Who is afraid of Data Publishing – The ESSD Experience

Hans Pfeiffenberger, Dave Carlson

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

APE2014, 2014-01-29, Berlin
Who should be Scared of Data Publishing?

Respectful (Jan Brase)

Everybody!!

- At least those who don’t like a tough challenge recognize
Who should really be Afraid of Data Publishing?

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

- Those who build business-models on a monopoly on knowledge or facts, e.g.
  - Non-OA Publishers
  - Institutes which consider data collections as “their” capital
Royal Society: Science as an Open Enterprise (2012)

- **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.

- RS applied this to data: **Intelligent Openness**
Scared in the 17th Century

Hooke, published 1676 by anagram „ceiiinossstttuv“

1678 in booklet
Meitner-Hahn-Strassmann Uran-Experiment, Berlin-Dahlem, 1938

The last big discovery by a small group with a lab notebook?
What we do today: ARGO, the biggest experiment in the world
ARGO is not Scared of „Data Publishing“!

What is really fascinating: There are

- More than 3,000 buoys
- from more than 30 countries, lots of companies

and yet there is:

- Co-ordinated (quality) data management
  - One („published“) standard for instruments
  - One („published“) standard for formats
  - One („published“?) standard for processing
  - Open access to data - (almost) no delay
The Dangers of Working in Closed Silos – „Does computation threaten the scientific method?“

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

- L. Hatton, A. Giordani

ISGTW

H. Pfeiffenberger GeoSim Seminar, 2013-02-15, Potsdam
Data Publishing Challenge #1

- **Quality of Data**
  - Royal Soc. “intelligent Openness” (2012):
    Data need to be “… assessable. Recipients need to be able to make some judgment or assessment of what is communicated.

  - “Guidelines on Data Management in Horizon 2020” (2013):
    “… are data provided in a way that judgments can be made about their reliability and the competence of those who created them”
Earth System Science Data (ESSD)

established 2008

Advisory Board:
Paul J. Crutzen
Sydney Levitus
Alexander Petrovich Lisitzin

Editors in Chief:
David Carlson
Hans Pfeiffenberger

Publishing House
Copernicus Publications
– OA Publisher, EGU
Estimate of Error and Data Provenance
- No fancy interpretations!!

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

$$C = \frac{X_D}{X_S}.$$  

Systematic differences and random errors of the electrochemical ozone sonde, type SC2P, has been estimated by analysing 28 random ozone soundings at the Aeronautical 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
2013: CO above Troll Station, Original Data

Mesospheric CO above Troll station, Antarctica observed by a ground based microwave radiometer

C. Straub\textsuperscript{1}, P. J. Espy\textsuperscript{1}, R. E. Hibbins\textsuperscript{1}, and D. A. Newnham\textsuperscript{2}

\textsuperscript{1}Norwegian University of Science and Technology (NTNU), Trondheim, Norway
\textsuperscript{2}British Antarctic Survey, Cambridge, UK

Abstract. This paper presents mesospheric carbon monoxide (CO) data acquired by the ground-based microwave radiometer of the British Antarctic Survey (BAS radiometer) stationed at Troll station in Antarctica (72° S, 2.5° E, 1270 a.m.s.l.). The data set covers the period from February 2008 to January 2010, however, due to very low CO
2013: CO above Troll Station, Original Data

Middle atmospheric carbon monoxide above Troll station, Antarctica from February 2008 - January 2010

GB/NERC/BAS/PDC/00789

Summary

Abstract:
This data set contains mesospheric carbon monoxide (CO) data acquired by the ground-based microwave radiometer of the British Antarctic Survey (BAS radiometer) stationed at Troll station in Antarctica (72 deg S, 2.5 deg E, 1270 amsl). The BAS radiometer has been designed in order to study the effects of energetic particle precipitation on the middle and upper atmosphere, using nitric oxide and ozone measurements. This data set contains the CO measurements carried out in order to study the dynamical context.

The data set covers the period from February 2008 to January 2010, however, due to very low CO concentrations...
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 Espy, tel: +47 73 55 10 95, email: patrick.espy@ntnu.no

date [UT]: 2009-10-19 10:44:06
apriori contribution: The profile is most reliable where the contribution from the a priori profile is less than approx. Negative values are a scaling artifact and should be regarded as close to 0.

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

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0.749894</td>
<td>50.679</td>
<td>0.060</td>
<td>-5.939</td>
<td>0.003</td>
<td>0.048</td>
<td>0.010</td>
</tr>
<tr>
<td>0.562341</td>
<td>53.021</td>
<td>0.065</td>
<td>-20.151</td>
<td>0.002</td>
<td>0.056</td>
<td>0.011</td>
</tr>
<tr>
<td>0.421697</td>
<td>55.337</td>
<td>0.072</td>
<td>-27.600</td>
<td>0.002</td>
<td>0.061</td>
<td>0.012</td>
</tr>
<tr>
<td>0.316228</td>
<td>57.609</td>
<td>0.080</td>
<td>-29.442</td>
<td>0.004</td>
<td>0.067</td>
<td>0.013</td>
</tr>
</tbody>
</table>

Smoothing error: This error only needs to be considered if the profiles of the BAS radiometer are compared to profiles with a significantly larger vertical resolution. For such a comparison the better way would be to convolve the high-resolution profile with the AVK of the retrievals.

Sum of errors: To build the sum of certain errors they are added up as follows sqrt(error1^2 + error2^2)
Data Publishing Challenge #2

- **Citability / Cite-worthy-ness / Reputation**
  - **NSF Proposal Preparation Instructions (2013)**
    Proposals / PIs’ CVs must contain:
    “A list of: (i) up to **five products** … Acceptable products must be **citable and accessible** including but not limited to **publications, data sets, software,** …”
  - **DFG “Rules of Good Scientific Practice” (2013):**
    Recommendation 12 on authorship: contribution may be “**preparation ... of data**”
Fluxes of sedimenting material from sediment traps in the Atlantic Ocean

S. Torres-Valdés¹, S. C. Painter¹, A. P. Martin¹, R. Sanders¹, and J. Felden²
¹Ocean Biogeochemistry and Ecosystems Research Group, University of Southampton, SO14 3ZH, UK
²Center for Marine Environmental Sciences, Universität Bremen, Germany

Abstract. We provide a data set assemblage of directly observed and derived fluxes of sedimenting material (total mass, POC, PON, BSiO₂, CaCO₃, PIC and lithogenic/terrigenous fluxes) obtained using sediment traps. This data assemblage contains over 5900 data points distributed across the Atlantic, from the Arctic Ocean to the Southern Ocean. Data from the Mediterranean Sea are also included. Data were compiled from a variety of sources: data repositories (e.g., BCO-DMO, PANGAEA), time series sites (e.g., BATS, CARIACO), published scientific papers and data provided by originating PI’s. All sources are specified within the combined data set. Data from the World Ocean Atlas 2009 were extracted to coincide with flux
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!
Data Publishing Challenge #3

- **Linking text and data ... and much more**
  - The Lancet “Reducing waste from incomplete or unusable reports of biomedical research” (2014)
  - “… studies of published trial reports showed that … 40–89% were non-replicable”
  - Offered a long laundry list of “Components of study documentation” to be published
Now this laundry list is really scary!

1 The protocol and related documents, such as details submitted for study registration

3 Supplementary materials, such as education materials for patients, clinician training resources, and videos

7 The primary data, data manuals, and statistical code for analyses

9 Reliable and stable bidirectional linkages between all these elements
2012: Nature CC & ESSD; Carbon data aggregation at global scale

The challenge to keep global warming below 2 °C

Glen P. Peters, Robbie M. Andrew, Tom Boden, Josep G. Canadell, Philippe Ciais, Corinne Le Quéré, Gregg Marland, Michael R. Raupach & Charlie Wilson

Affiliations | Contributions | Corresponding author

Published online 02 December 2012
The global carbon budget 1959–2011


1Tyndall Centre for Climate Change Research, University of East Anglia, Norwich Research Park, Norwich, NR4 7TJ, UK
2Carbon Dioxide Information Analysis Center (CDIAC), Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA
3National Oceanic & Atmosphere Administration, Earth System Research Laboratory (NOAA/ESRL), Boulder, Colorado 80305, USA
4Woods Hole Research Centre (WHRC), Falmouth, Massachusetts 02540, USA
5Cabot Institute, Dept of Geography, University of Bristol, UK
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25

This discussion paper is/has been under review for the journal Earth System Science Data (ESSD). Please refer to the corresponding final paper in ESSD if available.
### 2012: Nature CC & ESSD; Carbon data aggregation at global scale

- **Terrestrial CO\(_2\) 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\(_2\)), multiply by 10.
- 1PgC = 1 petagram of carbon = 1 billion tonnes C = 1 gigatonne C = 3.67 billion tonnes of CO\(_2\)

<table>
<thead>
<tr>
<th>Year</th>
<th>Terrestrial CO(_2) sink as a residual of the global carbon budget</th>
<th>Models</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>CLM4CN</td>
</tr>
<tr>
<td>17</td>
<td>0.42</td>
<td>0.79</td>
</tr>
<tr>
<td>18</td>
<td>1.14</td>
<td>0.75</td>
</tr>
<tr>
<td>19</td>
<td>1.20</td>
<td>0.30</td>
</tr>
<tr>
<td>20</td>
<td>1.76</td>
<td>0.79</td>
</tr>
<tr>
<td>21</td>
<td>1.72</td>
<td>-1.20</td>
</tr>
</tbody>
</table>

Cite as:
Linking Text and Data

Data (in repository)  Article in data journal  Article in „classical“ journal
But we do not despair!
Conclusions

- Socio-cultural change is on the way (may need just a few more decades)
  - Need for change/quality is recognized (Lancet)
  - NSF “5 products” rule offers the way out of the metrics dungeon
- “Technical” challenges remain, e.g.
  - Repositories for computer code etc.
  - Quality assessment for “protocols” etc.
  - bidirectional Linking of Everything Open (b-LEO)
  - And, I did not even mention versioning …