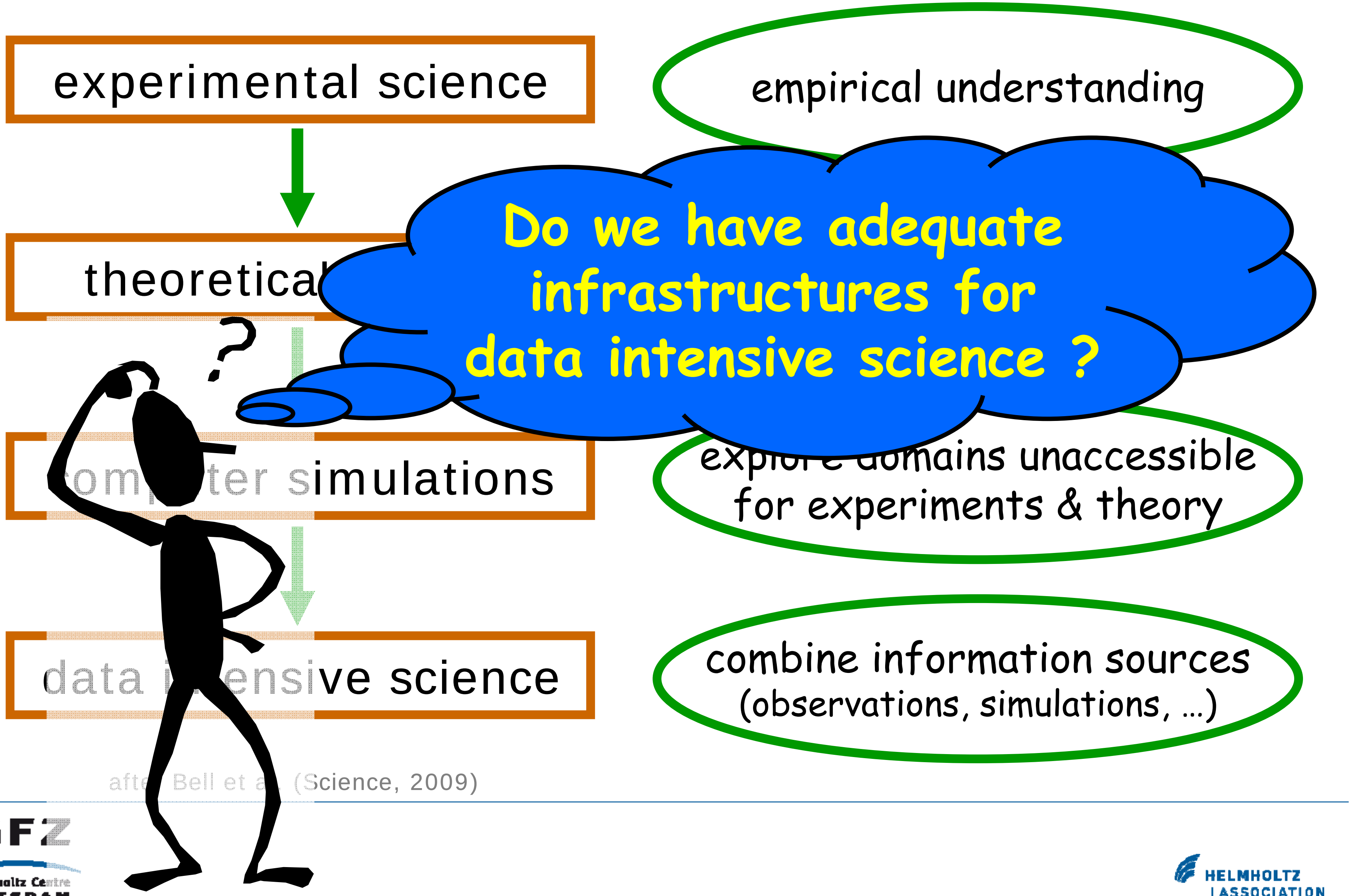
theoretical description

explore domains unaccessible for experiments & theory

combine information sources (observations, simulations, …)

after Bell et al. (Science, 2009)

Paradigms of science



experimental science

empirical understanding

theoretical

computer simulations

explore domains unaccessible for experiments & theory

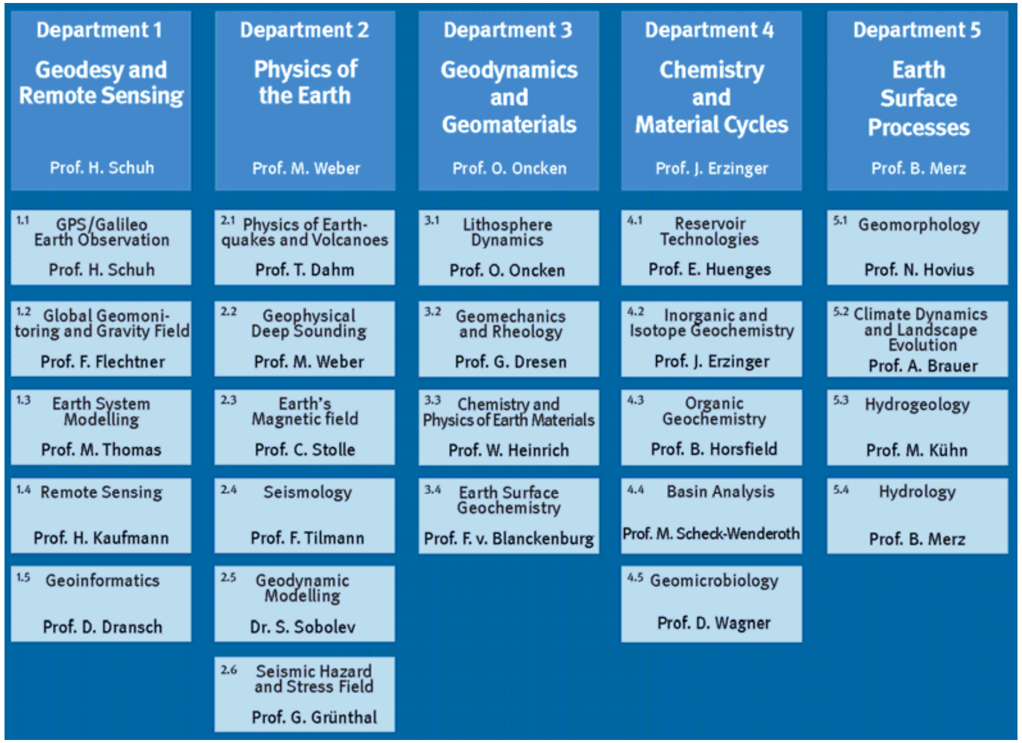
data intensive science

combine information sources (observations, simulations, ...)

?
Do we have adequate infrastructures for data intensive science ?

after Bell et al. (Science, 2009)

Example:
German Research Centre for Geosciences (GFZ)



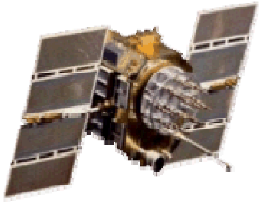
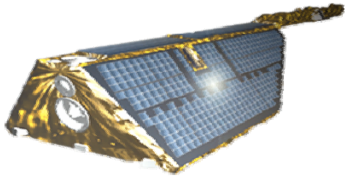
Department 1 Geodesy and Remote Sensing Prof. H. Schuh	Department 2 Physics of the Earth Prof. M. Weber	Department 3 Geodynamics and Geomaterials Prof. O. Oncken	Department 4 Chemistry and Material Cycles Prof. J. Erzinger	Department 5 Earth Surface Processes Prof. B. Merz
1.1 GPS/Galileo Earth Observation Prof. H. Schuh	2.1 Physics of Earthquakes and Volcanoes Prof. T. Dahm	3.1 Lithosphere Dynamics Prof. O. Oncken	4.1 Reservoir Technologies Prof. E. Huenges	5.1 Geomorphology Prof. N. Hovius
1.2 Global Geomonitoring and Gravity Field Prof. F. Flechtner	2.2 Geophysical Deep Sounding Prof. M. Weber	3.2 Geomechanics and Rheology Prof. G. Dresen	4.2 Inorganic and Isotope Geochemistry Prof. J. Erzinger	5.2 Climate Dynamics and Landscape Evolution Prof. A. Brauer
1.3 Earth System Modelling Prof. M. Thomas	2.3 Earth's Magnetic field Prof. C. Stolle	3.3 Chemistry and Physics of Earth Materials Prof. W. Heinrich	4.3 Organic Geochemistry Prof. B. Horsfield	5.3 Hydrogeology Prof. M. Kühn
1.4 Remote Sensing Prof. H. Kaufmann	2.4 Seismology Prof. F. Tilmann	3.4 Earth Surface Geochemistry Prof. F. v. Blanckenburg	4.4 Basin Analysis Prof. M. Scheck-Wenderoth	5.4 Hydrology Prof. B. Merz
1.5 Geoinformatics Prof. D. Dransch	2.5 Geodynamic Modelling Dr. S. Sobolev		4.5 Geomicrobiology Prof. D. Wagner	
	2.6 Seismic Hazard and Stress Field Prof. G. Grünthal			

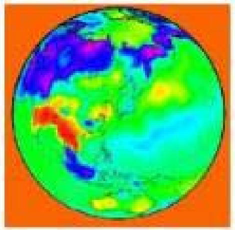
- 5 research areas (Departments) each with 4-6 sections and more than 1000 employees
- several research projects in every section
- more than 100 different data sources

Example: Data sources at GFZ Potsdam

Many data sets come from …

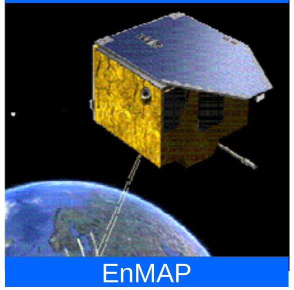
- **Earth monitoring systems**
  - generate large amounts of data
  - mostly homogeneous structures
  - partly automated workflows
- **Modelling systems**
  - produce medium-sized data amounts
  - partly homogeneous structures
  - but mostly no supporting workflows
- **Labs and field observations**
  - in most cases only small data amounts
  - but heterogeneous structures
  - data sets often generated in ad-hoc created structures



The Earth system: Monitoring, Analysis, Simulation

Hyperspectral mission

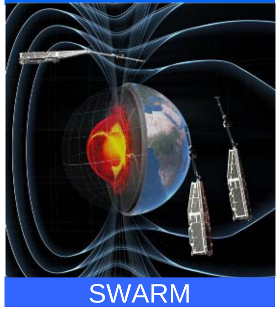


EnMAP

Terrestrial networks

Tide gauge GPS SG Geophone

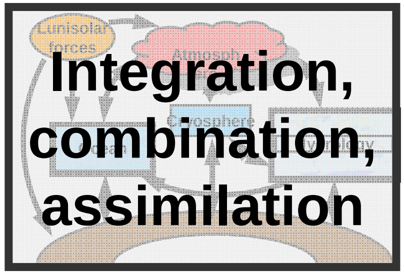
Magnetic field



SWARM

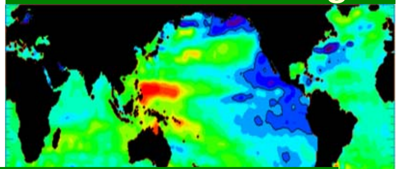
Gravimetry missions

GRACE-FO

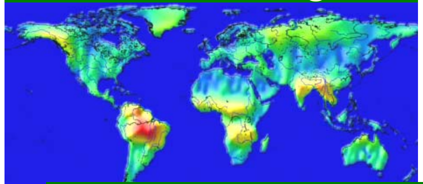


Integration, combination, assimilation

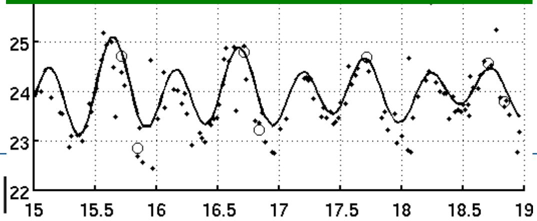
Sea level change



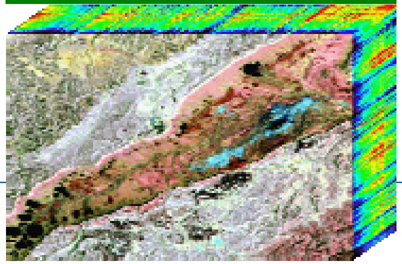
Water storage



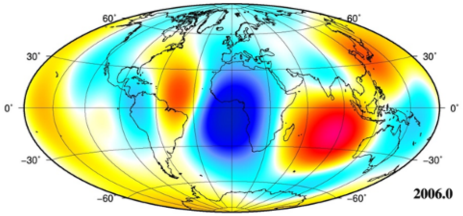
Ocean tides



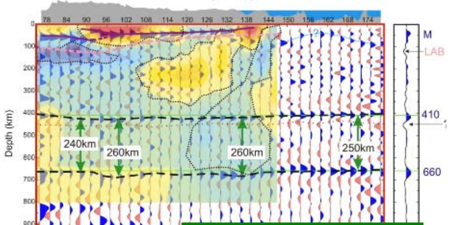
Surface parameters



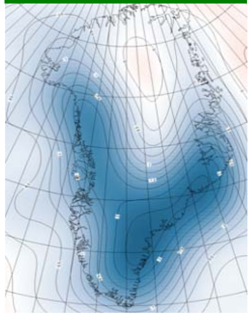
Magnetic field variations



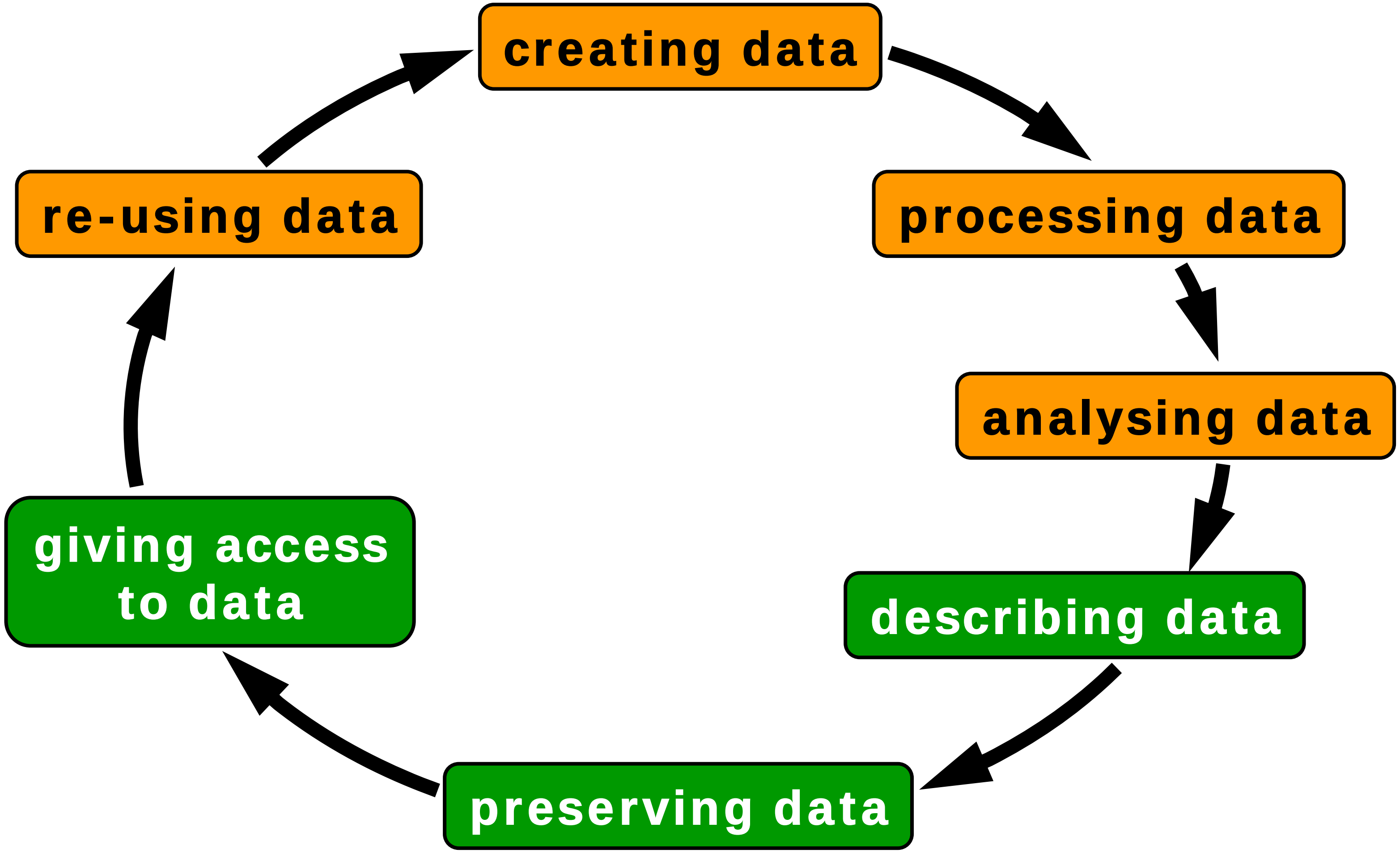
Lithospheric structure



Ice mass change

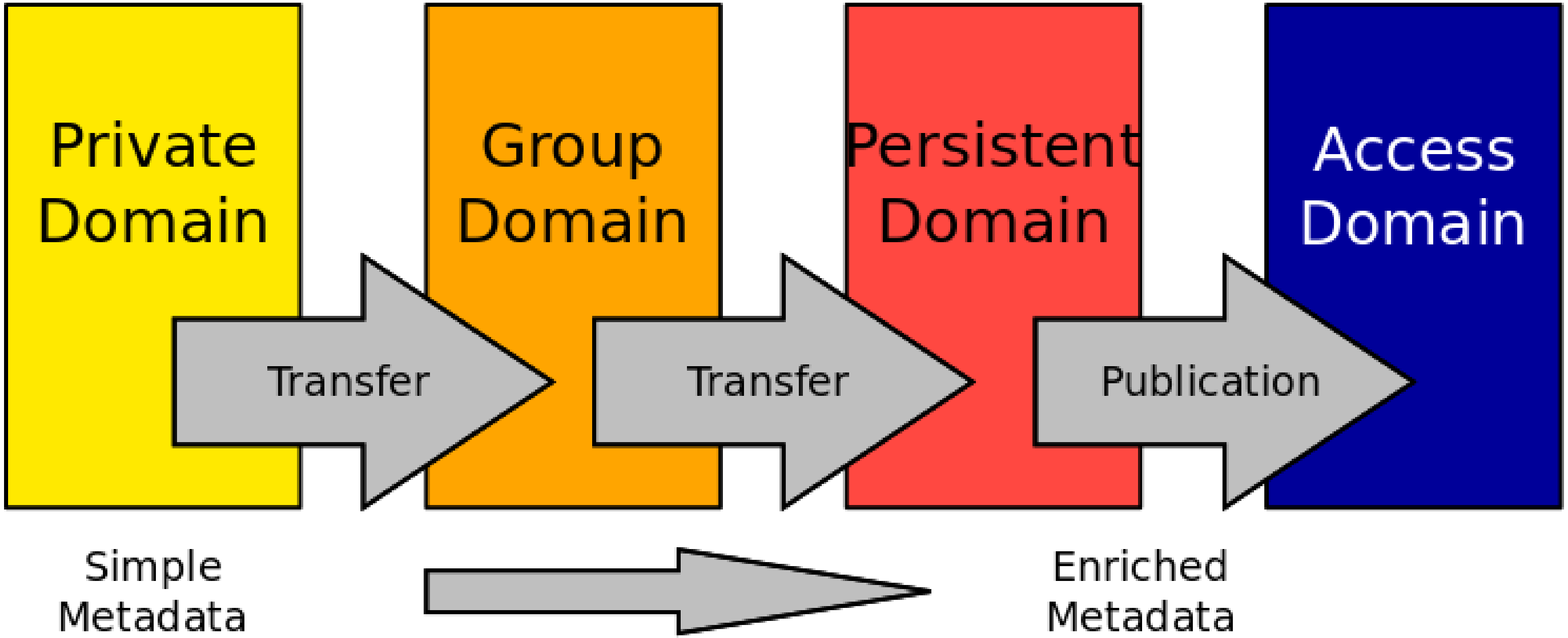


Data Life Cycle Model



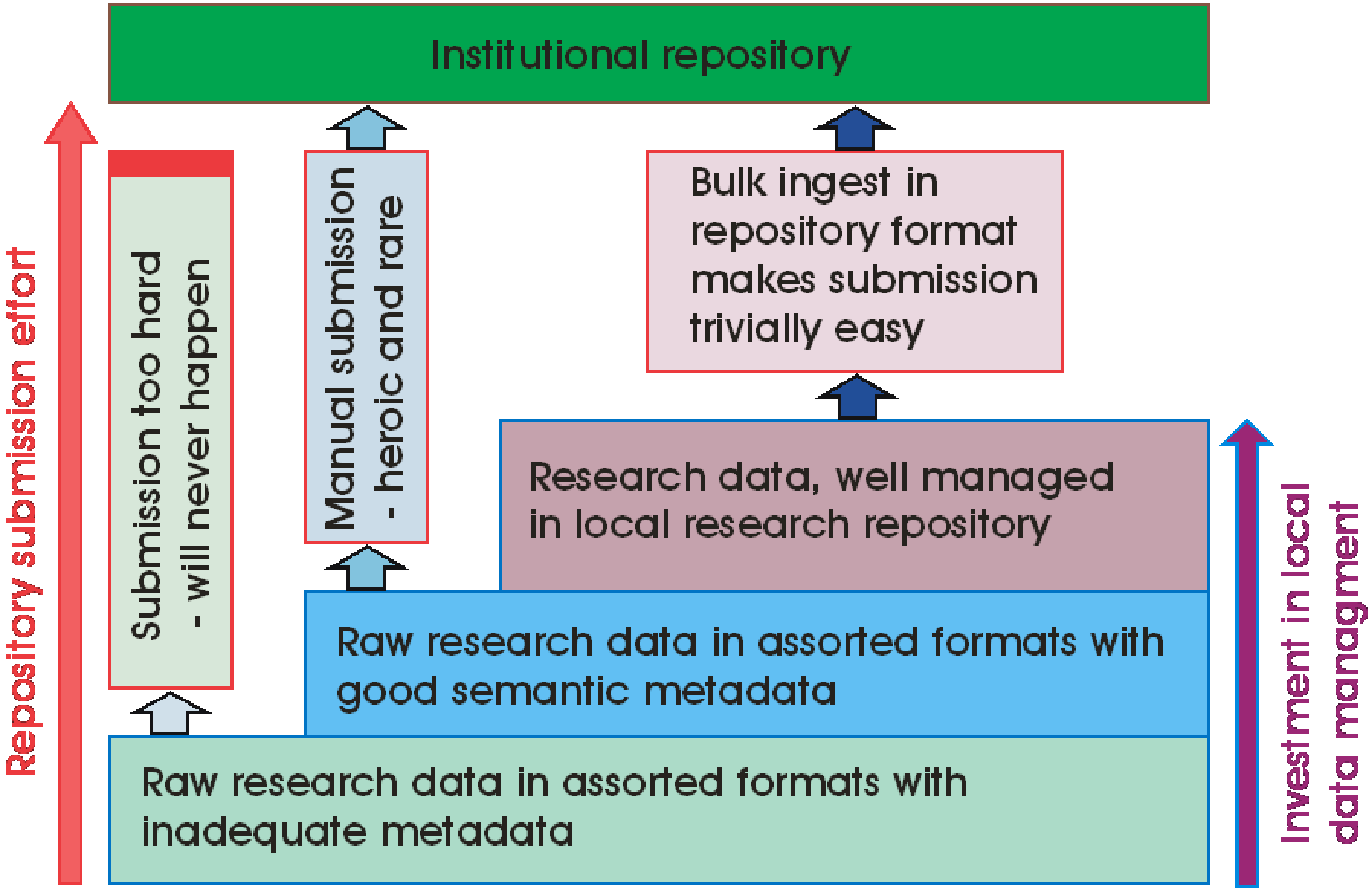
… defines demands on data management !

Data Life Cycle Model: From private to public



↳ **How to provide research data sets to external users ?**

Submitting data to repositories



Institutional repository

Repository submission effort

Submission too hard - will never happen

Manual submission - heroic and rare

Bulk ingest in repository format makes submission trivially easy

Research data, well managed in local research repository

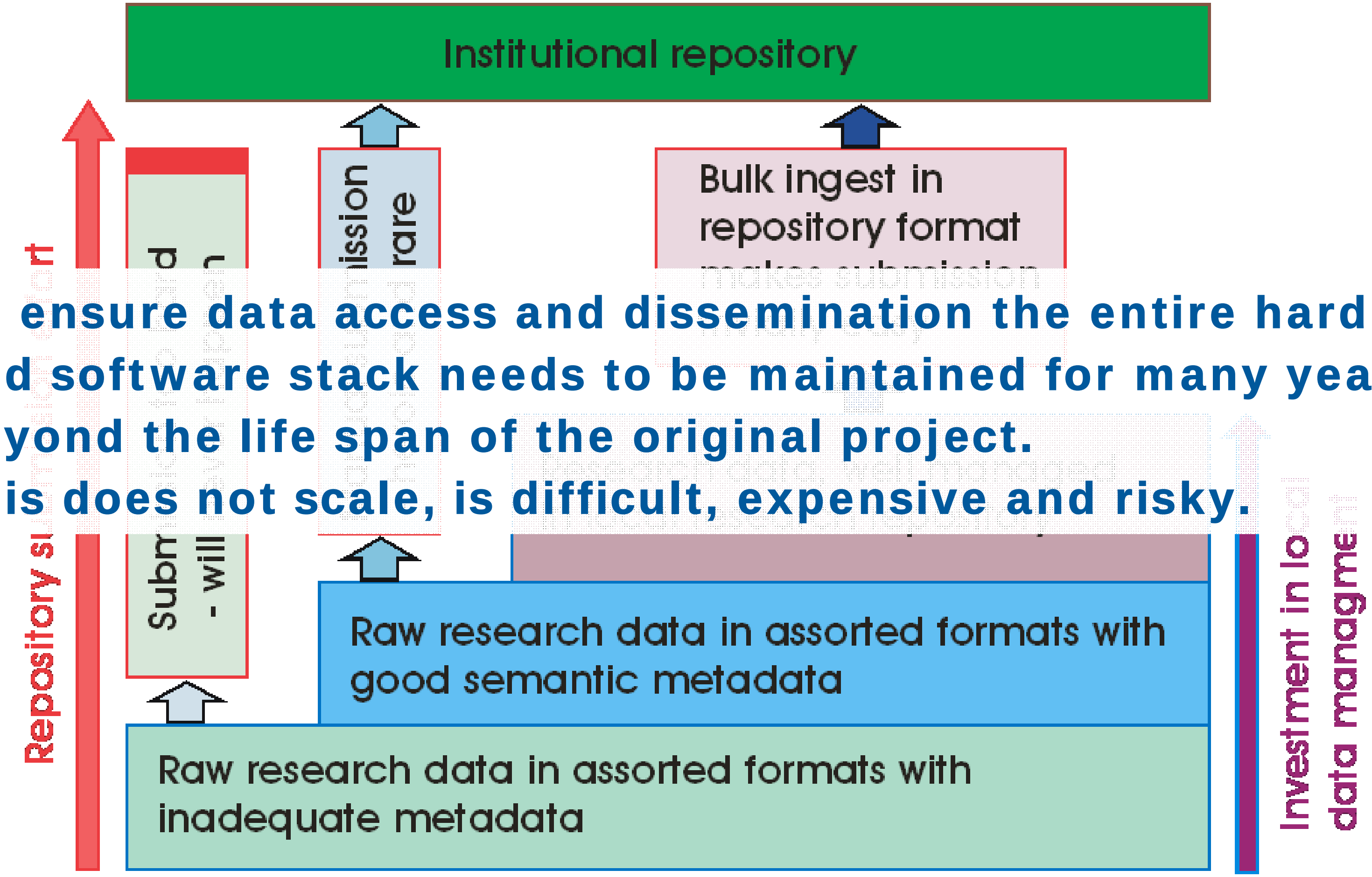
Raw research data in assorted formats with good semantic metadata

Raw research data in assorted formats with inadequate metadata

Investment in local data managment

D. Shotton, opencitations.wordpress.com

Submitting data to repositories

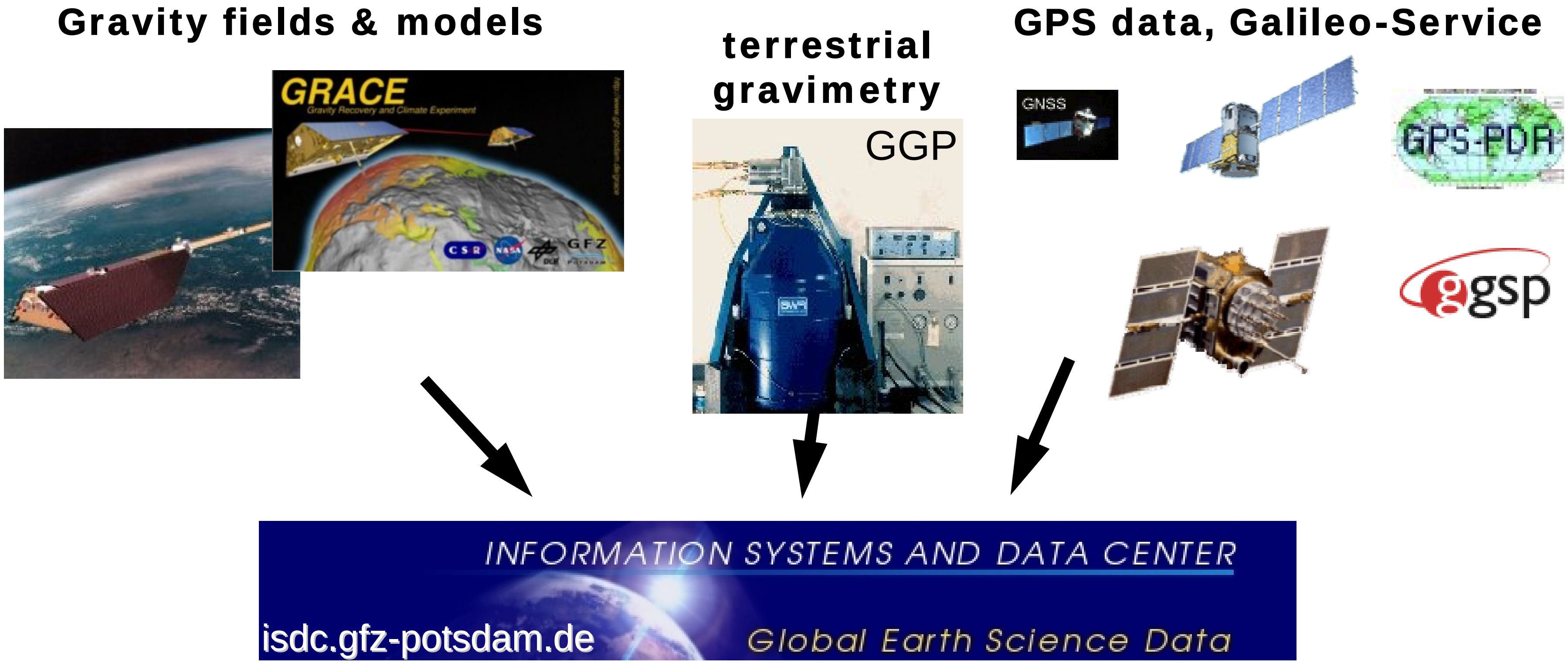


Institutional repository
Bulk ingest in repository format makes submission
Raw research data in assorted formats with good semantic metadata
Raw research data in assorted formats with inadequate metadata
Repository submission effort
Subm
- will
ission
rare
Investment in local data managment

D. Shotton, opencitations.wordpress.com

- To ensure data access and dissemination the entire hardware and software stack needs to be maintained for many years beyond the life span of the original project.
This does not scale, is difficult, expensive and risky.

An isolated application:
Information System and Data Center (ISDC)



Gravity fields & models

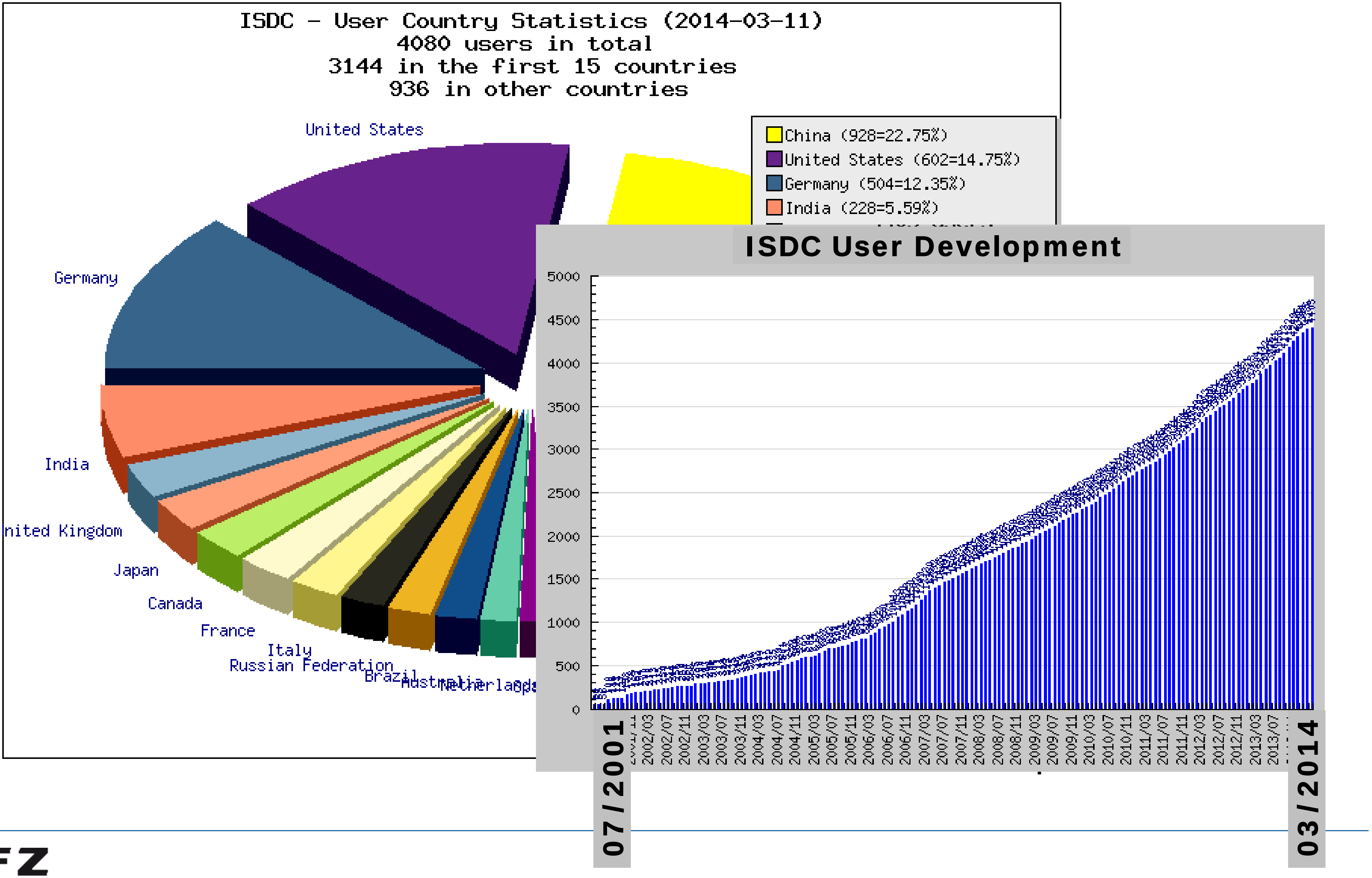
terrestrial gravimetry

GPS data, Galileo-Service

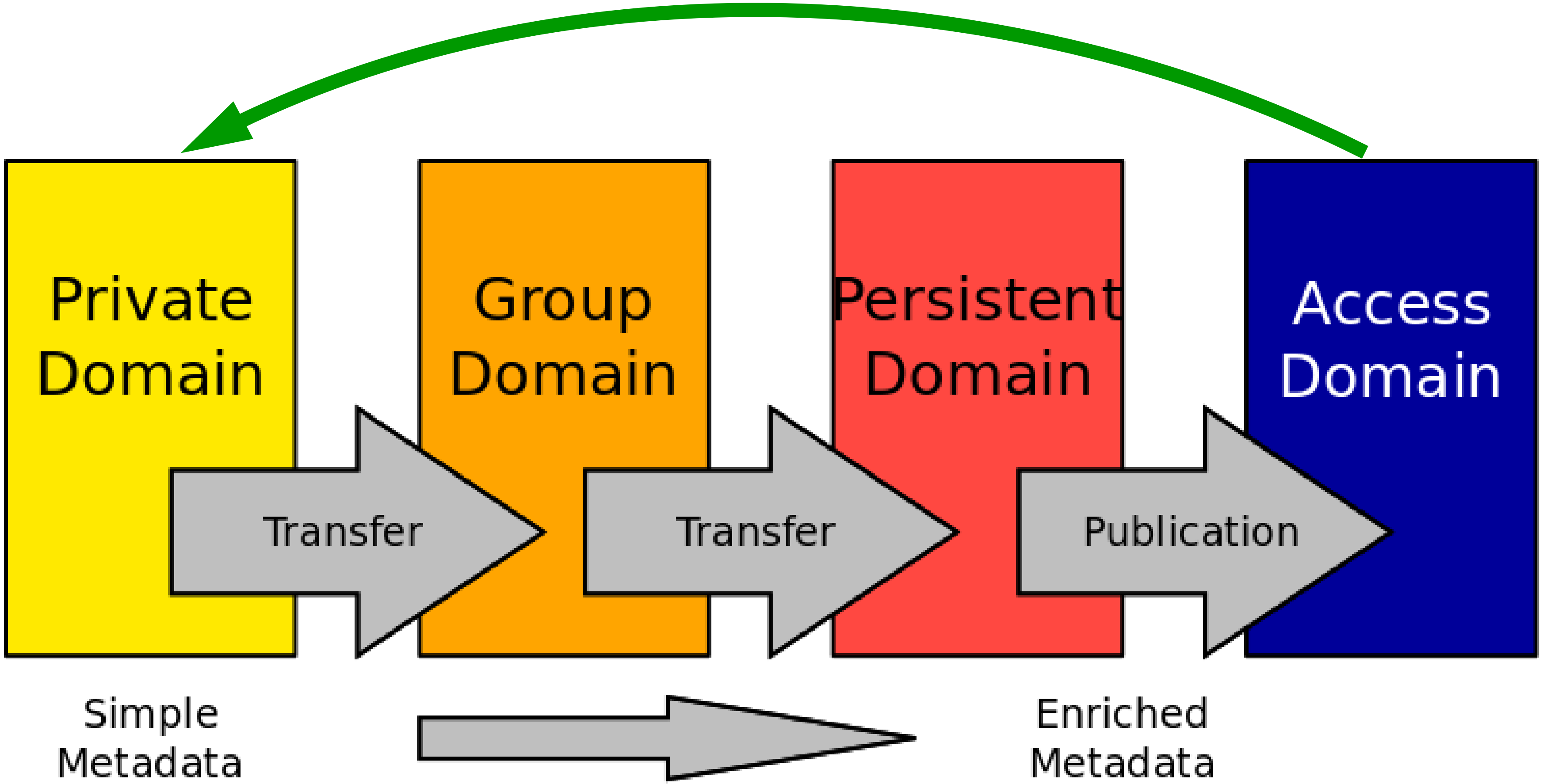
GRACE
Gravity Recovery and Climate Experiment
CSR
NASA
GFZ
GGP
GNSS
GPS-PDR
ggsp
INFORMATION SYSTEMS AND DATA CENTER
isdc.gfz-potsdam.de
Global Earth Science Data

↳ various data products with corresponding meta data, error estimates FAQs and user support, assigned to major research projects

ISDC: Statistics



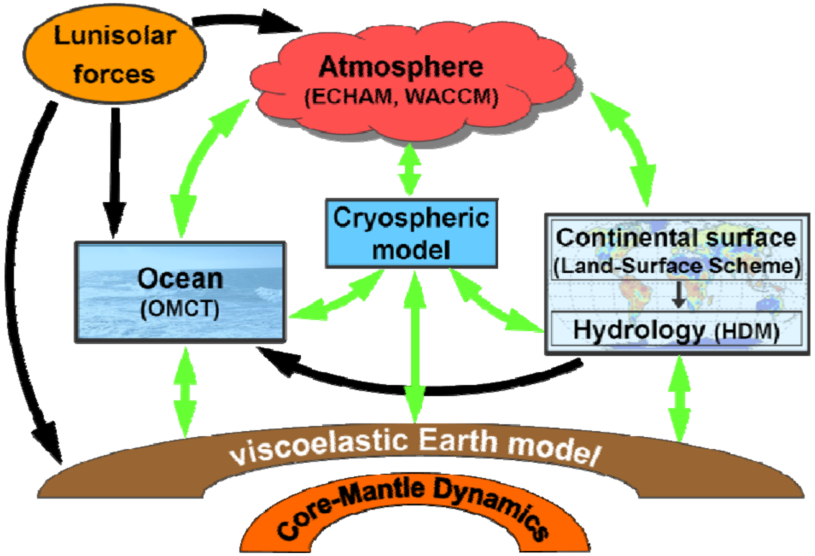
Data Life Cycle Model



↳ **What about the integration of data from neighboring research areas ?**

From multi- via inter- to transdisciplinarity

- **differentiation of disciplines** due to scientific specialization over the last decades
- significant **scientific progress** is presently expected from research across specialized disciplines
- recent **IT developments** principally provide excellent conditions **for rapprochement of scientific disciplines** (and research cultures ?)



Example:

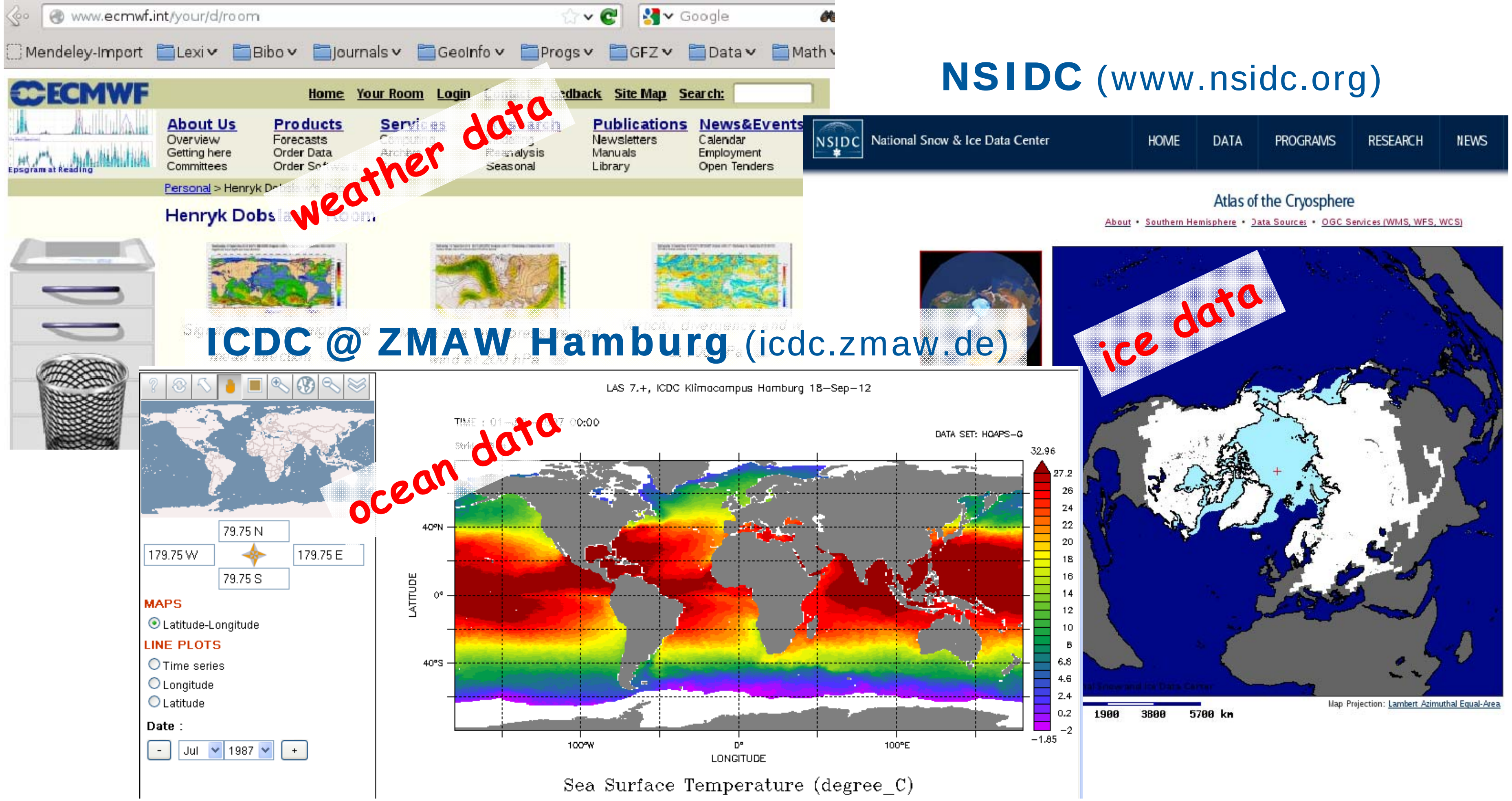
climate change and its socio-economic impacts

The geoscientific archipelago

MARS @ ECMWF (www.ecmwf.int)

NSIDC (www.nsidc.org)

ICDC @ ZMAW Hamburg (icdc.zmaw.de)



www.ecmwf.int/your/d/room

Mendeley-Import Lexi Bibo Journals GeoInfo Progs GFZ Data Math

ECMWF Home Your Room Login Contact Feedback Site Map Search:

About Us: Overview, Getting here, Committees

Products: Forecasts, Order Data, Order Software

Services: Computing, Archive

Research: Modelling, Reanalysis, Seasonal

Publications: Newsletters, Manuals, Library

News&Events: Calendar, Employment, Open Tenders

Personal > Henryk Dobslaw's Room

Henryk Dobslaw's Room

National Snow & Ice Data Center HOME DATA PROGRAMS RESEARCH NEWS

Atlas of the Cryosphere

About • Southern Hemisphere • Data Sources • OGC Services (WMS, WFS, WCS)

Map Projection: Lambert Azimuthal Equal-Area

LAS 7.+, ICDC Klimacampus Hamburg 18–Sep–12

DATA SET: HOAPS–G

79.75 N | 179.75 W | 179.75 E | 79.75 S

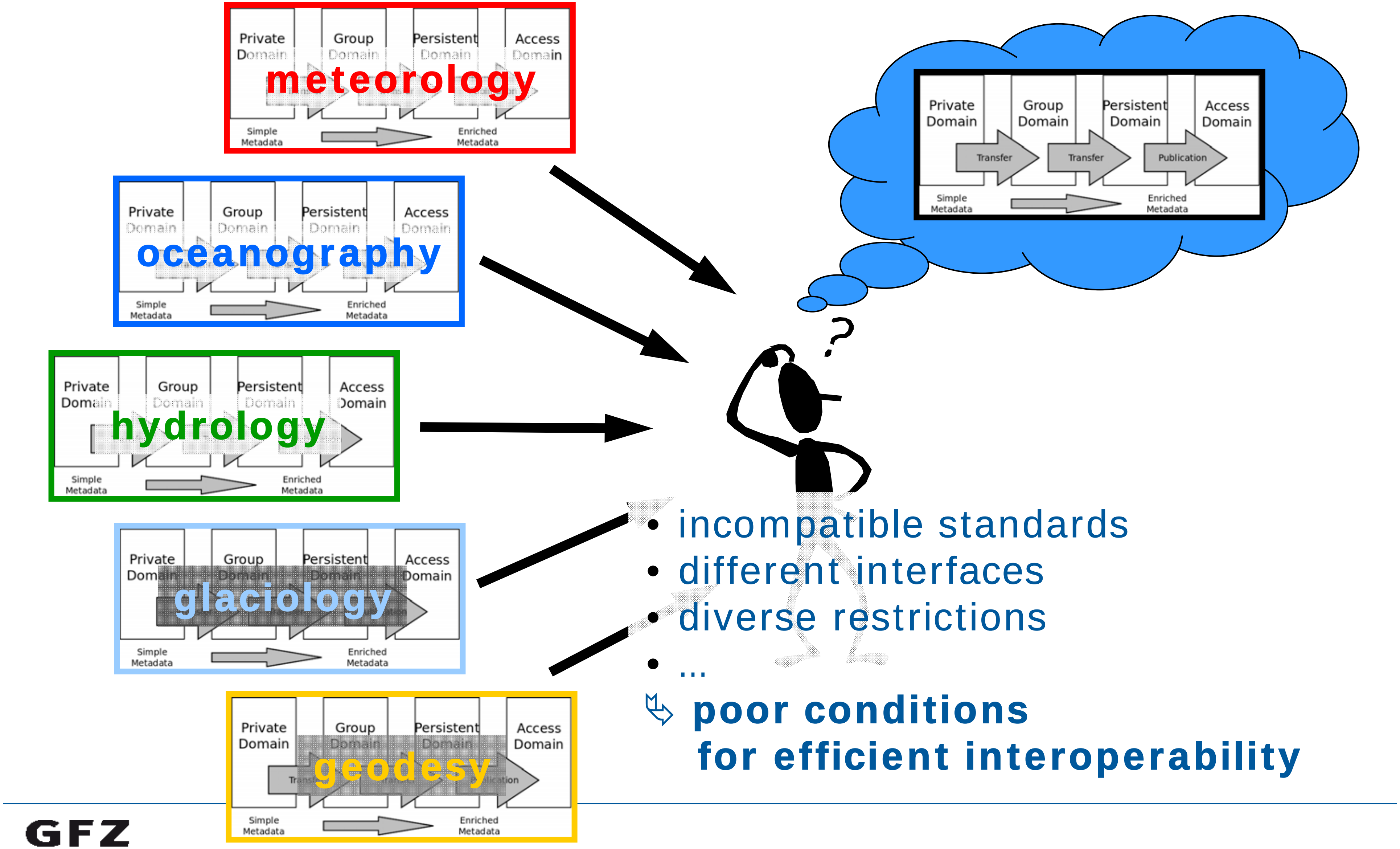
MAPS
● Latitude-Longitude

LINE PLOTS
○ Time series
○ Longitude
○ Latitude

Date : Jul 1987

Sea Surface Temperature (degree_C)

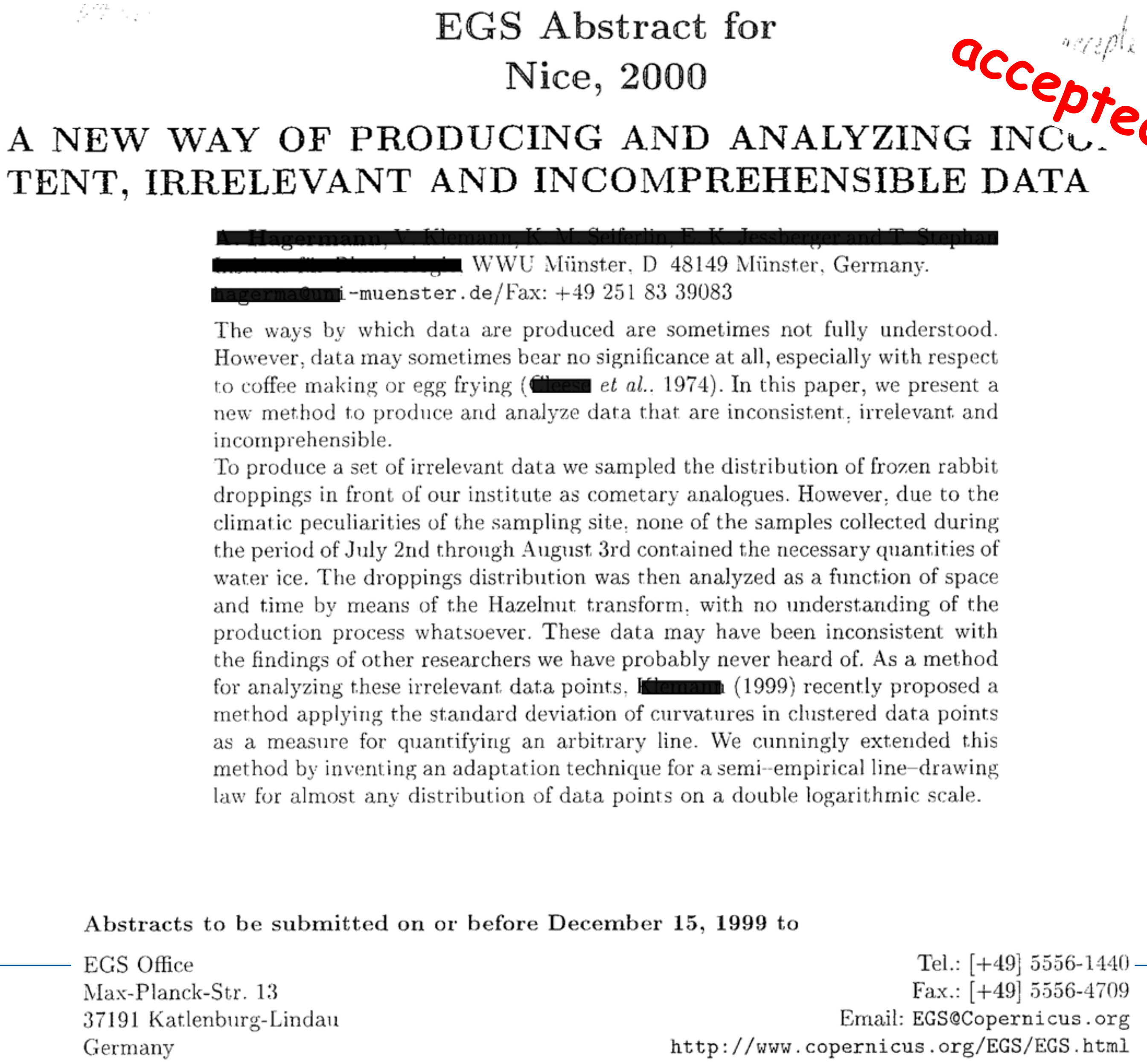
Many communities and even more „solutions“ …



- incompatible standards
- different interfaces
- diverse restrictions
- …

↳ **poor conditions for efficient interoperability**

Expected Results?



EGS Abstract for
Nice, 2000

accepte

accepted !

A NEW WAY OF PRODUCING AND ANALYZING INCO-TENT, IRRELEVANT AND INCOMPREHENSIBLE DATA

[redacted]
[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.

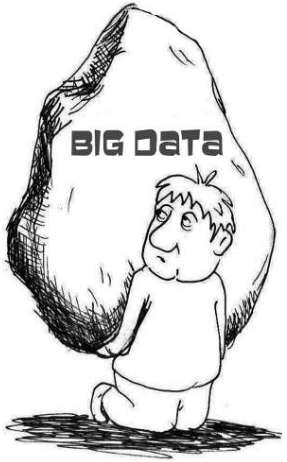
Abstracts to be submitted on or before December 15, 1999 to

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

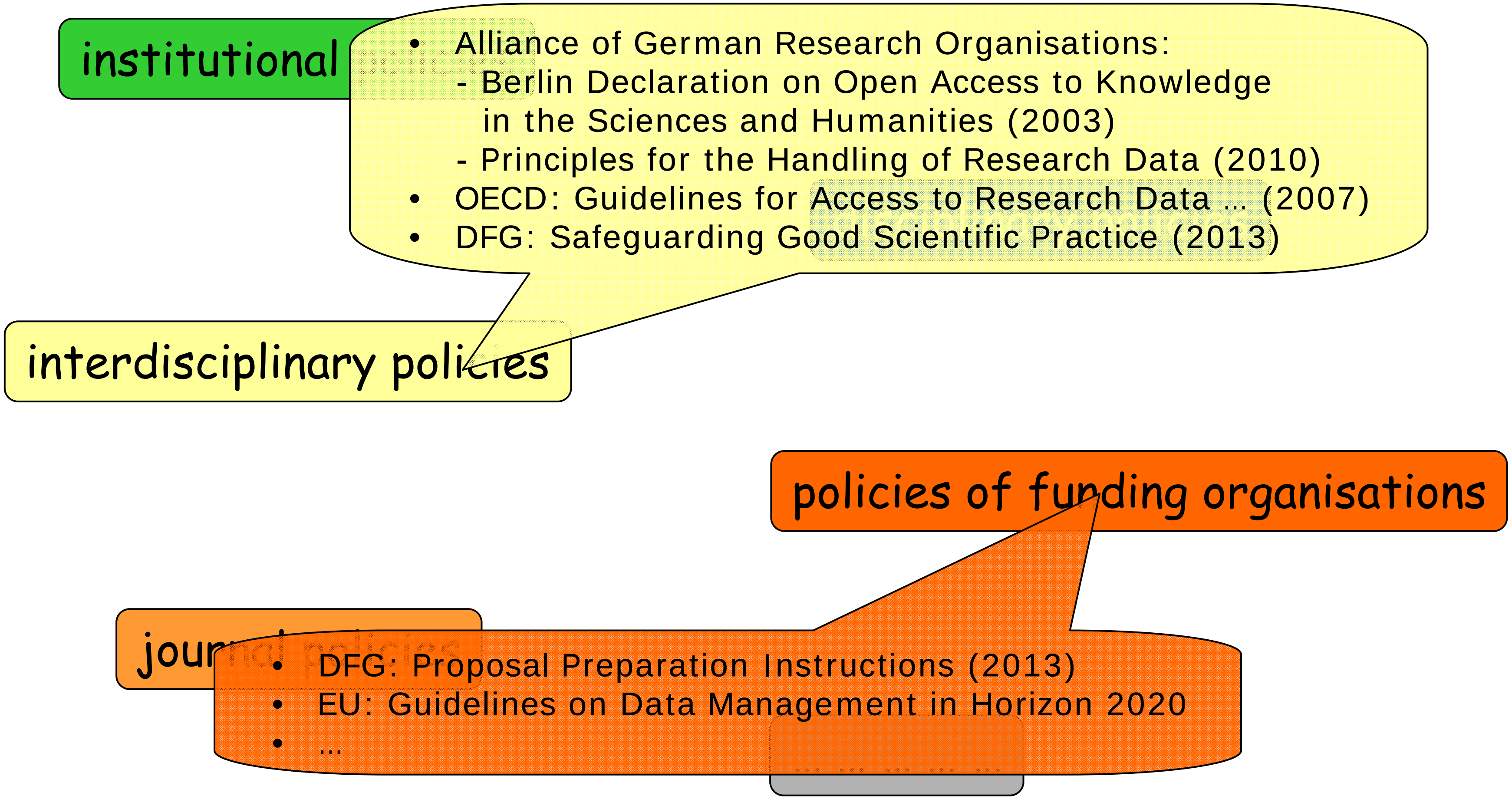
Data Science: From Big to Small

- **Usually:** Rules for data handling focus on large data sets, i.e., *„Big Data Science“*.
  ↳ Existing data management due to regulations of funding bodies.
- **„Big Data Science“:** large data amounts, homogeneous structures
- **But:** Vast investments of time (and brain work !) to generate an immense number of small data sets, i.e., *„Small Science Data“*.
  ↳ Systematic data management missing !
- **„Small Data Science“:** small data amounts, heterogeneous structures



(P. Vettiyattil, 2012)

Policies



institutional policies

- Alliance of German Research Organisations:
  - Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities (2003)
  - Principles for the Handling of Research Data (2010)
- OECD: Guidelines for Access to Research Data … (2007)
- DFG: Safeguarding Good Scientific Practice (2013)

disciplinary policies

interdisciplinary policies

policies of funding organisations

journal policies

- DFG: Proposal Preparation Instructions (2013)
- EU: Guidelines on Data Management in Horizon 2020
- …

… … … … …

Example: Policy at GFZ

- DFG-Rules of Good Scientific Practice (GSP) are standing instructions.
- Helmholtz-Association supports Open Access initiative (1st signee of Berlin Declaration!).
- GFZ‘s publication policy includes GSP and Berlin Declaration on Open Access and is part of the employment contract.

↳ *Since GSP and Open Access are components of employment contracts,* ***data should be accessible.*** ***But*** *…*

The spirit is willing …

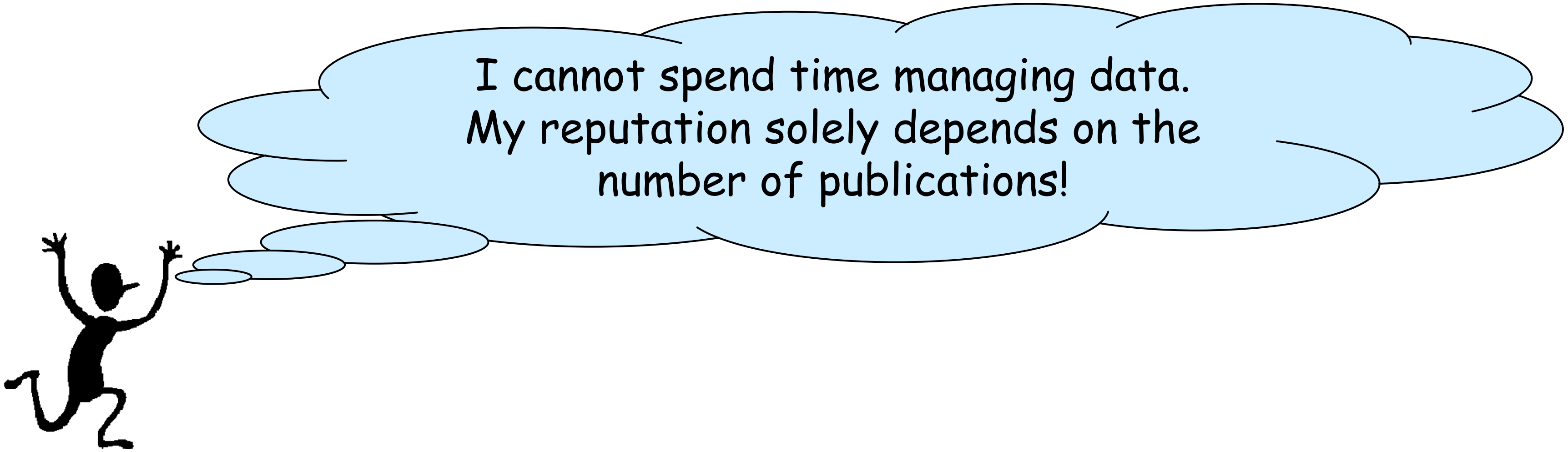
How can responsible data handling be established ?

- Imposition of sanctions fail in this context.
  ⇘ Incentives are indispensable for realization !

- Implementation of policies requires:
  - Organization of scientific workflows
  - Tools allowing integration into scientific workflows

- Currently, data handling is not a primary task of scientists. For the individual scientist responsible **data handling is more trouble than it‘s worth !**

Trusting the cat to keep the cream …


I cannot spend time managing data.
My reputation solely depends on the
number of publications!

On the other hand:

Every day several thousand eBay users add metadata to objects - and they do it correctly.

↳ **Are the tasks of players in data life cycles correctly assigned?**

Research data handling …

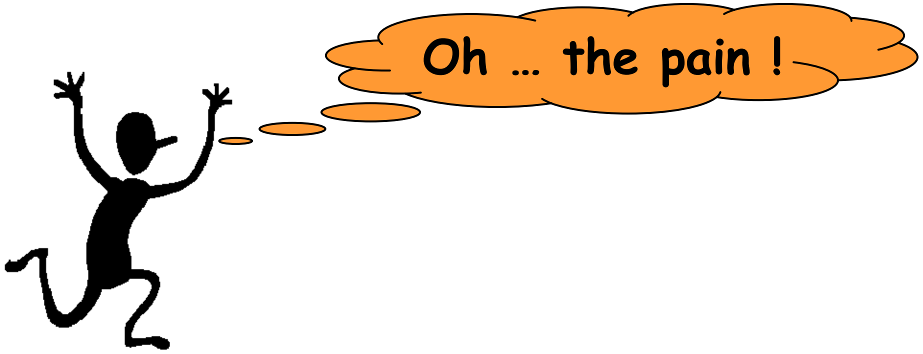
… is like driving a flock of cats over plains.

Example GFZ:

- ca. 150 simultaneous research projects
- annual fluctuation of …
  … research projects: ≈ 30
  … employees: ≈ 200

↳ **We need generic tools !**

Describing data



- Content useless without description
- How can I get metadata without sprawling „data bureaucracy“ ?

- Collect data descriptions (semi-)automatically as part of the workflow

Why does it work in other areas?

- Ebay:
  - In order to sell goods people spend much time on describing metadata.
  - motivation: money

Bild im Bild Studio - Fotolia

- Facebook:
  - People are willing to make vast amounts of personal data public.
  - motivation: self-expression, networking

- Wikipedia:
  - People write and edit articles on specific topics without being acknowledged as authors.
  - motivation: prospect of becoming a Wikipedia administrator

... sometimes it also works in the scientific world:

Institutional Workflows

Examples:

ZALF (Leibniz Centre for Agricultural Landscape Research)

- Requests for lab performances are fed online.
  ↳ Metadata of experiments are already recorded.

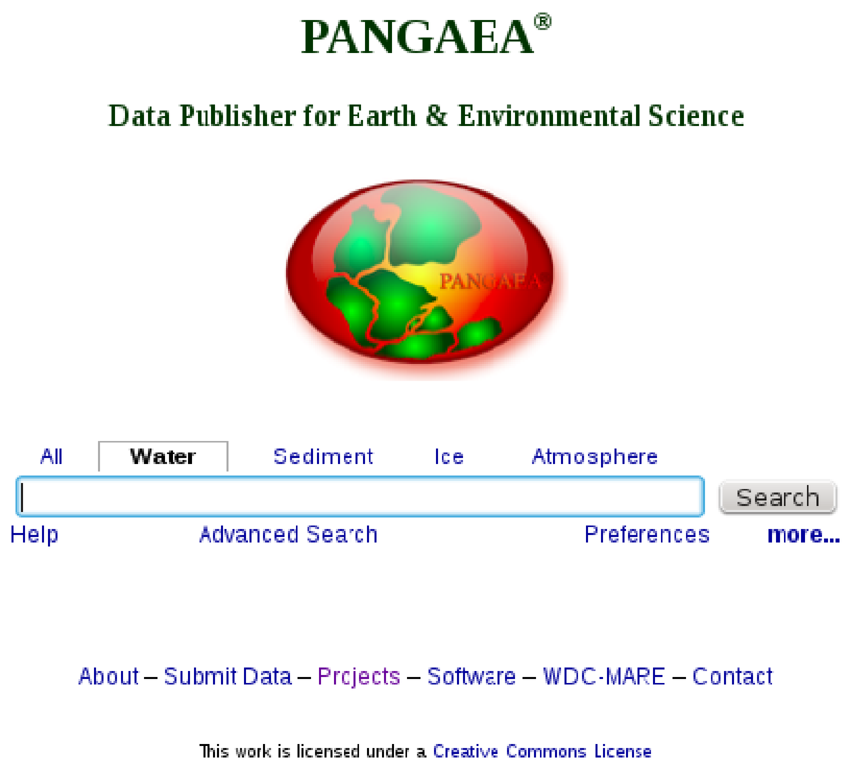
Rutherford Appleton Laboratory

- Applications for using instruments are made online.
  ↳ Metadata of experiments are automatically recorded.

University of Bremen

- Graduates in geology have to store their research data in the database PANGAEA before the certificate is delivered.

Example: PANGAEA (www.pangaea.de)



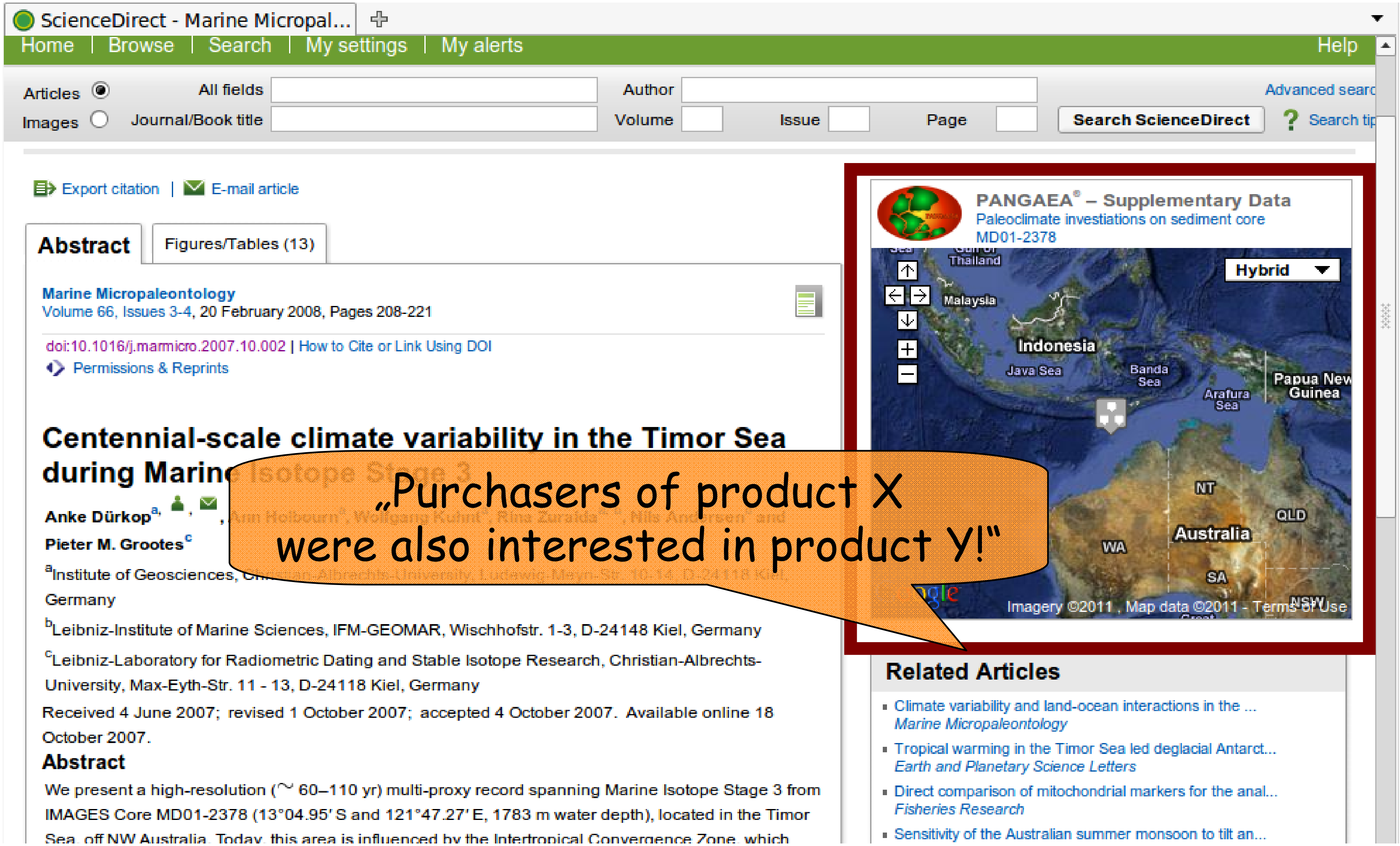
PANGAEA®

Data Publisher for Earth & Environmental Science

This work is licensed under a Creative Commons License

- Data sets are not sorted, but searched for.
- At the time of incorporation a DOI is added to the data sets.
- No incorporation of operationally increasing data, so far.

PANGAEA:
Cross-linking papers with research data



Marine Micropaleontology
Volume 66, Issues 3-4, 20 February 2008, Pages 208-221

doi:10.1016/j.marmicro.2007.10.002 | How to Cite or Link Using DOI
Permissions & Reprints

Centennial-scale climate variability in the Timor Sea during Marine Isotope Stage 3

Anke Dürkop[a], ..., Ann Holbourn[a], Wolfgang Kuhnt[a], Rina Zuraida[a, b], Nils Andersen[c] and Pieter M. Grootes[c]

[a] Institute of Geosciences, Christian-Albrechts-University, Ludewig-Meyn-Str. 10-14, D-24118 Kiel, Germany

[b] Leibniz-Institute of Marine Sciences, IFM-GEOMAR, Wischhofstr. 1-3, D-24148 Kiel, Germany

[c] Leibniz-Laboratory for Radiometric Dating and Stable Isotope Research, Christian-Albrechts-University, Max-Eyth-Str. 11 - 13, D-24118 Kiel, Germany

Received 4 June 2007; revised 1 October 2007; accepted 4 October 2007. Available online 18 October 2007.

Abstract

We present a high-resolution (~ 60–110 yr) multi-proxy record spanning Marine Isotope Stage 3 from IMAGES Core MD01-2378 (13°04.95′ S and 121°47.27′ E, 1783 m water depth), located in the Timor Sea, off NW Australia. Today, this area is influenced by the Intertropical Convergence Zone, which

PANGAEA® – Supplementary Data
Paleoclimate investiations on sediment core MD01-2378

Related Articles

- Climate variability and land-ocean interactions in the ...
  Marine Micropaleontology
- Tropical warming in the Timor Sea led deglacial Antarct...
  Earth and Planetary Science Letters
- Direct comparison of mitochondrial markers for the anal...
  Fisheries Research
- Sensitivity of the Australian summer monsoon to tilt an...

„Purchasers of product X were also interested in product Y!"

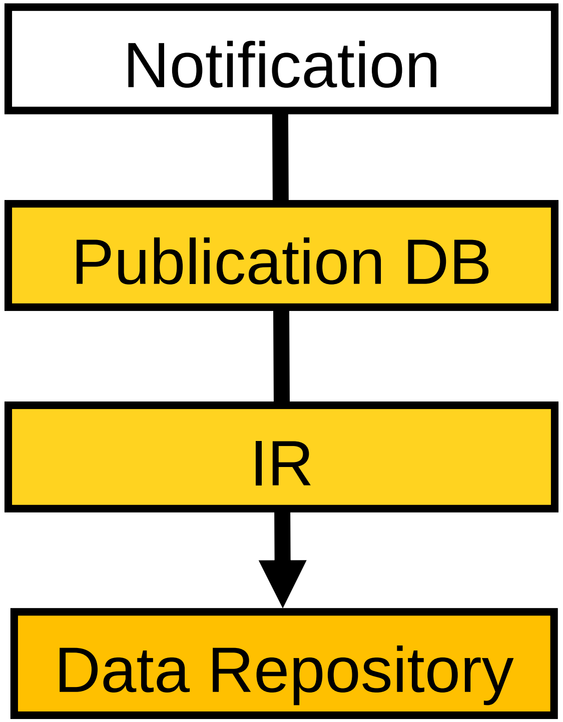
Basic institutional measures

- Implement infrastructure for systematic publication of research data by means of Digital Object Identifiers (DOI).

- Install virtual research environments to integrate data management into scientific work flows.

- Install adequate data repositories (centralized vs. decentralized; internal vs. external).

Once again: Institutional workflows



- In general, required policy components already exist at institutions.
- Bibliometric indicators are crucial factors for scientific reputation !
- Institutional support of workflows (library, administration, data center, ...)
- Platform for data publication (e.g., provided by library)

Vision

danblackonleadership.info

- **Vision** of research organisations:

  „Research data should be openly accessible for the long-term and distributed without restrictions.“

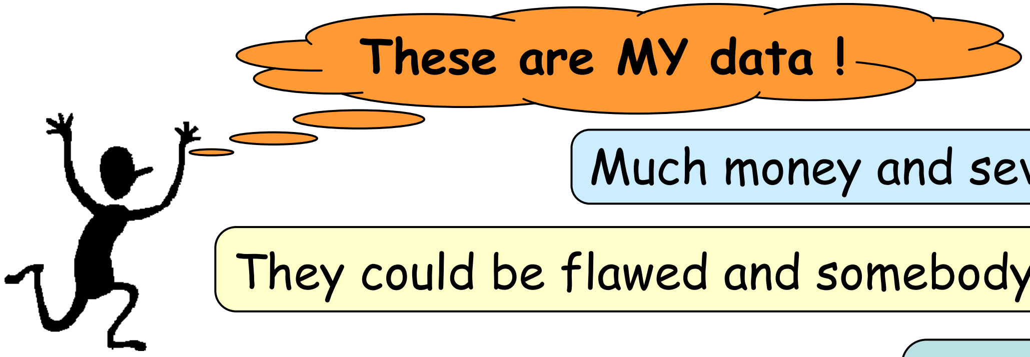
- **Intended purpose:** Open and long-term access ...
  - promotes new findings and research approaches,
  - allows new interpretations,
  - is of strategic importance for the competetiveness of science locations
  - enables syntheses of research results of different disciplines,

  ↳ accelerates the process of making scientific findings !

- **But ...**

… and practice

There are good reasons not to share data, that is „give away“:



Much money and several years of work were put into it.

They could be flawed and somebody else could find it.

They could easily be misinterpreted.

Someone could find something interesting in it.

Somebody could publish them before me.

How and where should I submit my data?

It needs too much time to put the data into a new context and a new format.

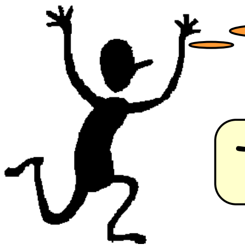
Data management generates costs (at the expense of research itself?).

I won't have any control on my data.

Data preparation for repositories often is too complicated.

… and practice

There are good reasons not to share data, that is „give away“:



These are MY data !

Much money and several years of work were put into it.

They could be flawed and somebody else could find it.

They could easily be misinterpreted.

Someone could find something interesting in it.

Somebody could publish them before me.

How and where should I submit my data?

It needs too much time to put the data into a new context and a new format.

Data management generates costs (at the expense of research itself?).

I won't have any control on my data.

Data preparation for repositories often is too complicated.

⇨ **Researchers need reasons to share !**

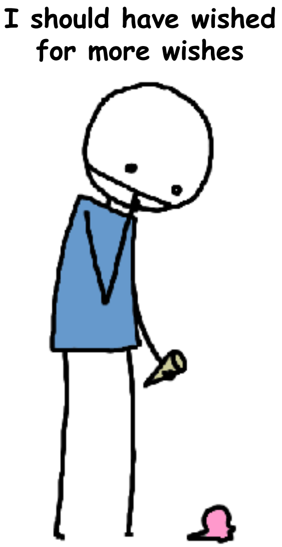
Challenges

123RF Limited

- “Temporary scientists generate the major share of research data.“ (and are gone …)
- Integration of data management into work flows
- Conflicting areas of „Big Data Science“ vs. „Small Data Science“
- Interoperability of discipline specific, international and across discipline structures
- Data management as obligatory component of training ?
- Establishment of incentive schemes
- Cultural change: availability of research data as a matter of course

Desires

- Visualization of mutual *„benefits“*
- Adapting data management, tools and services into typical work flows in disciplines
- Generic tools that are easily transferable (e.g. within temporary projects)
- Flexible interfaces for adaptation to needs and previous knowledge of users
- Transparency: *What happens with my data ?*
- Balance between information science and service



sam brown, explodingdog

Incentives:
On honour, fame and fortune

- **Reputation**:
  - Establishment of a culture of data publication
  - Integration in reputation systems (h-index, ...)
- **New professions**:
  - Bridging the discontinuity between science and information management
- **Provision of science-related tools**:
  - generic (and comprehensible) interfaces, virtual research environments, ...
- **Service-oriented infrastructures**

© SmartSign

© 2014 - Positioning Systems

↳ Recognition of the data provision for further use as a natural component in science

Imagine ...

... that backup and archiving, sharing and reuse of research data would be possible and is reality !

imageshostingfree.com

- ✓ We would have more data per discipline.
- ✓ We could compare many differing data.
- ✓ Doubling of studies could be avoided.
- ✓ Quality control would be guaranteed.
- ✓ Other data sets could support own findings.
- ✓ Own results could be supported by others.
- ✓ Unique, non-reproducible results would be documented.
- ✓ New interpretations would be possible.

© 2014 ViryaNet

↳ Make your data usable, not just accessible !

Thank you for your attention !

GFZ
Helmholtz Centre
POTSDAM

HELMHOLTZ
ASSOCIATION