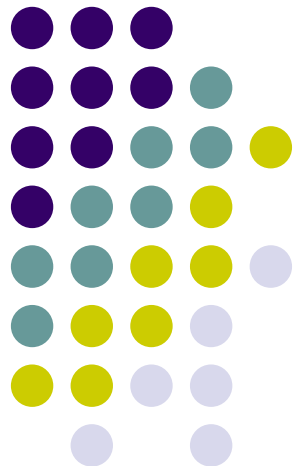
# DOIs for Tracking and Citing Scientific Data

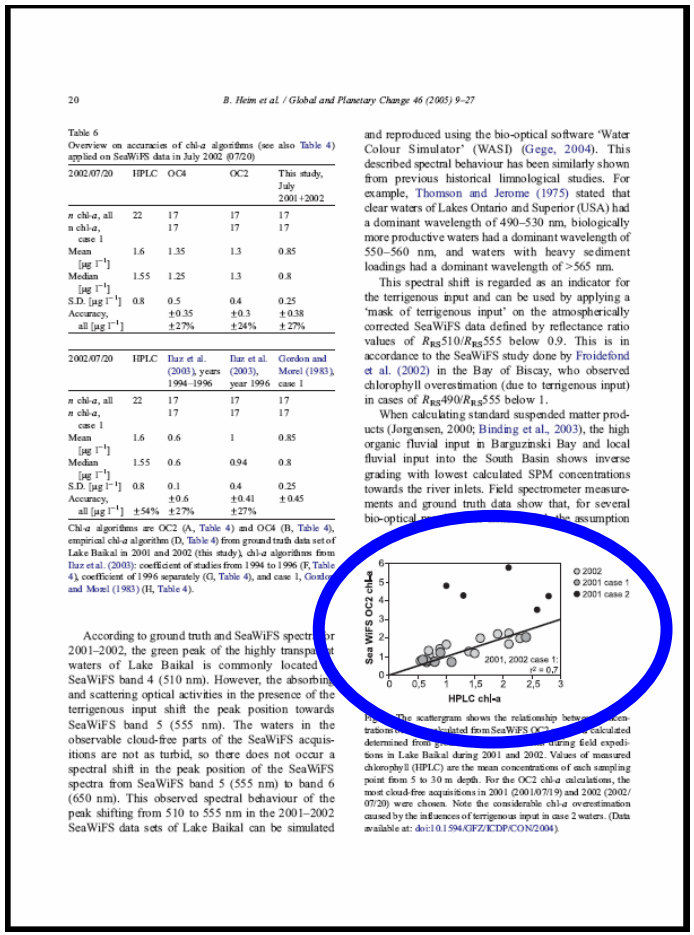
J. Klump, J. Wächter and M. Lautenschlager

CODATA Conference 2006

Beijing, PR China



# Data publication today



20 B. Heim et al. / Global and Planetary Change 46 (2005) 9–27

Table 6
Overview on accuracies of chl-*a* algorithms (see also Table 4) applied on SeaWiFS data in July 2002 (07/20)

| 2002/07/20 | HPLC | OC4 | OC2 | This study, July 2001+2002 |
|---|---|---|---|---|
| *n* chl-*a*, all | 22 | 17 | 17 | 17 |
| *n* chl-*a*, case 1 | | 17 | 17 | 17 |
| Mean [μg l$^{-1}$] | 1.6 | 1.35 | 1.3 | 0.85 |
| Median [μg l$^{-1}$] | 1.55 | 1.25 | 1.3 | 0.8 |
| S.D. [μg l$^{-1}$] | 0.8 | 0.5 | 0.4 | 0.25 |
| Accuracy, all [μg l$^{-1}$] | | ±0.35 ±27% | ±0.3 ±24% | ±0.38 ±27% |

| 2002/07/20 | HPLC | Ilaz et al. (2003), years 1994–1996 | Ilaz et al. (2003), year 1996 | Gordon and Morel (1983), case 1 |
|---|---|---|---|---|
| *n* chl-*a*, all | 22 | 17 | 17 | 17 |
| *n* chl-*a*, case 1 | | 17 | 17 | 17 |
| Mean [μg l$^{-1}$] | 1.6 | 0.6 | 1 | 0.85 |
| Median [μg l$^{-1}$] | 1.55 | 0.6 | 0.94 | 0.8 |
| S.D. [μg l$^{-1}$] | 0.8 | 0.1 | 0.4 | 0.25 |
| Accuracy, all [μg l$^{-1}$] | ±54% | ±0.6 ±27% | ±0.41 ±27% | ±0.45 |

Chl-*a* algorithms are OC2 (A, Table 4) and OC4 (B, Table 4), empirical chl-*a* algorithm (D, Table 4) from ground truth data set of Lake Baikal in 2001 and 2002 (this study), chl-*a* algorithms from Ilaz et al. (2003): coefficient of studies from 1994 to 1996 (F, Table 4), coefficient of 1996 separately (G, Table 4), and case 1, Gordon and Morel (1983) (H, Table 4).

According to ground truth and SeaWiFS spectra of 2001–2002, the green peak of the highly transparent waters of Lake Baikal is commonly located at SeaWiFS band 4 (510 nm). However, the absorbing and scattering optical activities in the presence of the terrigenous input shift the peak position towards SeaWiFS band 5 (555 nm). The waters in the observable cloud-free parts of the SeaWiFS acquisitions are not as turbid, so there does not occur a spectral shift in the peak position of the SeaWiFS spectra from SeaWiFS band 5 (555 nm) to band 6 (650 nm). This observed spectral behaviour of the peak shifting from 510 to 555 nm in the 2001–2002 SeaWiFS data sets of Lake Baikal can be simulated and reproduced using the bio-optical software 'Water Colour Simulator' (WASI) (Gege, 2004). This described spectral behaviour has been similarly shown from previous historical limnological studies. For example, Thomson and Jerome (1975) stated that clear waters of Lakes Ontario and Superior (USA) had a dominant wavelength of 490–530 nm, biologically more productive waters had a dominant wavelength of 550–560 nm, and waters with heavy sediment loadings had a dominant wavelength of >565 nm.

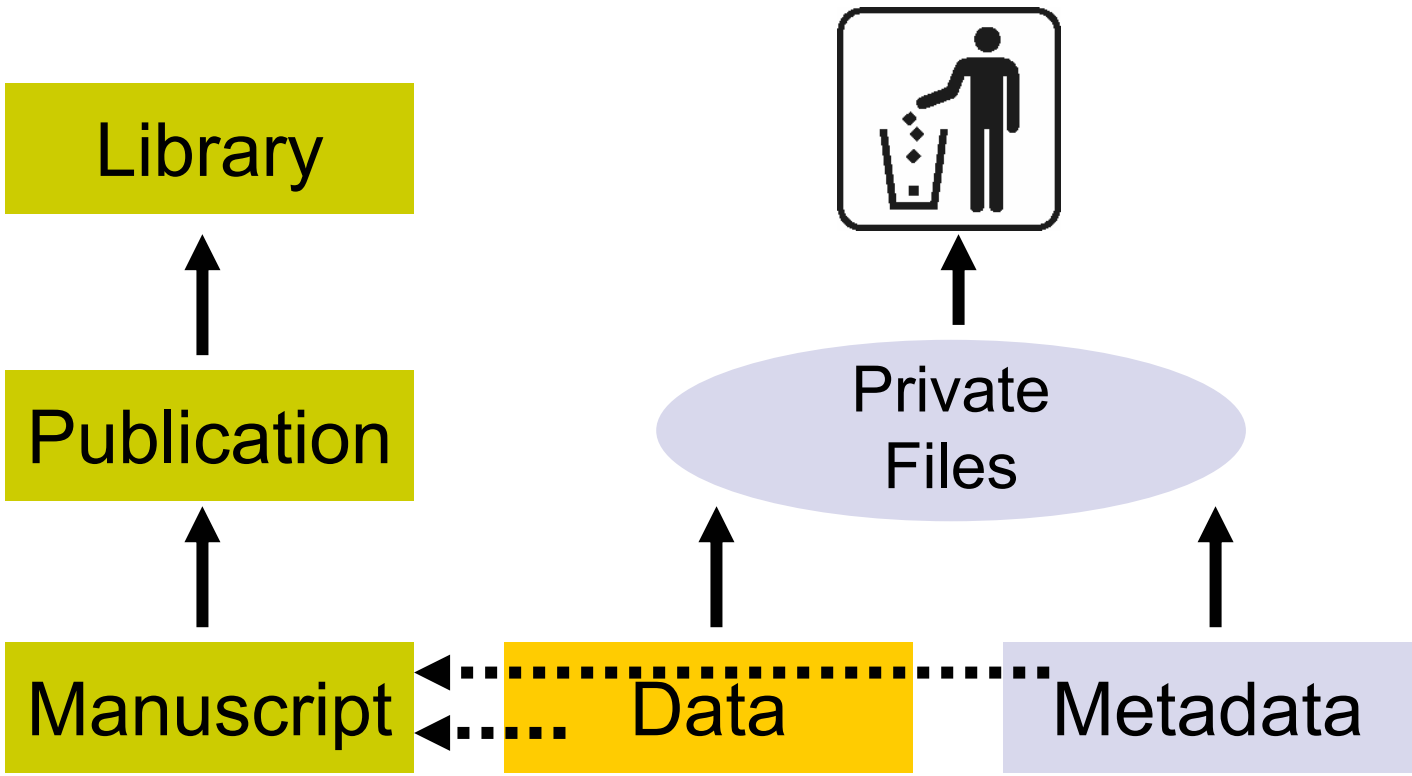
This spectral shift is regarded as an indicator for the terrigenous input and can be used by applying a 'mask of terrigenous input' on the atmospherically corrected SeaWiFS data defined by reflectance ratio values of $R_{RS}510/R_{RS}555$ below 0.9. This is in accordance to the SeaWiFS study done by Froidefond et al. (2002) in the Bay of Biscay, who observed chlorophyll overestimation (due to terrigenous input) in cases of $R_{RS}490/R_{RS}555$ below 1.

When calculating standard suspended matter products (Jørgensen, 2000; Binding et al., 2003), the high organic fluvial input in Barguzinski Bay and local fluvial input into the South Basin shows inverse grading with lowest calculated SPM concentrations towards the river inlets. Field spectrometer measurements and ground truth data show that, for several bio-optical [illegible] the assumption



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)

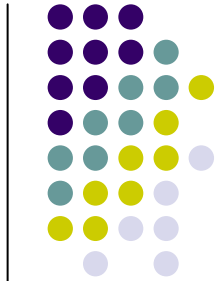
# The consequences

- 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.

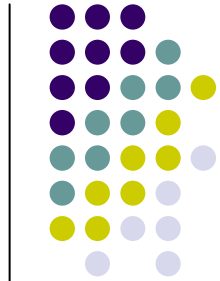
# Example CHAMP

*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.

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

# 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.

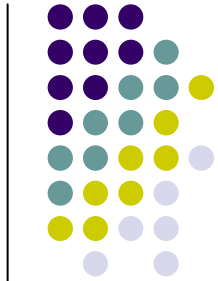
# Necessary steps

- 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 indentifiers and long-term archives.
- Intellectual property rights need to be secured.
  - Authors need full control other their publications.

# Project “Publication and Citation of Scientific Primary Data”

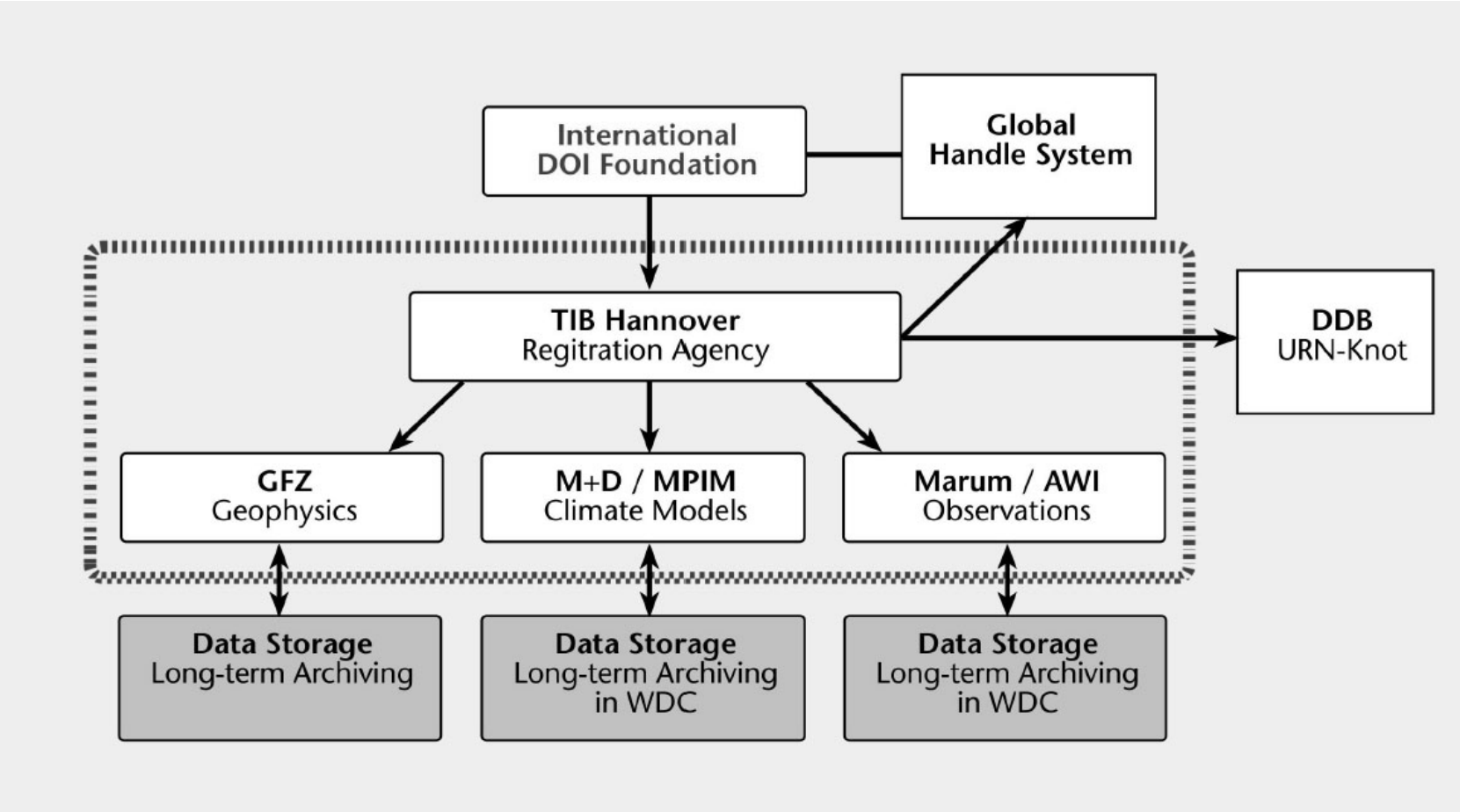
- Funded by the German Science Foundation.
- Implementation of services for the publication of data.
- DOI registration agency at German National Library for Science and Technology (TIB Hannover).
- To date 6 DOI registration agents.
- Project partners:
  - WDC-MARE (Bremen/Bremerhaven)
  - WDC Climate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

# Was is a DOI?

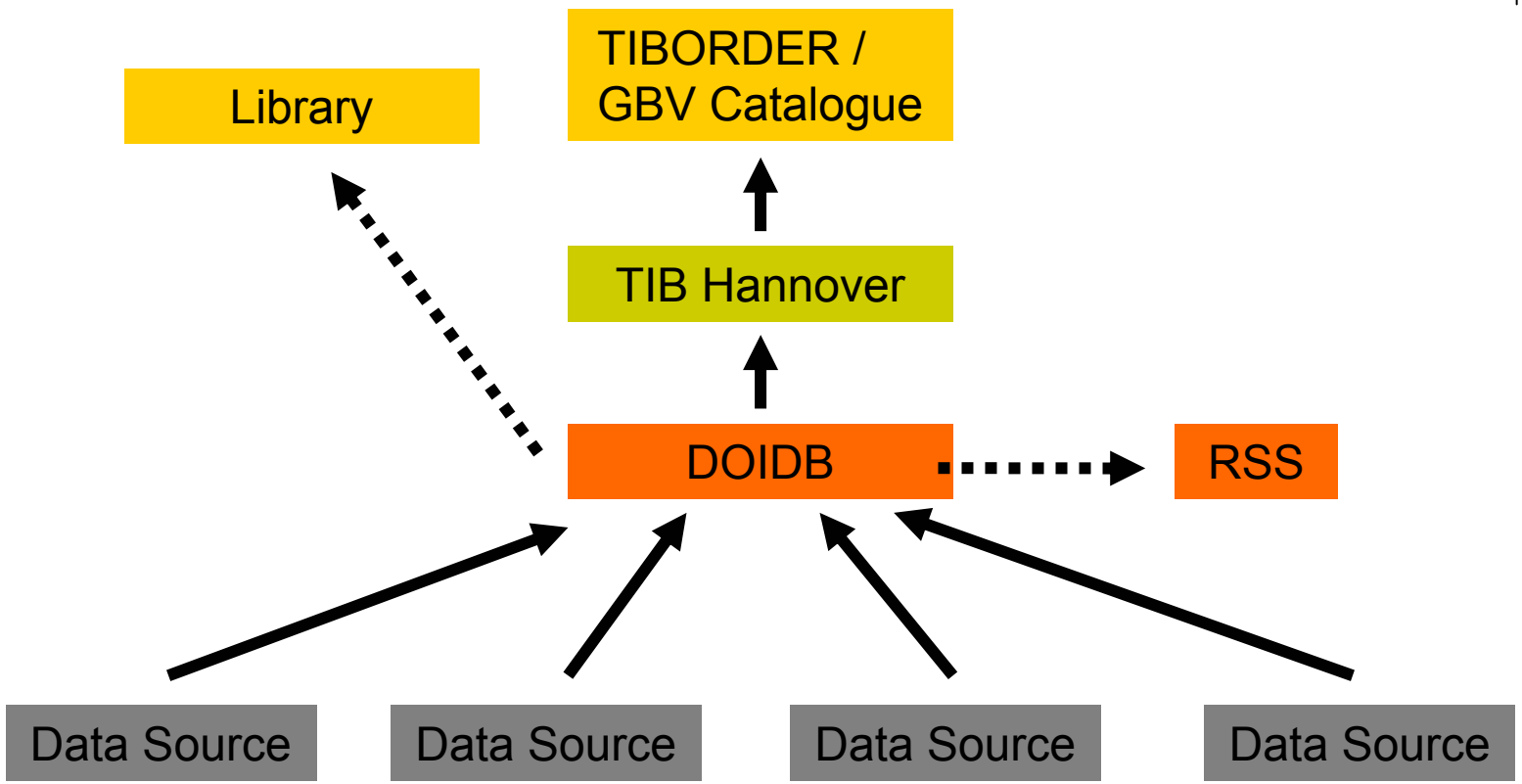


- DOI = Digital Object Identifier, a persistent, digital identifier for an object.
- DOI = Name of object, URL = Location of object.
- The location may change, the name persists, irrespective of the location of the object.
- Global resolving mechanism (handle.net) “translates” DOIs to URLs.

# STD-DOI system architecture



# System architecture at GFZ



# Example

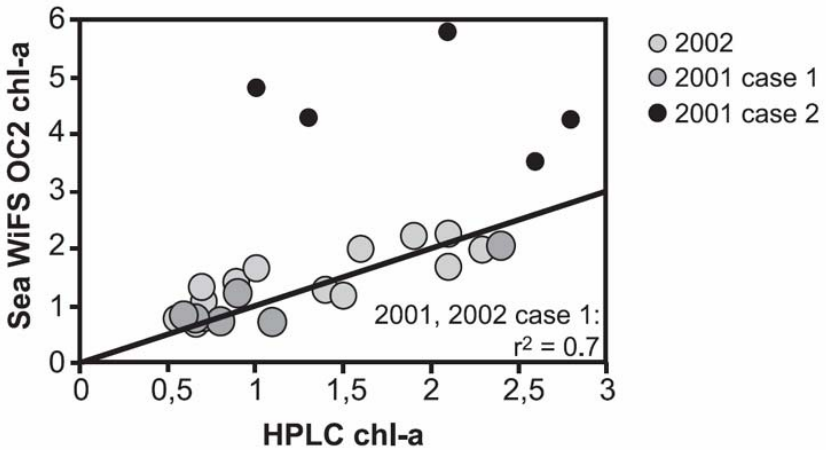
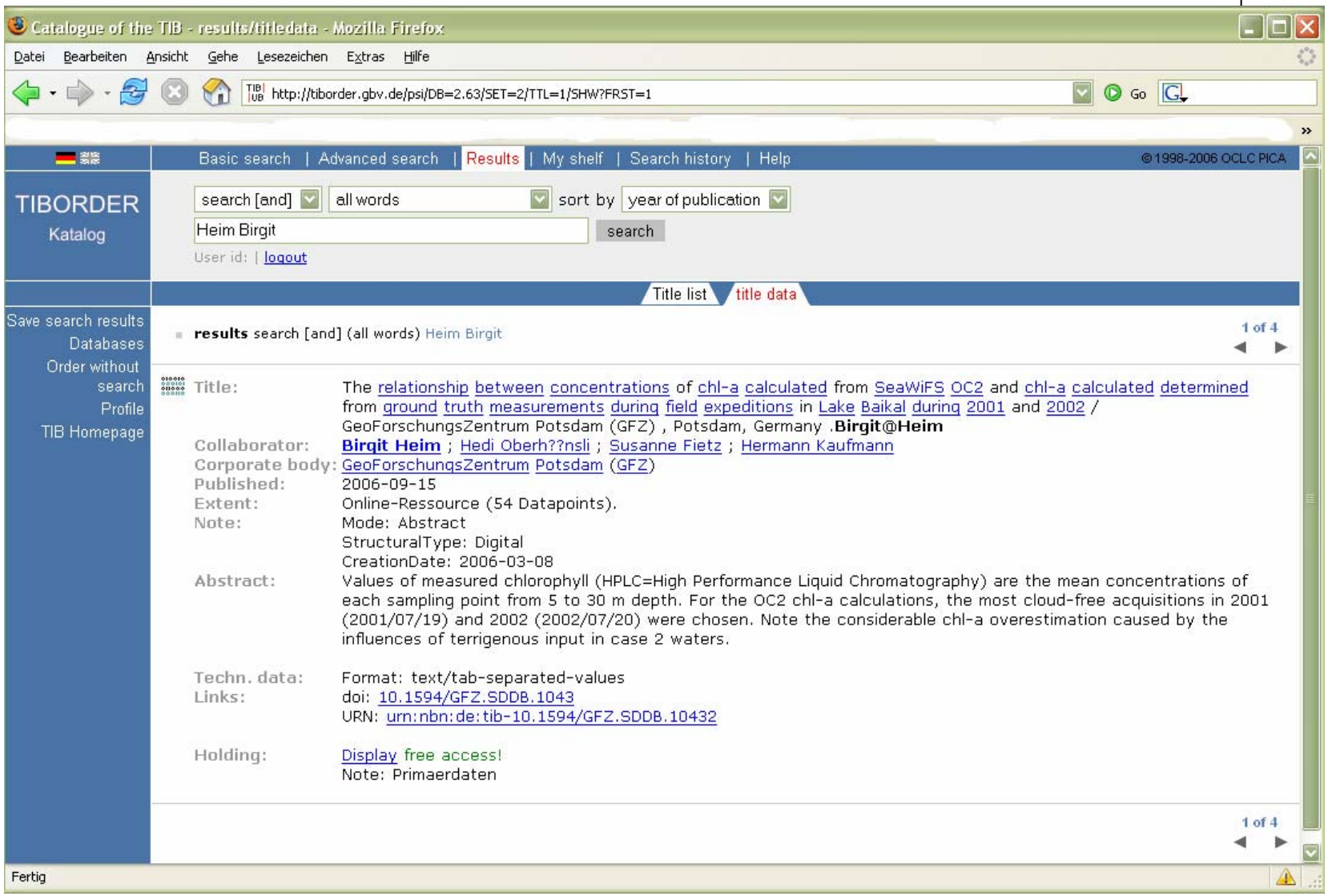



Fig. 2. The scattergram shows the relationship between concentrations of chl-*a* calculated from SeaWiFS OC2 and chl-*a* calculated determined from ground truth measurements during field expeditions in Lake Baikal during 2001 and 2002. Values of measured chlorophyll (HPLC) are the mean concentrations 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.

- GFZ Library
- TIBORDER
- doi:**10.1594/GFZ.SDDB.1043**

# TIBORDER / GBV Catalogue



# How to cite data

Scientific Drilling Database

Data from Deep Earth Sampling and Monitoring

Citation: Heim, Birgit; Oberhänsli, 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)

+ Sampling Gear

+ Parameters

+ Admin

GFZ POTSDAM icdp

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 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.

Show in Google Earth

Activities: CON01-501-1

| | |
|---|---|
| Latidude: | 52.6667 |
| Longitude: | 107 |
| Elevation: | -1250 |
| Date/Time: | 2001-07-16 00:52:00 |
| Program: | High-resolution CONTINENTal paleoclimate record in Lake Baikal |
| Expedition: | CON01-5 |
| Platform: | R/V Vereshchagin |
| Gear: | Water sample |

CON01-502-1

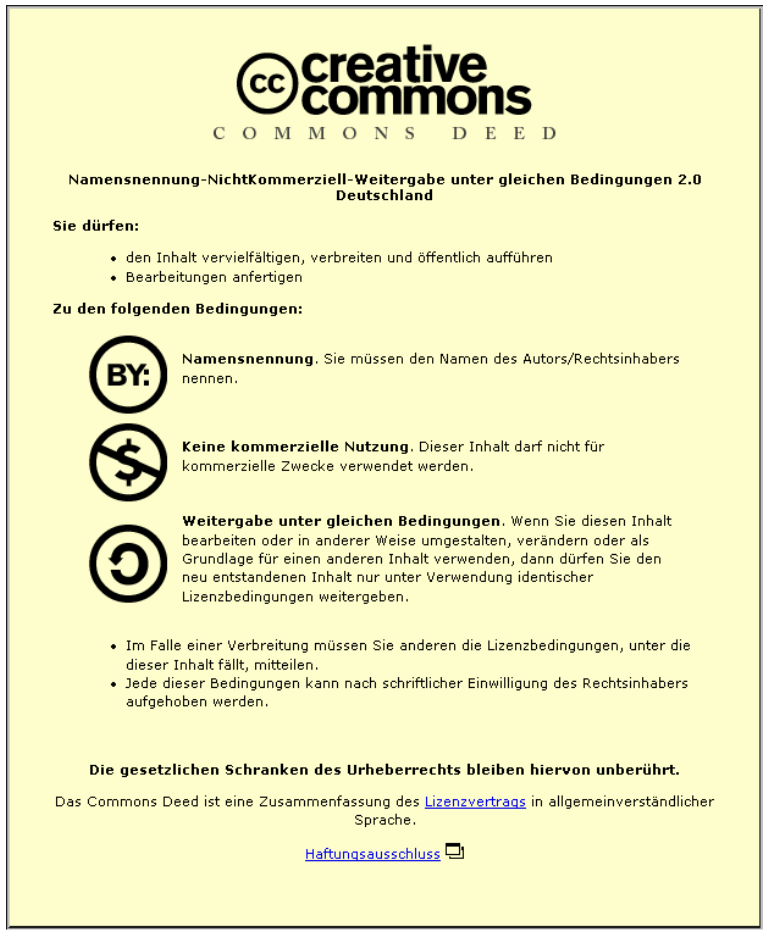
| | |
|---|---|
| Latidude: | 52.9561 |

# 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)



creative commons

COMMONS DEED

**Namensnennung-NichtKommerziell-Weitergabe unter gleichen Bedingungen 2.0 Deutschland**

**Sie dürfen:**

- den Inhalt vervielfältigen, verbreiten und öffentlich aufführen
- Bearbeitungen anfertigen

**Zu den folgenden Bedingungen:**

BY: **Namensnennung**. Sie müssen den Namen des Autors/Rechtsinhabers nennen.

**Keine kommerzielle Nutzung**. Dieser Inhalt darf nicht für kommerzielle Zwecke verwendet werden.

**Weitergabe unter gleichen Bedingungen**. Wenn Sie diesen Inhalt bearbeiten oder in anderer Weise umgestalten, verändern oder als Grundlage für einen anderen Inhalt verwenden, dann dürfen Sie den neu entstandenen Inhalt nur unter Verwendung identischer Lizenzbedingungen weitergeben.

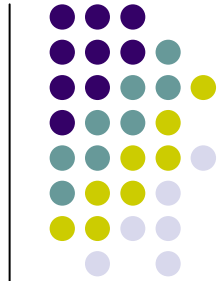
- Im Falle einer Verbreitung müssen Sie anderen die Lizenzbedingungen, unter die dieser Inhalt fällt, mitteilen.
- Jede dieser Bedingungen kann nach schriftlicher Einwilligung des Rechtsinhabers aufgehoben werden.

**Die gesetzlichen Schranken des Urheberrechts bleiben hiervon unberührt.**

Das Commons Deed ist eine Zusammenfassung des Lizenzvertrags in allgemeinverständlicher Sprache.

Haftungsausschluss

# 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!