DOIs for Tracking and Citing Scientific Data

J. Klump, J. Wächter and M. Lautenschlager
CODATA Conference 2006
Beijing, PR China
Data publication today

According to ground truth and SeaWiFS spectra for 2001-2002, the green peak of the highly renewing waters of Lake Baikal is commonly located in the SeaWiFS band at 555 nm. However, the observed and modeled spectral absorptions in the presence of the sea surface winds and turbidity in the water suggest that the observed spectral absorptions are not as inflected, yet more sensitive to the impact of the SeaWiFS bands at 555 nm. The waters in the observed clear-water parts of the SeaWiFS acquisition are not as high, yet more sensitive to the impact of the SeaWiFS bands at 555 nm. The observed spectral absorptions of the peak shifting from 510 to 555 nm in the 2001-2002 SeaWiFS data sets of Lake Baikal can be correlated and reproduced using the bi-optical software “Water Colour Simulation” (WCSIM) (Kirsanov, 2003). This described spectral behaviour has been already observed from previous historical spectroscopic studies. For example, Hargrave and Jerlov (1970) related that the waters of Lakes Ontario and Superior (USA) had a dominant wavelength of 500-550 nm, benthically more productive waters had a dominant wavelength of 550-600 nm, and waters with heavy sediment loadings had a dominant wavelength of 600 nm.

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.
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.
Specific situation at GFZ Potsdam

- GFZ produces not only closed data sets but also time series from monitoring systems and observatories.
  - Satellites (CHAMP, GRACE, future missions)
  - Earth magnetic field variations
  - Seismology
  - Geodetic services (e.g. rotation, GPS baseline)
- Operation of these systems is labour intensive but is not fully appreciated in the scientific literature.
Example CHAMP

- No citation, only acknowledgement.
- The data sources need to be deduced from the paper. No Metadata.
- Often the source of data is not acknowledged.

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.
Why data are 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.
  - Access through persistent identifiers and long-term archives.
- Intellectual property rights need to be secured.
  - Authors need full control over their publications.
Project “Publication and Citation of Scientific Primary Data”

- Funded by the German Science Foundation.
- 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.
- Project partners:
  - WDC-MARE (Bremen/Bremerhaven)
  - WDC Climate (Hamburg)
  - GFZ Potsdam (proposed WDC-TERRA)
  - WDC-RSAT (Oberpfaffenhofen)
- Inclusion of data publications into library catalogues.
Was is a DOI?

- DOI = Digital Object Identifier, a persistent, digital identifier for an object.
- DOI = Name of object, URL = Location of object.
- The location may change, the name persists, irrespective of the location of the object.
- Global resolving mechanism (handle.net) “translates” DOIs to URLs.
STD-DOI system architecture

Diagram showing the architecture with the following components:

- International DOI Foundation
- Global Handle System
- TIB Hannover Registration Agency
- GFZ Geophysics
- M+D / MPIM Climate Models
- Marum / AWI Observations
- Data Storage Long-term Archiving
- Data Storage Long-term Archiving in WDC
- Data Storage Long-term Archiving in WDC

Connections show the flow and relationships between these components.
System architecture at GFZ

Library

TIBORDER / GBV Catalogue

TIB Hannover

DOIDB

RSS

Data Source

Data Source

Data Source

Data Source
Fig. 2. The scattergram shows the relationship between concentrations of chl-α calculated from SeaWiFS OC2 and chl-α 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-α calculations, the most cloud-free acquisitions in 2001 (2001/07/19) and 2002 (2002/07/20) were chosen. Note the considerable chl-α 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

Collaborator: GeoForschungsZentrum Potsdam (GFZ)

Published: 2006-09-15

Extent: Online-Resource (54 Datapoints).

Abstract: Values of measured chlorophyll (HPLC=High Performance Liquid Chromatography) 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.

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

Links: doi: 10.1594/GFZ.SDDB.1043
URN: urn:nbn:de:bib-10.1594/GFZ.SDDB.1043

Holding: Display free access!
Note: Primaerdaten
How to cite data

Citation: Heim, Birgit; Oberhansli, 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

Download Citation (EndNote)

DOI: 10.1594/GFZ.SDDB.1043
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
Abstract: Values of measured chlorophyll (HPLC=High Performance Liquid Chromatography) are the mean concentrations of each sampling point from 0 to 50 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.

Activities:
- CON01-501-1
  - Lat/Long: 52.6987/107
  - Elevation: -1260
  - Date/Time: 2001-07-15 00:52:00
  - Program: High-resolution CONTINENTal paleoclimate record in Lake Baikal
  - Expedition: CON01-5
  - Platform: R/V Verevshchegin
  - Gear: Water sample

- CON01-502-1
  - Lat/Long: 52.9561
  - Date/Time: 2001-07-15 00:52:00
  - Program: High-resolution CONTINENTal paleoclimate record in Lake Baikal
  - Expedition: CON01-5
  - Platform: R/V Verevshchegin
  - Gear: Water sample
Fair Use

- “Fair Use” of electronic sources is one of the most contested issues surrounding the use of the internet.
- Scientific publication are acknowledged by a “citation”.
- The citation is part of good scientific conduct.
- In that sense, data publications are analogous to “classical” publications.
The Creative Commons Licence

- Toolbox for the configuration of a custom licence.
- Our recommendation for scientific data:
  - By attribution (citation)
  - Non-commercial
  - Share alike (derivative works have to be published under the same licence)
Questions remain

- Data publication attempts to change existing scientific practice.
- How does review of data publications work?
- What do trusted data repositories look like?
- What are the requirements of different scientific disciplines?
Thank you!

Please visit our project website at

http://www.std-doi.de

Thank you for your attention!