

Доверяй, но проверяй

Russian Proverb

Used by **R.Reagan** and **W.I.Lenin**

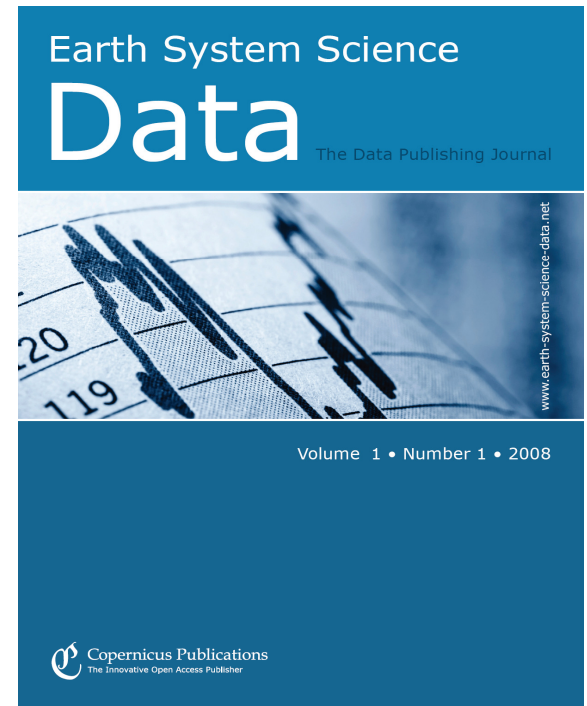
=> Trust, but verify

Data Libraries – A Matter of Trust (in reliability, quality,...)

Hans Pfeiffenberger

Alfred-Wegener-Institute for Polar and Marine Research,
Helmholtz Association - Germany

DL.org Workshop, The British Academy, 2011-02-04, London



Agenda

- **The analog analogue**
 - In what we (need to) trust
 - The **LIBRARY**
 - A trustworthy book
- **The data challenge**
 - A scalability challenge
 - Your government already knows: It is hard!
- **Trust – Data – Digital Libraries**
 - criteria for a trustworthy digital library landscape, serving research data

On The Shoulders of Giants

- Isaac **Newton** famously remarked in a letter to his rival Robert **Hooke** dated February 5, 1676 that:

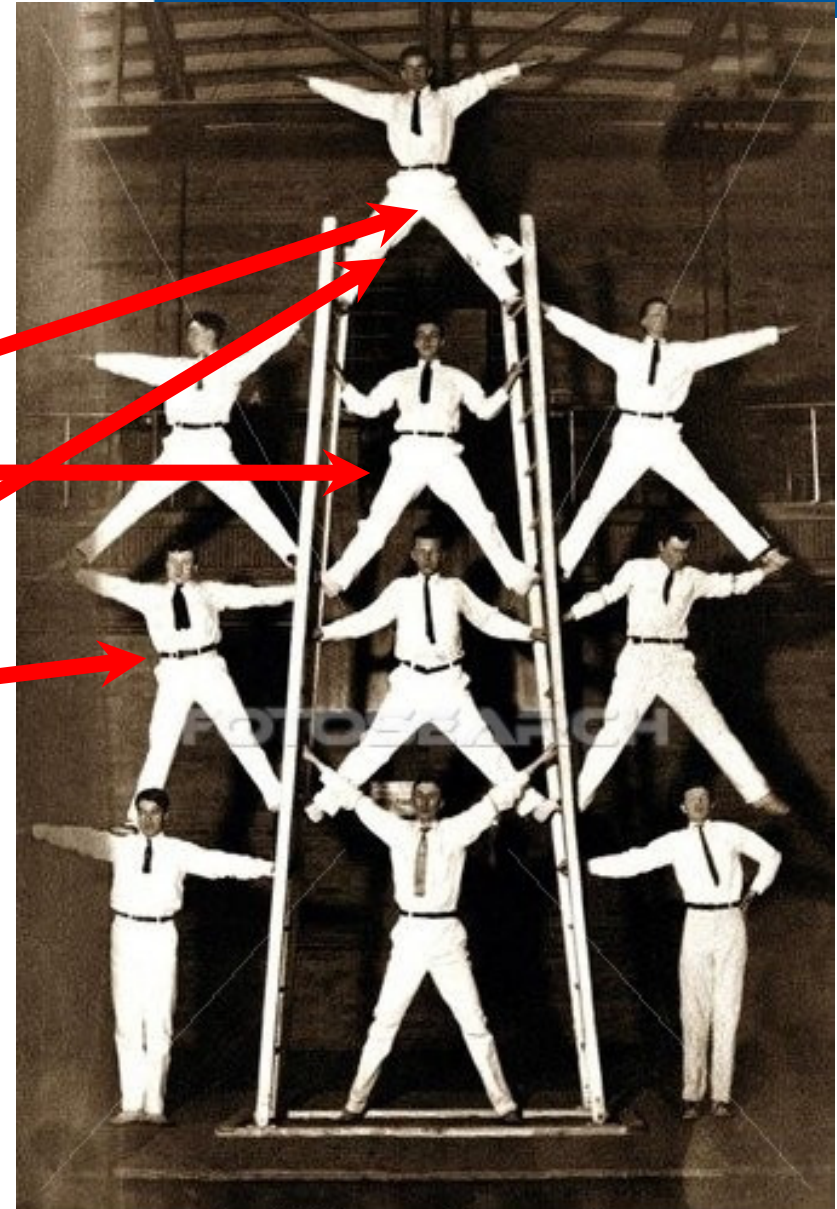
„What **Descartes** did was a good step. You have added much several ways, and especially in taking the colours of thin plates into philosophical consideration.

If I have seen a little further it is by standing on the shoulders of Giants.“

en.wikipedia.org/wiki/Standing_on_the_shoulders_of_giants

Giants today are ...

- probably more like ordinary humans ;-))
- Authors of articles
- Creators of datasets
- **May datasets be less reliable** than articles, books? Why treat them different?



bxp22772 www.fotosearch.de

Schmalkalden, Germany, 2010, sagging of the ground



To pull the ground from under s.o.'s feet (German Proverb)

- The proverb shows:
It is important to us to have solid ground below our feet and a solid foundation for our houses, roads,
- We **do not expect** things to happen as in Schmalkalden
- We **trust in** the **survey** that they would not let us build where the ground is treacherous
- We know: Nobody is perfect, so “stuff happens”
- But the survey **does their job professionally** and **according to rules** ...

Cologne, Germany, 2009; Historical Archive drops into tube construction site



- North-Rhine-Westphalia
- 10 Mio population
- 2 civil engineers for oversight of building
- **This is what must not happen to the building of scientific knowledge!**

DL.Org workshop , policy working group

- What is / constitutes a Digital Library?
- s.o. (McKenzie Smith?): *When you enter a real library, you know it is a **Library***
- This actually means two things:
 - **“The library” is a well managed brand**
 - We **expect** the library will deliver certain (minimum) **quality services:**
 - **Books, journals have a minimum quality**
 - **There are no falsifications**
 - **...**

A trustworthy book

Julio Gonzalo Costantino Thanos
M. Felisa Verdejo Rafael C. Carrasco (Eds.)

Research and Advanced Technology for Digital Libraries

10th European Conference, ECDL 2006
Alicante, Spain, September 17-22, 2006
Proceedings

 Springer

Volume Editors

Julio Gonzalo
M. Felisa Verdejo
Universidad Nacional de Educación a Distancia (UNED)
Departamento de Lenguajes y Sistemas Informáticos
c/Juan del Rosal, 16, 28040 Madrid, Spain
E-mail: {julio.felisa}@lsi.uned.es

Costantino Thanos
Consiglio Nazionale delle Ricerche
Istituto di Scienza e Tecnologie dell'Informazione
Via Moruzzi, 1, 56124, Pisa, Italy
E-mail: Costantino.Thanos@isti.cnr.it

Rafael C. Carrasco
Universidad de Alicante
Departamento de Lenguajes y Sistemas Informáticos
03071 Alicante, Spain
E-mail: carrasco@dlsi.ua.es

Library of Congress Control Number: Applied for

CR Subject Classification (1998): H.3.7, H.2, H.3, H.4.3, H.5, J.7, J.1, I.7

LNCS Sublibrary: SL 3 – Information Systems and Application, incl. Internet/Web and HCI

ISSN	0302-9743
ISBN-10	3-540-44636-2 Springer Berlin Heidelberg New York
ISBN-13	978-3-540-44636-1 Springer Berlin Heidelberg New York

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

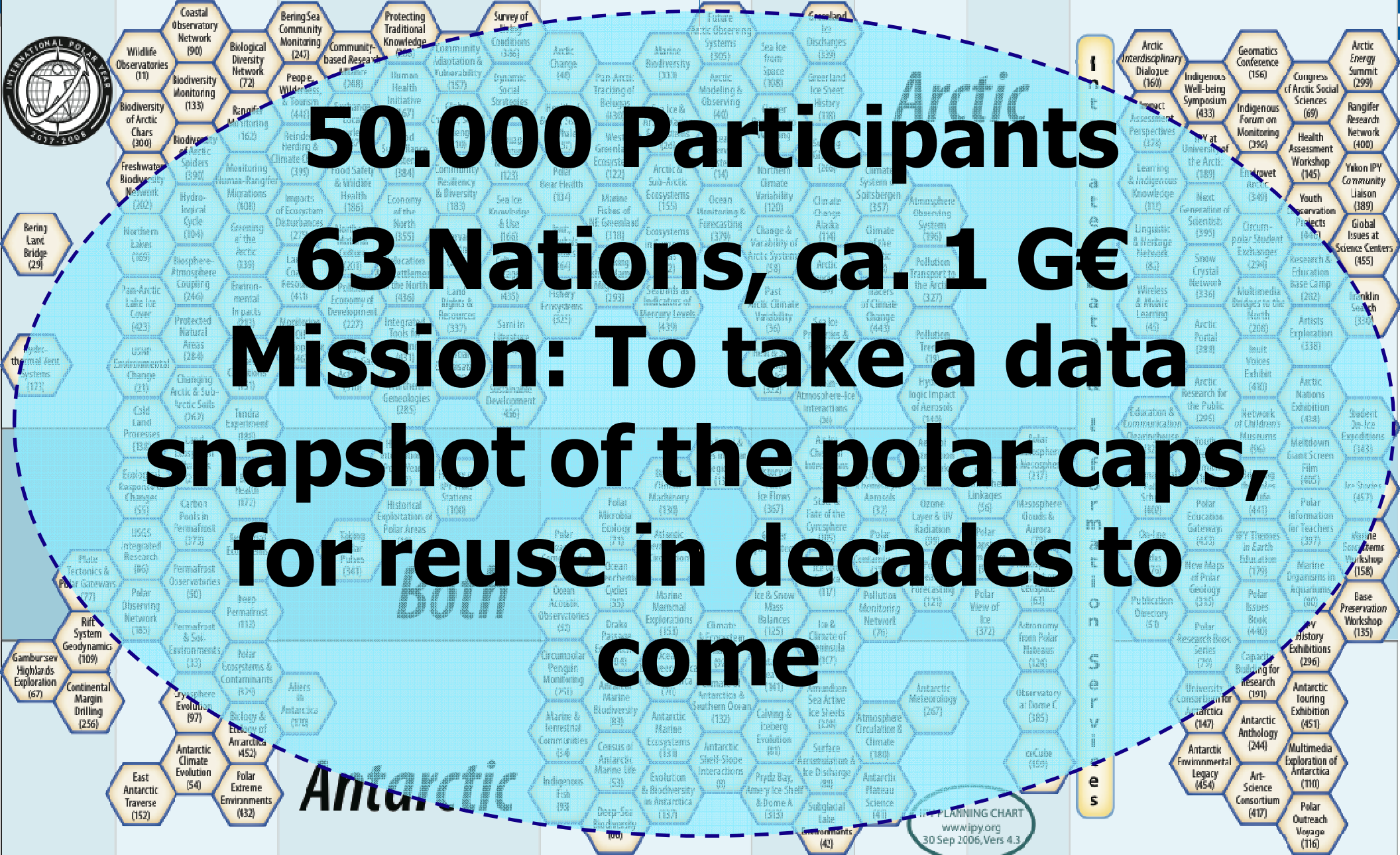
Springer is a part of Springer Science+Business Media

springer.com

© Springer-Verlag Berlin Heidelberg 2006
Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 11863878 063142 543210

Earth Land People Ocean Ice Atmosphere Space Education & Outreach



Earth Land People Ocean Ice Atmosphere Space Education & Outreach

Making this a reality is a more difficult task than it may seem. To collect, curate, preserve and make available ever-increasing amounts of scientific data, new types of infrastructures will be needed. The potential **benefits are enormous but the same is true for the costs.**

Neelie Kroes, VP European Commission

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


PANGAEA - Elsevier

 [Purchase PDF \(743 K\)](#) |  [Export citation](#)

[Abstract](#) | [Article](#) | [Figures/Tables](#) | [References](#)




Marine Micropaleontology
 Volume 66, Issues 3-4, 20 February 2008, Pages 192-207

doi:10.1016/j.marmicro.2007.09.002 | [How to Cite or Link Using DOI](#)
 Copyright © 2007 Elsevier B.V. All rights reserved.

 Cited By in Scopus (2)

 [Permissions & Reprints](#)

Organic matter rain rates, oxygen availability, and vital effects from benthic foraminiferal $\delta^{13}\text{C}$ in the historic Skagerrak, North Sea

Sylvia Brückner  ^a,  and Andreas Mackensen ^a, 

^aAlfred Wegener Institute for Polar and Marine Research,
 Columbusstr., D-27568 Bremerhaven, Germany

Received 27 March 2007; revised 21 September 2007;
 accepted 24 September 2007. Available online 4 October 2007.

Abstract

The sediment cores 225514 and 225510 were recovered from 420 and 285 m water depth, respectively. They were investigated for their benthic foraminiferal $\delta^{13}\text{C}$ during the last 500 years.

Purchase the
 full-text article

- ▶ PDF and HTML
- ▶ All references
- ▶ All images
- ▶ All tables





PANGAEA® – Supplementary Data

Stable carbon isotope composition of benthic foraminifera from sediments of the Skagerr...



Related Articles

-  [The tropical rainbelt and productivity changes off north...
Marine Micropaleontology](#)
-  [Temporal variability in living deep-sea benthic foramin...
Earth-Science Reviews](#)
-  [Early Maastrichtian benthic foraminiferal assemblages f...
Marine Micropaleontology](#)

PANGAEA®

Publishing Network for Geoscientific & Environmental Data

Always quote citation when using data!

Data Description

[Show Map](#) [Google Earth](#)

Citation: **Boström, Kurt (2005):** Water content, carbon and carbonate analysis of sediment core Y80_122SGC. doi:10.1594/PANGAEA.230011

Reference(s): **Boström, Kurt; Thiede, Jörn; Bock, W (1984):** YMER-80, Swedish Arctic Expedition. *Meddelanden från Stockholms Universitets Geologiska Institution, Department of Geology, Stockholm, 260*, 123 pp

Coverage: *Latitude:* 79.450000 * *Longitude:* 1.366667
Date/Time Start: 1980-08-24T00:00:00 * *Date/Time End:* 1980-08-24T00:00:00
Minimum DEPTH, sediment: 0.0 m * *Maximum DEPTH, sediment:* 3.8 m

Event(s): **Y80_122SGC (122SGC)** * *Latitude:* 79.450000 * *Longitude:* 1.366667 * *Date/Time:* 1980-08-24T00:00:00 * *Elevation:* -3175.0 m * *Recovery:* 4.08 m * *Location:* Arctic Ocean * *Campaign:* YMER-80 * *Basis:* Ymer * *Device:* Gravity corer

Parameter(s):

#	Name	Short Name	Unit	Principal Investigator	Method	Comment
1	DEPTH, sediment	Depth	m			Geocode
2	Water content of wet mass	Water wm	%	Boström, Kurt		
3	Carbon, organic, total	TOC	%	Boström, Kurt		
4	Calcium carbonate	CaCO3	%	Boström, Kurt		

License: Creative Commons Attribution 3.0 Unported

Size: 66 data points

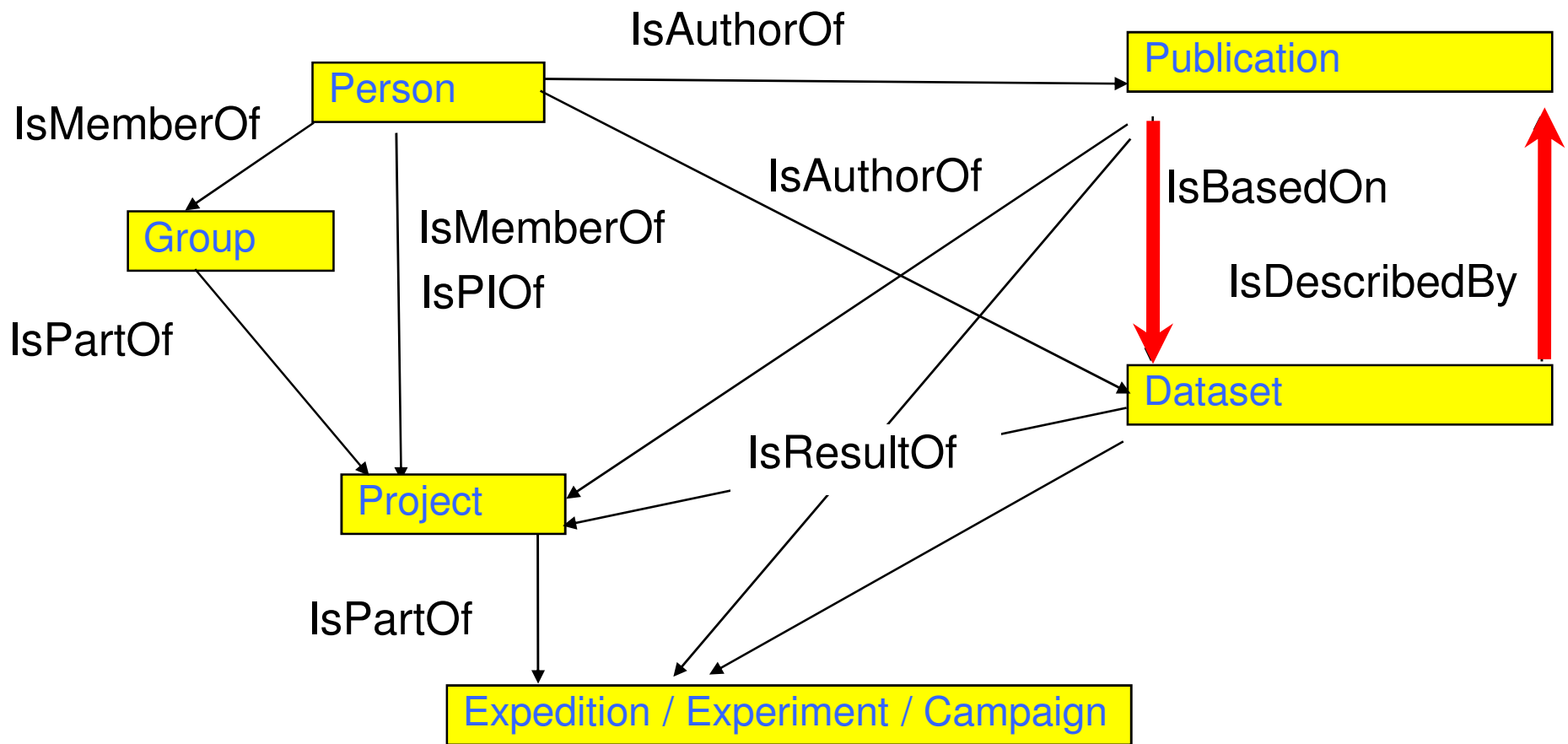


Download Data

Download dataset as tab-delimited text (use the following character encoding:)

View dataset as HTML

Pfeiffenberger, Macario, Text, Data and People, OAI4, CERN 2005



Who is who...

Advisory Board:

Paul J. Crutzen

Sydney Levitus

Alexander Petrovich Lisitzin

Editors in Chief:

David Carlson

Hans Pfeiffenberger

Publishing House

Copernicus Publishers – OA Publisher, EGU

Managing Editor

Suenje Dallmeier-Tiessen



Estimate of Error and Data Provenance

For balloon-borne ozone profile measurements a pump correction has to be applied in order to compensate the decreasing pump efficiency with increasing height and changing air temperature. Both, an inadequate pump correction and an erroneous estimate of residual ozone above the height of balloon burst may contribute to the overall measurement error of the ozone profile. Usually an independent column ozone observation X_D by spectrometer measurement is compared with the integrated column ozone X_S between the ground level and the height of balloon burst plus estimated residual ozone above that level to adjust the recorded profile values. The correction factor is

$$C = X_D / X_S.$$

Systematic differences and random errors of the electrochemical ozone sonde, type OSF, has been estimated by analysing 20 tandem ozone soundings at the Aerological Observatory Lindenberg in 1982 (Feister et al., 1985). Random errors are at their maximum of about 10 to 13% in the troposphere and above 32 km, and reach a minimum of 2 to 5% between 20 and 28 km. The mean random error is 11.5% in the troposphere, 7% in the stratosphere beneath the ozone maximum height (ca. 22 km), and 5.6% above that height.

2 Data Provenance and Structure

The first permanently operated German research base – later named Georg-Forster-Station – was established in 1976 in the Schirmacher Oasis at 70°46' S, 11°41' E. Since then the station was permanently used and operated as an annex to the Russian station Novolazarevskaya until 1987, and then as a German Antarctic station named after

ESSDD

1, 1–13, 2008

**Antarctic
ozonesonde profiles**

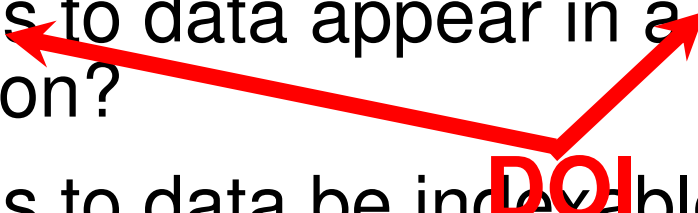
 G. König-Langlo and
H. Gernandt

[Title Page](#)
[Abstract](#)
[Instruments](#)
[Data Provenance & Structure](#)
[Tables](#)
[Figures](#)

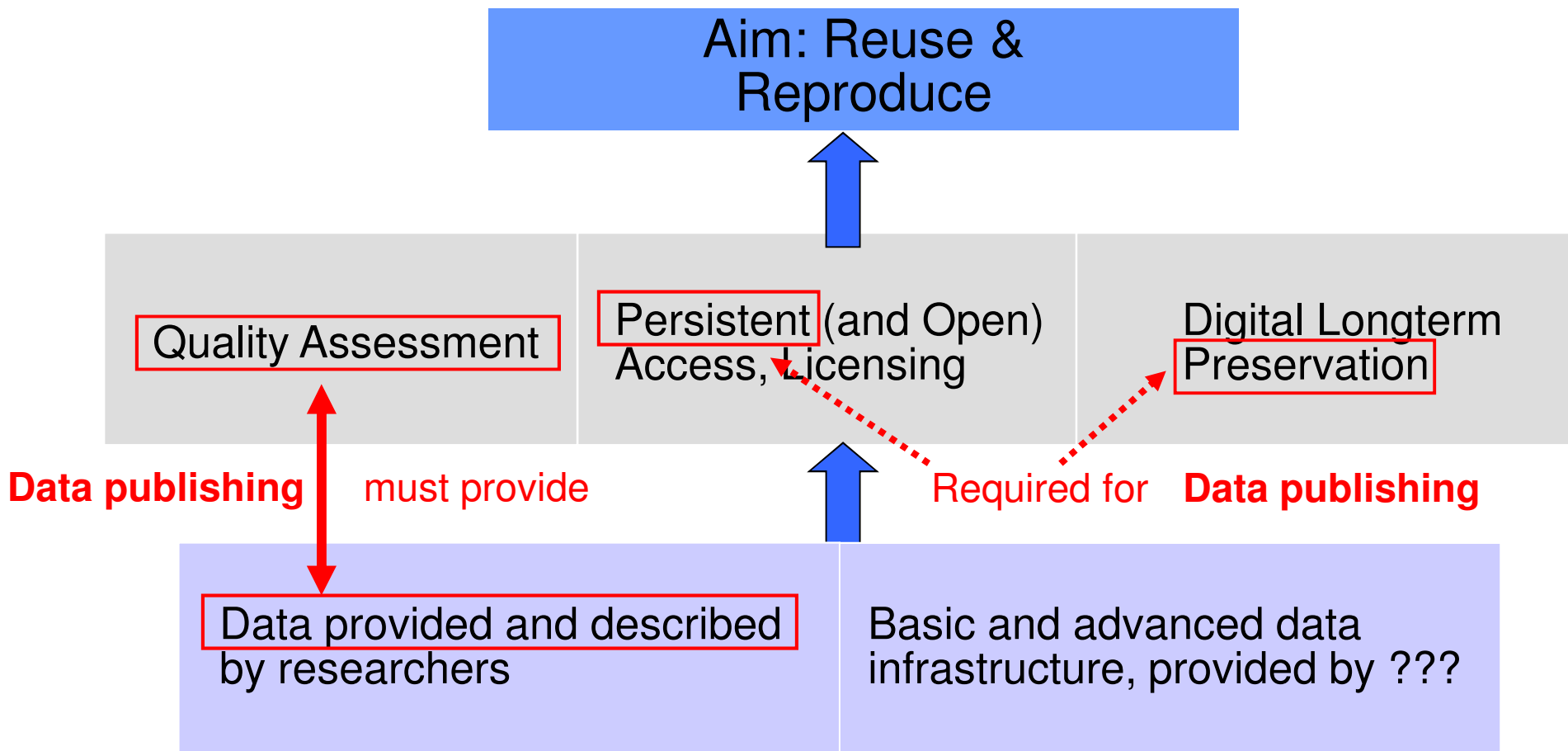



[Back](#)
[Close](#)
[Full Screen / Esc](#)
[Printer-friendly Version](#)
[Interactive Discussion](#)


May 2011, Institute for Quantitative Social Science, Harvard

- The goal of the **Data Citation Principles workshop** is to ... articulate common answers to **core questions** such as:
 - **Should data** that is used to support published research results **be cited** in peer-reviewed articles presenting those results
 - Should each **citation include sufficient information to allow a unique dataset to be identified over time?**
 - Should citations to data appear in a consistent place within publication?
 - Should citations to data be **indexable** so that they can be used for linkage and impact analysis?
- 
- DOI

ESF: “... permanent access to ... quality assured research data”



DOI = a handle (a technology) with a policy

- The CNRI “Handle System” resolves “ID”s to URLs
- The DOI foundation operates “a” handle server, and approves registration agencies
- DataCite is a DOI registration agency for research data
 - DataCite is now considering to ask **data repositories** for some kind of **certification** (or will they **accredit** them?)
- => you need a **professional** organization with some **policy** to permanently deliver on the technologies’ promise
- DataCite will foster **global interoperability** about a specific **policy issue**

data repositories today

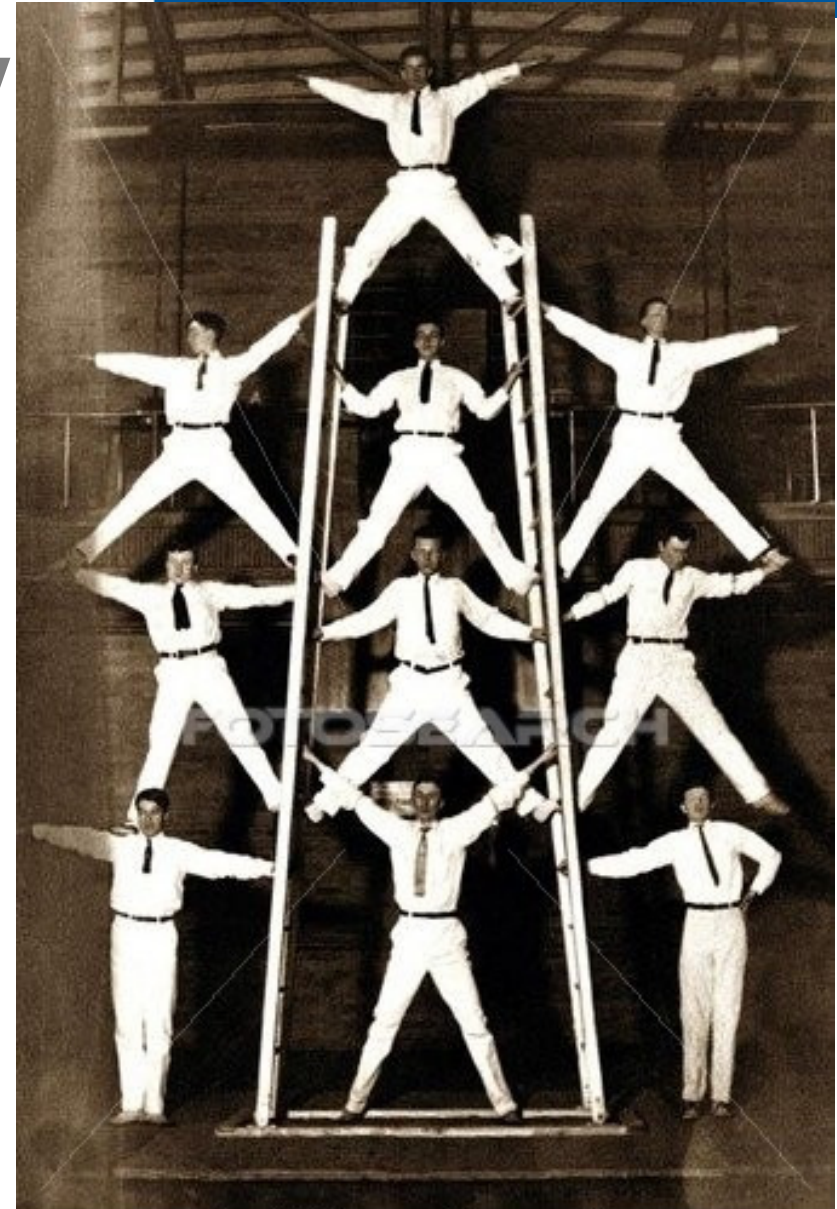
- There are lots of data repositories today...
- Most operate as **projects**, on a **best effort** basis
- They are highly incompatible regarding, e.g.:
 - (access) protocols and formats supported
 - **content qualities** (QA, granularities, ...)
 - rights/licensing
- **interoperability at a global scale is hard/impossible**
 - integration of data (**don't mix high/low quality data**)
 - trust about **long term availability**

Digital Data Library = a data repository with a policy

- **ICSU World Data System** is about **global interoperability** on a number of **policy issues**
 - **long term availability: handover of data in case of default**
 - would be so much more helpful, if DOIs were employed
 - **Open Access**
 - makes things so much easier
 - What about endangered species, social science data?
 - **operates by accreditation, considers certification**
 - Which certification? **ICSU World Data System** is about **global interoperability** on a number of **policy issue**
 - ICSU expects contributions from **developing countries...**

Conclusions: Data - Trust - Quality

- two most important elements **stability of the knowledge architecture** of science :
 - **Quality of each building block**
 - quality assurance,
 - encoding of quality indicators
 - **provenance**
 - **Persistent availability, accessibility of each block**
 - **Handover / Mirroring**
 - Persistent IDs
 - Checksums?



bxp22772 www.fotosearch.de

Thank you!

Criteria for service adoption by communities

1 year horizon (Individual, piece of software)

- Usefulness (cost of learning/adoption vs. short term success)
- Availability (Access)
- Usability (No certificates! Classical GIS vs. Google Maps!)

10+ years horizon (Community, Infrastructure)

- **Trust** (Proven **scientific validity** / quality, conferred by branding)
- **Reliability** (Persistence, to become part of **scientific practise**)

Catch 22 (for service providers, visionaries)

- **Real big improvements for complex problems** (Earth System)
- **Trust builds slowly, bottom up** (per narrow discipline)