Novel approaches in open data access and science -

Publisher’s viewpoint

Hans Pfeiffenberger

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

3rd Nordic ENVRI workshop on data management and interoperability, Lund, 2015-10-28

Publishing - with a capital P

- A Marketing Meme – to emphasize a distinction
  - to publish: make publicly available, e.g. upload to website
  - to Publish - as in Scientific Publishing:
    - is a quite formal, “ritualised” process requires systematic 3rd party scrutiny
    - => QA supposedly yields higher quality
    - establishes priority
  - quality + priority => reputation => willingness to share
  - published items become part of “The Records of Science”
(Data) Publishing - with a capital P

- Standardized, well known process
  => TRUST => we can build on the work of others !!

- Apply these mechanisms to data
  - get all the benefits !! ??
  - thus: “Data Publishing with a capital P”

- In a broader sense, this is about:
  To make data a legitimate part of research culture

Agenda

- A little bit of “historical perspective”

- Impression from recent publishers’ workshop
  (COPDESS)

- ESSD, a Data Publishing journal, est. 2008
  with a note on presentation of data

- Linked Scientific Information
  (all Capital Letters) – but what are the objects?
An impression from history of science

Modern Science is based on data – since Renaissance!

- **1606 - 1618**: Kepler‘s Laws
  - reduced Tycho Brahe‘s quality data
- **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>( d )</th>
<th>( T^2 )</th>
<th>( d^3 )</th>
<th>( T^2/d^3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Merkur</td>
<td>0,241</td>
<td>0,387</td>
<td>0,056081</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>140,8515004</td>
<td>0,999142846</td>
</tr>
<tr>
<td>Saturn</td>
<td>29,458</td>
<td>9.555</td>
<td>867,773764</td>
<td>872,3626289</td>
<td>0,994751131</td>
</tr>
</tbody>
</table>

\( T \) = siderische Umlaufzeit in trop. Jahren, \( d \) = große Halbachse in astronomischen Einheiten (Abstand Erde–Sonne)
The biggest experiment on this planet - ARGO

An important, “typical” Experiment

- **EISENEX / EIFEX**: Two expeditions of “**Polarstern**”: With a few tons of iron fertilizer, south of Capetown ….
- **EIFEX (2004)**:
  - 54 scientists and students from
  - 14 institutes and 3 companies from
  - 7 EU countries and South Africa
  - Oceanographers
  - Biologists
  - Chemists…..
- “Biogeochemistry”
- + Satellite observations!
Scared in the 17th Century

Hooke, published his law 1676 by anagram „cellinossssttu“ 1678 in booklet

Impressions from COPDESS

Coalition on Publishing Data in the Earth and Space Sciences -
**Journals’ Transparency Criteria**

- TOP: modular, agnostic to disciplines, low barrier to entry; categories are:
  - a) **data citation**
  - b) design transparency (standard operating procedures, protocols)
  - c) materials transparency
  - d) **data transparency**
  - e) analytical methods (code) transparency
  - f) preregistration of studies
  - g) preregistration of analysis plans
  - h) replication
- 3 levels
  - level 1: article states whether data are available and where to access
  - level 2: data has to be in trusted repository
  - level 3: data in TR, but reported analyses will be reproduced independently
Journals' Transparency Criteria

- > 526 journals signed + 50 organizations, commitment to review guidelines

Directory of trustworthy repositories (for ESS):

- **Discipline-specific, because only those**
  - can catch the needed metadata
  - can be regarded sustainable
- **In order of specificity:**
  - re3data.org
  - „something“ developed by AGU/NSF/COS
  - journals’ lists (ESSD, Scientific Data)
- **Certification**
  - WDS (actually an accreditation process)
  - Data Seal of Approval
Identifiers

- For data: persistent identifiers for data,
  ⇨ preferably DOIs, implying fixity, integrity!
- For authors, contributors: ORCID
  ⇨ all publishers to use ORCID, even the ORCID AuthN service
- For samples: IGSN

- For grants: FundRef (http://www.crossref.org/fundref/)
  ⇨ http://dx.doi.org/10.13039/501100000780,
  "European Commission",
  - "narrower": [{"resource": "http://dx.doi.org/10.13039/501100000889"},

Transparency = Openness?

- The “How” of Open
  - Is registration acceptable ??
  - Which kind of license (if any)

- The “When” of Open
  - Some time after end of data acquisition?
  - Some time after end of project / funding?
  - At time of publication?!
    - After acceptance?
      - If there is an on-going embargo, how can reviewers have access?
ESSD – Data Publishing in practise

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

<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>-5.399</td>
<td>0.003</td>
<td>0.048</td>
<td>0.010</td>
</tr>
<tr>
<td>0.562241</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>

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 concentrations.
ESSD Principles and Criteria

- ESSD expects data to be at a repository and be
  - Open Access, static, with a DOI
- ESSD expects authors to describe in the article
  - provenance, methods, limitations, estimates of error
- ESSD expects reviewers to
  - actually look at the data
  - assess consistency of article and dataset

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, 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

A huge work to find, assess, collate (quality) data;
24 out of 43 text pages are source data references!
Global Carbon Atlas - Visualization

- Note the **two panels** in www.globalcarbonatlas.org
  - data journalism ("emissions")
  - scientific visualization ("research")

- At the project website www.globalcarbonproject.org
  “real life” and linkage between publications and data
Conclusions

- „Earth Science“ is a „Big Data“ problem („Variety“ in 3 Vs)
  - finding and exploiting patterns in metadata and data

- Still needs Publishing processes
  - For quality assurance and recognition
  - Published text is the best „metadata“ one can have
  - Articles are still the linking hubs in the digital assets ecosystem

- Needs trustworthy infrastructures for data, software, …
  - most of all skilled people doing all the curatorial stuff etc.

- Clever systems exploiting all this for discovery, aggregation, analysis, … real time alerting, disaster mitigation, …
Thank you!

copdess.org

earth-syst-sci-data.net

expedition.awi.de

oa.helmholtz.de

Bonus Material
- (almost) not shown in session

There is more than separate publications and data!

Let’s link it!
Pfeiffenberger, Macario, Text, Data and People, OAI4, CERN 2005

H. Pfeiffenberger, 3rd Nordic ENVRI workshop on data management and interoperability, Lund, 2015-10-28

PANGAEA – Elsevier

Abstract

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

Related Articles

- The tropical rainfall and productivity changes off nort...
- Marine Micropaleontology
- Temporal variability in living deep-sea benthic foramin...
- Earth-Science Reviews
- Early Miocene benthic foraminiferal assemblages f...
- Marine Micropaleontology
eXpedition (in production since 2005)

Related Information:
- Reports on Polar and Marine Research (1982 to date)
- Primary data (all polarstern datasets in PANGAEA)
- Handbook and scientific device documentation (in deutsch)
- DShip (Polarstern Data Acquisition System)
- VirtualPS: Virtual Polarstern Tour

<table>
<thead>
<tr>
<th>Expedition</th>
<th>Date</th>
<th>Region</th>
<th>Research</th>
<th>Publications &amp; Primary Data</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANT-XXI/3</td>
<td>21.01.2004 - 25.03.2004</td>
<td>Atlantic/Indian Ocean, Polar frontal zone</td>
<td>Biology</td>
<td>ePIC: Publications</td>
<td>EIFEX</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ePIC: Reports on Polar and Marine Research</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ePIC: Weekly reports</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>PANGAEA: Stations</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(Note: Publications and datasets for recent cruises may not yet be available)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Meteorology</td>
<td></td>
</tr>
</tbody>
</table>

H. Pfeiffenberger, 3rd Nordic ENVRI workshop on data management and interoperability, Lund, 2015-10-28

---

Alfred Wegener Institut

Home | Latest expeditions | Search | About

- Current expedition
- Selected expeditions

Platform: Polarstern
- Expedition ANT-XXI/3
- Start: 2004-01-21
- End: 2004-02-25
- Location: Capetown

- Near real time data
- Latest data and publications

Archived data
- Continuous liquid water path (LWP) and integrated water vapour (IWV) measurements during POLARSTERN cruise PS83 (ANT-XXI/10) on 2014-04-11 (2014)
- Continuous liquid water path (LWP) and integrated water vapour (IWV) measurements during POLARSTERN cruise PS83 (ANT-XXI/10) on 2014-04-10 (2014)
- Continuous liquid water path (LWP) and integrated water vapour (IWV) measurements during POLARSTERN cruise PS83 (ANT-XXI/10) on 2014-04-09 (2014)

Publications
- New insights into the past glaciation of the northeast Greenland continental shelf (2014)
- Taxon-specific epibenthic foraminiferal 518O in the Arctic Ocean: relationship to water masses, deep circulation, and brine release (2014)

Reports
- Expedition Programme PS88 (2014)
- Expedition Programme Polarsz. Nr. 93 PS84, PS85, PS86, PS87 (APK-XXVIII/1-4) (2014)
- Effects of cold glacier ice crystal anisotropy on seismic data (2014)

H. Pfeiffenberger, 3rd Nordic ENVRI workshop on data management and interoperability, Lund, 2015-10-28
eXpedition – Publications and Data network

Late Quaternary climatic cycles as recorded in sediments from the Antarctic continental margin.

Pfeiffenberger, Carlson, TR32 DM WS Cologne 2014

- ORCID
- ePIC / Publishers
- (STD-)DOI - PANGAEA
- DataCite
- OpenAire
- fundRef
- Sample, Specimen
- IGSN
- ZENODO