Data are First Class Products of Science – Data Journals are one Way to Establish this

Hans Pfeiffenberger
Alfred-Wegener-Institute for Polar and Marine Research, Helmholtz Association - Germany
Helmholtz OA Workshop DESY, 2013-06-10, Hamburg

First, a Question Asked by Funding Agencies: Who is Afraid of Open Access to Data?

- Those who
  - **Invented** their data (Stapel),
  - **Selected data with a bias** (notorious: Clinical trials)
  - **Read wrong or to much** from their data (Reinhart/Rogoff)

- Those who build **business-models** on a *monopoly on knowledge or facts*
  - Publishers (“article of the future”)
  - Thomson Reuters (“Web of Knowledge”)
Royal Society: Science as an Open Enterprise (2012)

- Geoffrey Boulton, *Open your minds and share your results* Nature (486) 441, doi:10.1038/486441a

- **Open enquiry has been at the heart of science** since the first scientific journals were printed in the seventeenth century. …

- **Science's capacity for self-correction** comes from this openness to scrutiny and challenge.

- « **Intelligent Openness** »

  Access to data

---

Modern Science is based on data – since Renaissance!

- **1546 - 1601**: Tycho Brahe, *quality* data
- **1606 - 1618**: Kepler’s Laws
- **1684 - 1687**: Newton *De Motu* – Principia
  - explained (!) Kepler’s laws

\[ F = ma \]

\[ F \sim \frac{mM}{r^2} \]

<table>
<thead>
<tr>
<th>Planet</th>
<th>(T)</th>
<th>(a)</th>
<th>(T^2)</th>
<th>(a^3)</th>
<th>(T^2 / a^3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Merkur</td>
<td>0,241</td>
<td>0,387</td>
<td>0,058081</td>
<td>0,057960603</td>
<td>1,002077221</td>
</tr>
<tr>
<td>Venus</td>
<td>0,615</td>
<td>0,723</td>
<td>0,378225</td>
<td>0,377933067</td>
<td>1,000772446</td>
</tr>
<tr>
<td>Erde</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Mars</td>
<td>1,881</td>
<td>1,524</td>
<td>3,538161</td>
<td>3,539605824</td>
<td>0,999591812</td>
</tr>
<tr>
<td>Jupiter</td>
<td>11,863</td>
<td>5,203</td>
<td>140,730769</td>
<td>140,6515004</td>
<td>0,999142846</td>
</tr>
<tr>
<td>Saturn</td>
<td>29,458</td>
<td>9,555</td>
<td>867,773764</td>
<td>872,3526289</td>
<td>0,994751131</td>
</tr>
</tbody>
</table>

\(T\) = siderische Umlaufzeit in trop. Jahren \(a\) = große Halbachse in astronomischen Einheiten (Abstand Erde–Sonne)
Paradigms of modern (17th century) Science

- Experiment / Observation
  - Repeatable, (documented), ...
  - quantified
- Theory
  - mathematically expressed
  - checked against facts
- Published
- Governed by “Rules of good scientific conduct”
Meitner-Hahn-Strassmann Uran-Experiment, Berlin-Dahlem, 1938

The last big discovery by a „single“ person with a lab notebook?

The biggest experiment in the world (not at CERN!)
The Fourth Paradigm (2009)  
Jim Gray (Microsoft !!)

- **Research using big data holdings** as a new, distinct way to do research
- **Third P. was numeric modelling / supercomputing**
- **How is it done (properly)?**  
  What are the non-technical challenges?
- **How to communicate, how to preserve its results?**

Does computation threaten the scientific method? Reproducibility?

The scientific method has been the most successful contributor to systematic progress in the history of human endeavour. One of the key elements of the method is that if the result cannot be reproduced, it is discarded. Models are then

- using the same processed data from eight other companies, the same algorithms in the same programming language, same input data, but coded independently

- L. Hatton, A. Giordani
Intelligent, qualified Openness (™ Royal Soc. 2012)

- **Accessible** (such that it can readily be found)
- **Intelligible** (to those wishing to understand or scrutinise them)
- **Assessable** (to be able to make some judgment)
- **Usable** (data should be able to be reused)

- There are legitimate boundaries of openness which must be maintained in order to protect commercial value, privacy, safety and security.

(siehe Science as an Open Enterprise)

Socio-Cultural Change (2)

- **NSF Post Award Requirements**
  - Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, ...
  - in a form that protects the privacy of individuals and subjects involved. General adjustments and, where essential, exceptions to this sharing expectation may be specified by the funding NSF Program or Division/Office for a particular field or discipline …

(http://www.nsf.gov/bfa/dias/policy/dmp.jsp)
Socio-Cultural Change (1)

- NSF Proposal Preparation Instructions (Jan 2013)
  Proposals / PIs’ CVs must contain:
  - “A list of: (i) up to five products most closely related to the proposed project; … Acceptable products must be citable and accessible including but not limited to publications, data sets, software, patents, and copyrights.”
  - “Plans for data management and sharing of the products of research. … no more than two pages”.

- see San Francisco declaration … DFG Quality not Quantity

(www.nsf.gov/pubs/policydocs/pappguide/nsf13001/gpg_2.jsp#lIC2fic)

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
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 \( X_0 \) by spectrometer measurement is compared with the integrated column ozone \( X_2 \) 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

\[
C = \frac{X_0}{X_2}.
\]

Systematic differences and random errors of the electrochemical ozone sonde, type O3R, has been estimated by analysing 20 tandem ozone soundings at the Aerological Observatory Lindenberg in 1962 (Feisler et al., 1965). Random errors are at their maximum of about 10 to 15% 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

2013: CO above Troll Station, Original Data

BAS microwave radiometer CO profiles acquired at Troll station, Antarctica between Feb 2008 and Jan 2010

Contact: Patrick Pfeiffer, Tel: +49 73 55 10 55, Email: patrick.pfeiffer@tms.de

The 2-sigma systematic error bars provided have been determined using perturbation calculations:

- temperature error: error induced by the temperature profile (estimated error = 5K) needed as additional information for the retrieval, mainly random
- calibration error: error induced by the calibration of the measured spectrum (estimated error = 10 percent), can be small
- spectrometer error: usually lineintensity from EXTRAN 2004 with an estimated error of 2 percent, systematic
- channel shape error: uncertainty due to the use of a modified channel response function in the retrieval in order to improve the accuracy of the calibrations for the radiometer's local oscillators after 2008-08-09.

Sum of errors: The error bars give the sum of certain errors that are added up as follows sqrt(error1^2 + error2^2)

- pressure [hPa]
- altitude [km]
- vmr [ppmv]
- a priori contribution [ppmv]
- temperature error [percent]
- calibration error [ppmv]
- spectroscopy error [ppmv]

The data set covers the period from February 2008 to January 2010, however, due to very low CO concentrations.
2012: Nature CC & ESSD; Carbon data aggregation at global scale

Terrestrial CO\textsubscript{2} sink (positive values represent a flux from the atmosphere to the land)

All values in petagrams of carbon per year (PgC/yr), for the globe. For values in carbon dioxide (CO\textsubscript{2}), multi

1PgC = 1 petagram of carbon = 1 billion tonnes C = 1 gigatonne C = 3.67 billion tonnes of CO\textsubscript{2}

Cite as:

<table>
<thead>
<tr>
<th>CLM4CN</th>
<th>HYLAND</th>
<th>LPJ-GUESS</th>
<th>LPJ</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.79</td>
<td>2.02</td>
<td>0.42</td>
<td>-0.83</td>
</tr>
<tr>
<td>0.75</td>
<td>1.53</td>
<td>1.16</td>
<td>0.81</td>
</tr>
<tr>
<td>0.30</td>
<td>1.71</td>
<td>-0.07</td>
<td>-0.55</td>
</tr>
<tr>
<td>0.79</td>
<td>2.37</td>
<td>1.25</td>
<td>0.57</td>
</tr>
<tr>
<td>-1.20</td>
<td>1.81</td>
<td>0.26</td>
<td>-0.37</td>
</tr>
</tbody>
</table>

Reviewer: “no effort appears to have been made to engage the specialist scientists who have spent months or years at sea collecting such data. “ - not knowing that:

Authors asked 164 potential contributors – got answer from 13!
Conclusions

- Intelligent Openness will come

- It is really **scientific work to prepare data** (and software) accordingly

- **Peer reviewed Data Publication** is *one way to demonstrate this* (and earn reputation on it)