Every bit counts

Data management and data publication in the earth sciences

Jens Klump et al.

International Data Exchange Workshop
Kiel, 10 May 2007
Autors

- Jens Klump¹, Robert Huber², Jan Brase³, Michael Diepenbroek², Hannes Grobe⁴, Beate Hildenbrand⁵, Heinke Höck⁶, Michael Lautenschlager⁶, Uwe Schindler², Irina Sens³ and Joachim Wächter¹

- GFZ Potsdam (proposed WDC-TERRA)
- WDC-MARE, Univ. Bremen
- TIB Hannover (Nat. Lib. Sci. & Tech. Germany)
- WDC-MARE, AWI Bremerhaven
- WDC-RSAT, DLR-DFD Oberpfaffenhofen
- WDC-Climate, MPI-MET Hamburg
Data publication today

Table 6. Development of various data algorithms (see also Table 4) applied on SeaWiFS data in July 2002 (2003).

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>2002</th>
<th>2001 case 1</th>
<th>2001 case 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st Order</td>
<td>17</td>
<td>18</td>
<td>18</td>
</tr>
<tr>
<td>2nd Order</td>
<td>15</td>
<td>17</td>
<td>17</td>
</tr>
<tr>
<td>Mean</td>
<td>1.6</td>
<td>1.35</td>
<td>1.3</td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>0.6</td>
<td>1.2</td>
<td>1.2</td>
</tr>
<tr>
<td>Median</td>
<td>1.6</td>
<td>1.2</td>
<td>1.2</td>
</tr>
<tr>
<td>Mode</td>
<td>1.6</td>
<td>1.3</td>
<td>1.3</td>
</tr>
</tbody>
</table>

According to ground truth and SeaWiFS spectra, the highest percentiles of Lake Baikal in 2001-2002 were similar. However, the in situ chlorophyll concentrations determined by different algorithms during field expeditions in Lake Baikal during 2001 and 2002. Values of measured chlorophyll (HPLC) are the mean concentrations of each sampling point from 5 to 30 m depth. For the OC2 chl-a calculations, the most cloud-free acquisitions in 2001 (2001/07/19) and 2002 (2002/07/20) were chosen. Note the considerable chl-a overestimation caused by the influences of terrigenous input in case 2 waters.
Use of Published Data

• Often, the source of data is not acknowledged.
• No citation of the data source.
• The data source needs to be deduced from the paper. No Metadata.

Acknowledgements. We thank H.-F. Tsai, T. Horinouchi, T. Nakamura, Y. Shibagaki for their fruitful discussions and comments on the manuscript. We also wish to thank GFZ (GeoForschungsZentrum), Potsdam for providing CHAMP/GPS RO data through the ISDC (INTEGRAL Science Data Centre ISDC) data center. One of the authors (MVR) wishes to thank JSPS (Japan Society for Promotion of Sciences) for providing fellowship to carry out this work. This work is also supported partially by MEXT (Japanese Ministry of Education, Culture, Sports, Science and Technology) using Grant-in-Aid for Scientific Research on Priority Areas (Grant number: (A03) 13136206; (A04) 13136203).

Topical Editor U.-P. Hoppe thanks two referees for their help in evaluating this paper.
Data in the publication process today

After Helly et al. (2003)
The consequences

- Most data remain underutilised because they are not accessible.
  → Unnecessary duplication
- Research results cannot be verified.
  → Falsification of results.
- Calls to make data accessible and share data were welcomed but did not give any results.
Why are data not made accessible?

- Data publication is hampered by structural barriers in the publication process:
  - Journals do not devote space to data tables due to economic constraints and have no interest in archiving data.
  - Authors do not receive professional recognition for publishing data because the datasets cannot be cited in a reliable way.
  - Data are not cited because their location (URL), in many cases, is transient.
Necessary steps

- Data need to be citeable to be “valuable”.
  - “Reputation” is the currency of science.
- Authors will only prepare data for publication if the effort is worthwhile.
  - Data publication is labour intensive.
- Data must be accessible to be re-used.
  - Access through persistent identifiers and long-term archives.
- Existence of data must be known.
  - Dissemination of metadata to catalogues and portals.
- Intellectual property rights need to be secured.
  - Authors need full control of their publications.
Project “Publication and Citation of Scientific Primary Data”

- Funded by the German Science Foundation.
- Project partners:
  - German Nat. Lib. Science and Technology (Hannover)
  - WDC-MARE (Bremen/Bremerhaven)
  - WDC Climate (Hamburg)
  - GFZ Potsdam (proposed WDC-TERRA)
  - WDC-RSAT (Oberpfaffenhofen)
- Implementation of services for the publication of data.
- DOI registration agency at German National Library for Science and Technology (TIB Hannover).
- To date 6 DOI registration agents. Inclusion of data publications into library catalogues.
What is a DOI?

- DOI = Digital Object Identifier, a persistent, digital identifier of an object.
- DOI = Name of object, URL = Location of object.
- The location may change, the name persists, irrespective of the location of the object.
STD-DOI System Architecture
Fig. 2. The scattergram shows the relationship between concentrations of chl-a calculated from SeaWiFS OC2 and chl-a calculated determined from ground truth measurements during field expeditions in Lake Baikal during 2001 and 2002. Values of measured chlorophyll (HPLC) are the mean concentrations of each sampling point from 5 to 30 m depth. For the OC2 chl-a calculations, the most cloud-free acquisitions in 2001 (2001/07/19) and 2002 (2002/07/20) were chosen. Note the considerable chl-a overestimation caused by the influences of terrigenous input in case 2 waters.
Title: The relationship between concentrations of chl-a calculated from SeaWIFS OC2 and chl-a calculated determined from ground truth measurements during field expeditions in Lake Baikal during 2001 and 2002 / Geoforschungszentrum Potsdam (GFZ), Potsdam, Germany. Birgit Helm

Collaborator: Birgit Helm; Heidi Oberfringsli; Susanne Fleitz; Hermann Kaufmann

Corporate body: Geoforschungszentrum Potsdam (GFZ)

Published: 2006-01-15

Extant: Online-Resource (94 Datapoints)

Note: Abstract

StructuralType: Digital

CreationDate: 2006-03-09

Abstract: Values of measured chlorophyll (HPLC=High Performance Liquid Chromatography) are the mean concentrations of each sampling point from 0 to 20 m depth. For the OC2 chl-a calculations, the most cloud-free acquisitions in 2001 (2001/07/19) and 2002 (2002/07/20) were chosen. Note the considerable chl-a overestimation caused by the influences of terrigenous input in case 2 waters.

Techn. data: Format: text/tab-separated-values

DOI: 10.1594/GFZ.SDD8.1043

URN: urn:nbn:de:10-1594/GFZ.SDD8.10432

Holding: Display free access

Note: Primaedaten

P f e t t
Citation: 
Heim, Birgit; Oberhänsl, Hedi; Fietz, Susanne; Kaufmann, Hermann; (2006): The relationship between concentrations of chl-a calculated from SeaWiFS OC2 and chl-a calculated determined from ground truth measurements during field expeditions in Lake Baikal during 2001 and 2002, Scientific Drilling Database, 10.1594/GFZ.SDDB.1043

Activities:
**CON01.501.1**
- Latitude: 52.9957
- Longitude: 107
- Elevation: -1280
- Date/Time: 2001-07-16 00:52:00
- Program: High-resolution CONTINENTal paleoclimate record in Lake Baikal
- Expedition: CON01-5
- Platform: R/V Veenenboegen
- Gear: Water sample
- **CON01.502.1**
- Latitude: 52.9931
Data Syndication

- Many available data remain underutilised because their existence is not known.
- Metadata can be harvested and indexed by data portals using OAI-PMH or RSS.
- Both OAI-PMH and RSS can be used to carry discipline specific metadata.
- Examples:
  - PANGAEA Data Portals (OAI-PMH)
  - Open Geospatial Consortium GeoRSS
- Portals can provide specific views on existing data.
DOI metadata

- The STD-DOI metadata are mainly Dublin Core elements, plus system specific elements.
- The metadata transmitted to the National Library via web service (HTTP/SOAP) and incorporated into the library catalogue.
- The metadata may contain references to other objects.
  - Element `<RelatedIdentifier>`
  - `isCitedBy`, `isDuplicateOf`, `isAlsoPublishedAs`, …
The element `<RelatedIdentifier>` can be used to point to other electronic objects:

- Point to the literature where the data set is interpreted.
- Point to samples, from which the data were derived.
- Point to other datasets that belong to the same collection of datasets.
- Improve data discovery.
Information Discovery

Link to publication

Citation of data

IGSN points to sample
SDDB Metadata Management

Management of Metadata is kept simple by offering a data upload assistant.

Data upload process is styled in analogy to eBay sales upload assistant.

Most metadata are in the system already.
Internal Semantics

- SDDB metadata are held in a fully normalised relational database.
- SDDB metadata are fully browseable to allow iterative search.
- SDDB has (so far) no full-text search.
Putting data into context

- Parameters are organised in a tree-structure.
- Homonyms are allowed – the parameter tree is ordered according to scientific context.
The spatial dimension

Will soon be supplemented by OGC Web Services.
SDDB Plans for 2007/2008

- Integration of GSI components (deegree2).
- Introduction of links to IGSNs.
- Migration of web frontend from PHP to Java/JSP/JSR168.
- Introduction of Fedora Repository as middleware to integrate data, publications and services.
More Information

- Project “Publication and Citation of Scientific Primary Data”
  
  http://www.std-doi.de

- ICDP Scientific Drilling Database
  
  http://www.scientificdrilling.org

Thank you!

This work is licensed under a Creative Commons by Attribution 2.0-de License.