# Every bit counts

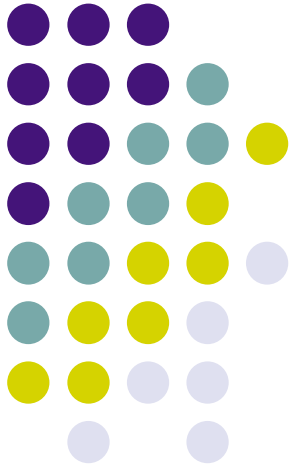
Data management and data publication in the earth sciences

Jens Klump *et al.*

International Data Exchange Workshop
Kiel, 10 May 2007

GFZ POTSDAM

dr.z daten- und rechenzentrum



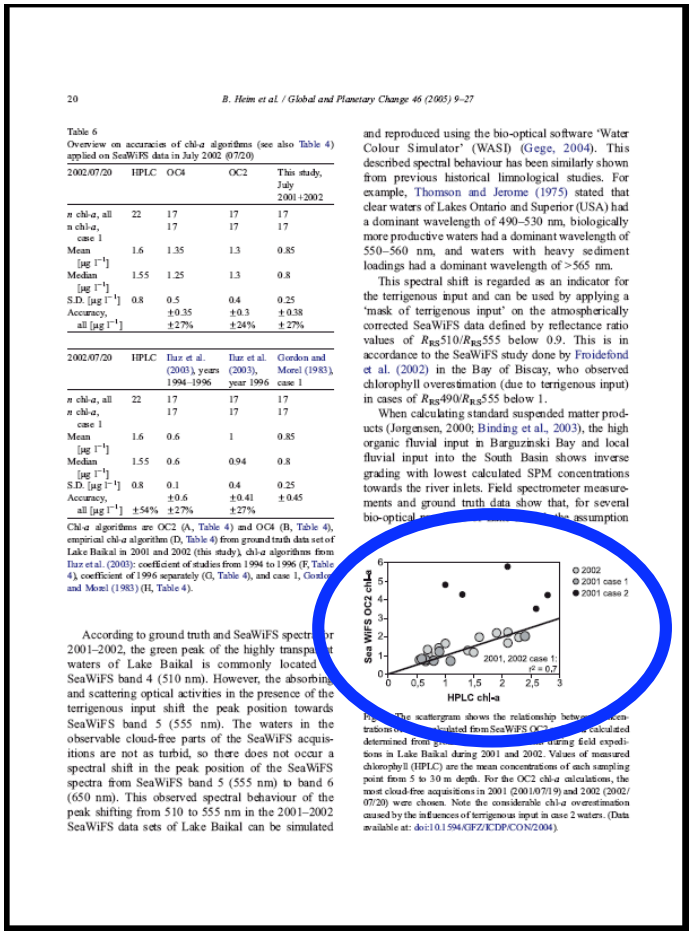
HELMHOLTZ ASSOCIATION

# Autors


- Jens Klump[1], Robert Huber[2], Jan Brase[3], Michael Diepenbroek[2], Hannes Grobe[4], Beate Hildenbrand[5], Heinke Höck[6], Michael Lautenschlager[6], Uwe Schindler[2], Irina Sens[3] and Joachim Wächter[1]

- 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



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 for 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] assumption




[illegible] 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 available at: doi:10.1594/GFZ/ICDP/CON/2004).

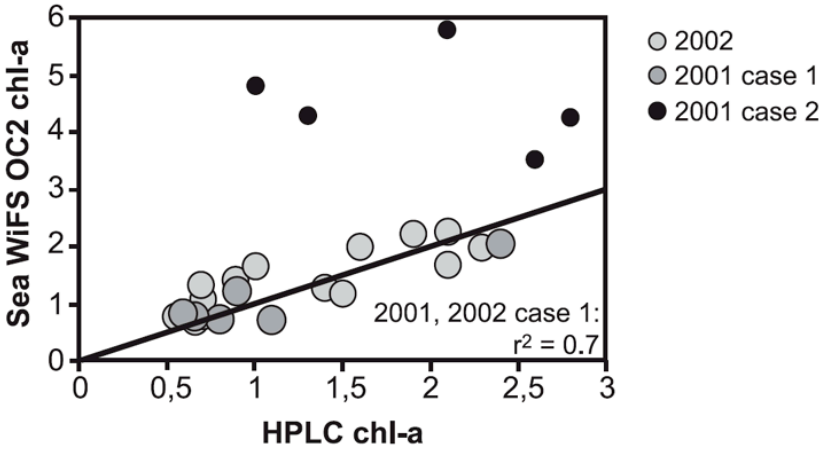



Fig. 2. The scattergram shows the relationship between concentrations of chl-*a* calculated from SeaWiFS OC2 and chl-*a* calculated determined from ground truth measurements during field expeditions in Lake Baikal during 2001 and 2002. Values of measured chlorophyll (HPLC) are the mean concentrations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

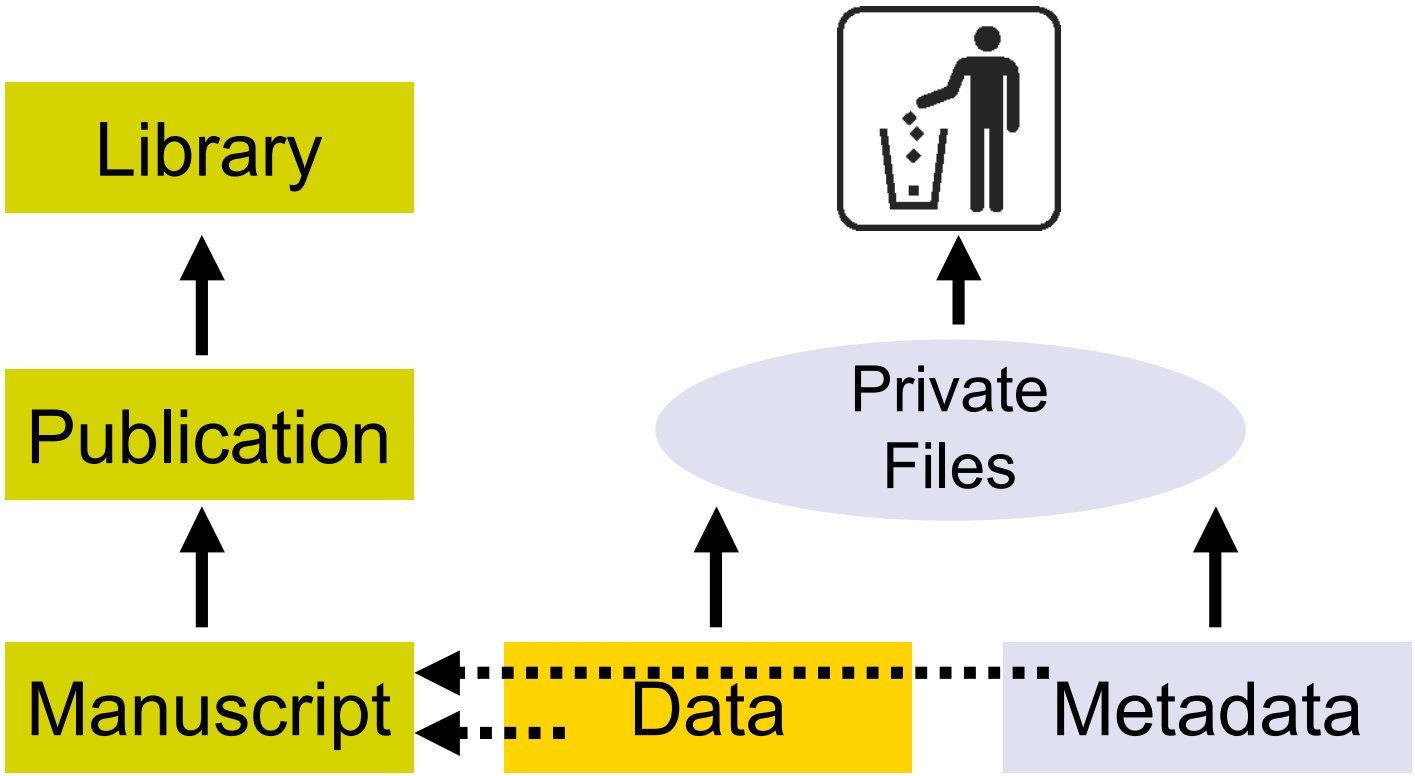
# Use of Published Data

*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.

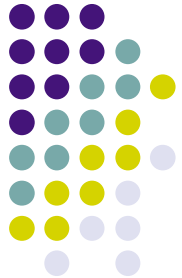
- Often, the source of data is not acknowledged.
- No citation of the data source.
- The data source needs to be deduced from the paper. No Metadata.

# Data in the publication process today




After Helly *et al*. (2003)

# The consequences



- Most data remain underutilised because they are not accessible.

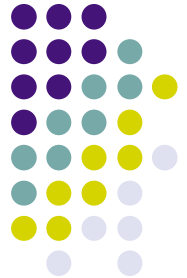
  →Unnecessary duplication

- Research results cannot be verified.

  →Falsification of results.

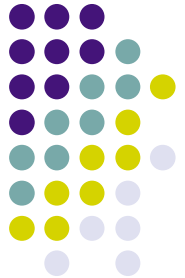
- Calls to make data accessible and share data were welcomed but did not give any results.

# Why are data not made accessible?



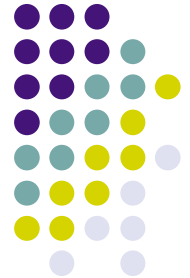
- Data publication is hampered by structural barriers in the publication process:
  - Journals do not devote space to data tables due to economic constraints and have no interest in archiving data.
  - Authors do not receive professional recognition for publishing data because the datasets cannot be 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.

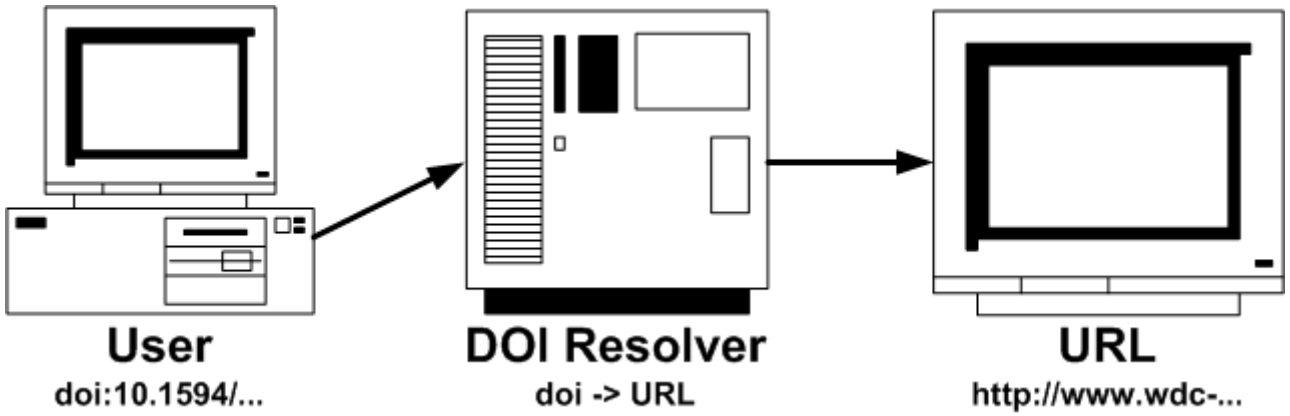
# Project “Publication and Citation of Scientific Primary Data”



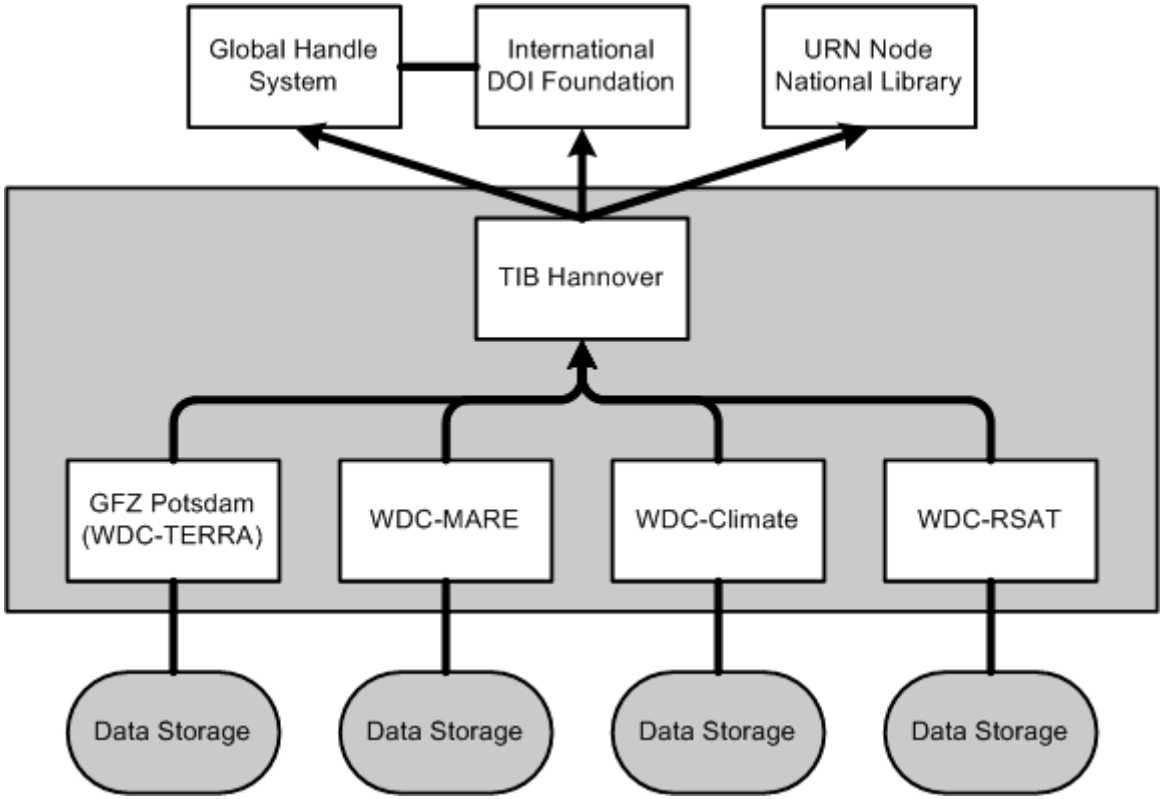
- Funded by the German Science Foundation.
- Project partners:
  - German Nat. Lib. Science and Technology (Hannover)
  - WDC-MARE (Bremen/Bremerhaven)
  - WDC Climate (Hamburg)
  - GFZ Potsdam (proposed WDC-TERRA)
  - WDC-RSAT (Oberpfaffenhofen)
- Implementation of services for the publication of data.
- 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



# Example Data Publication

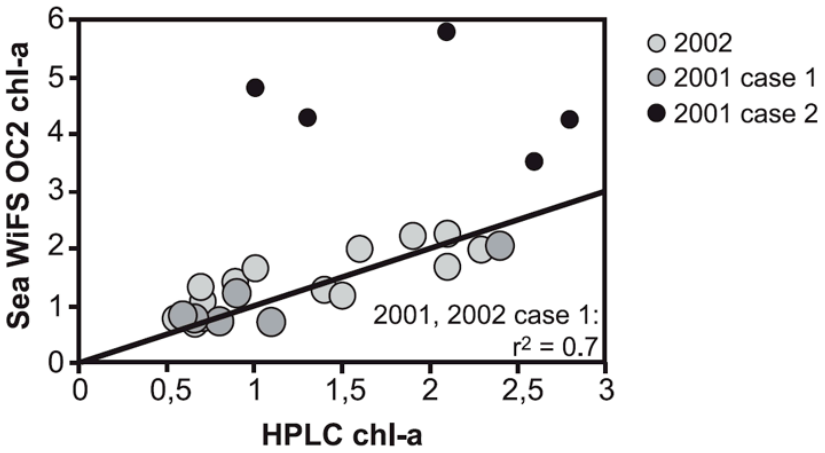
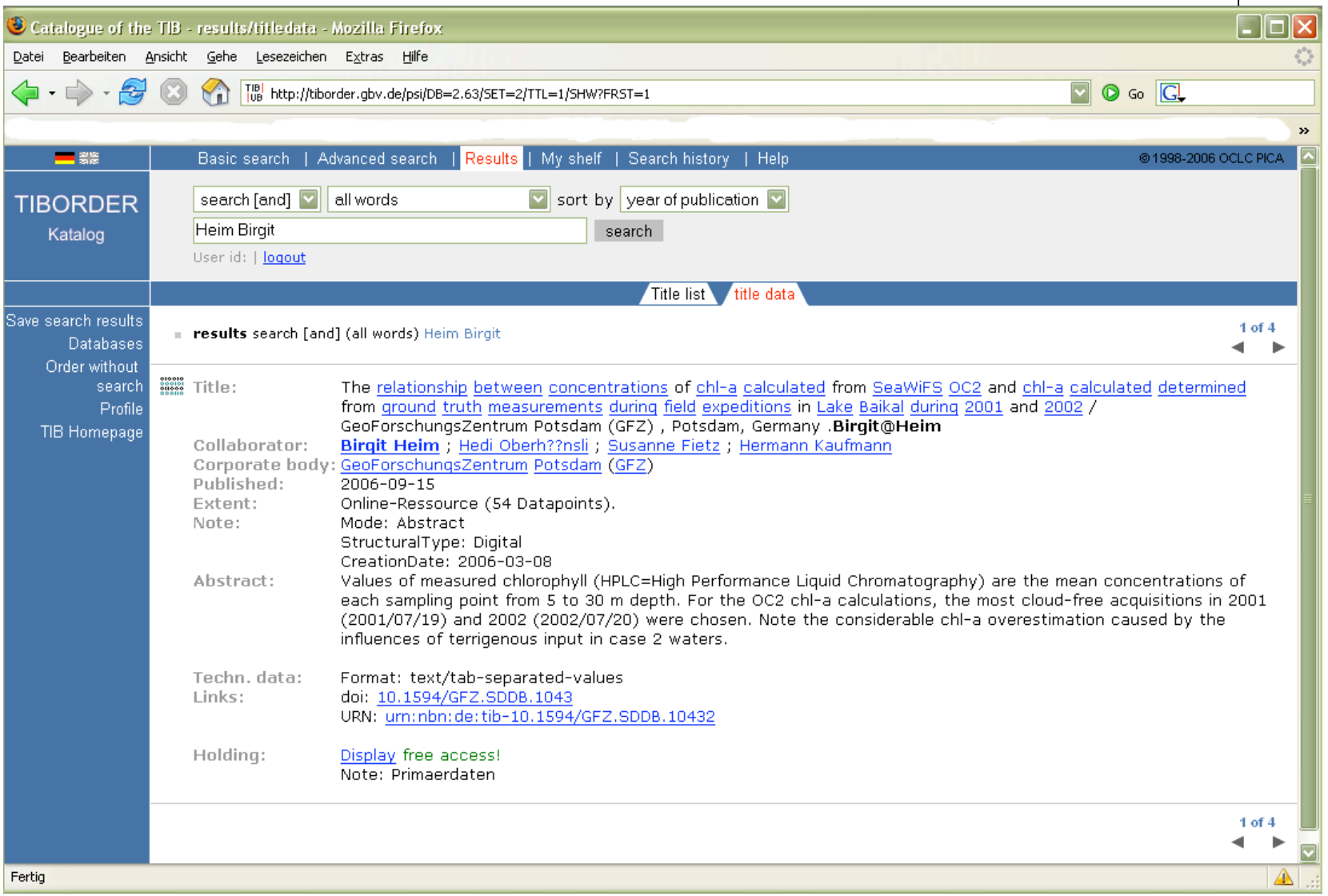


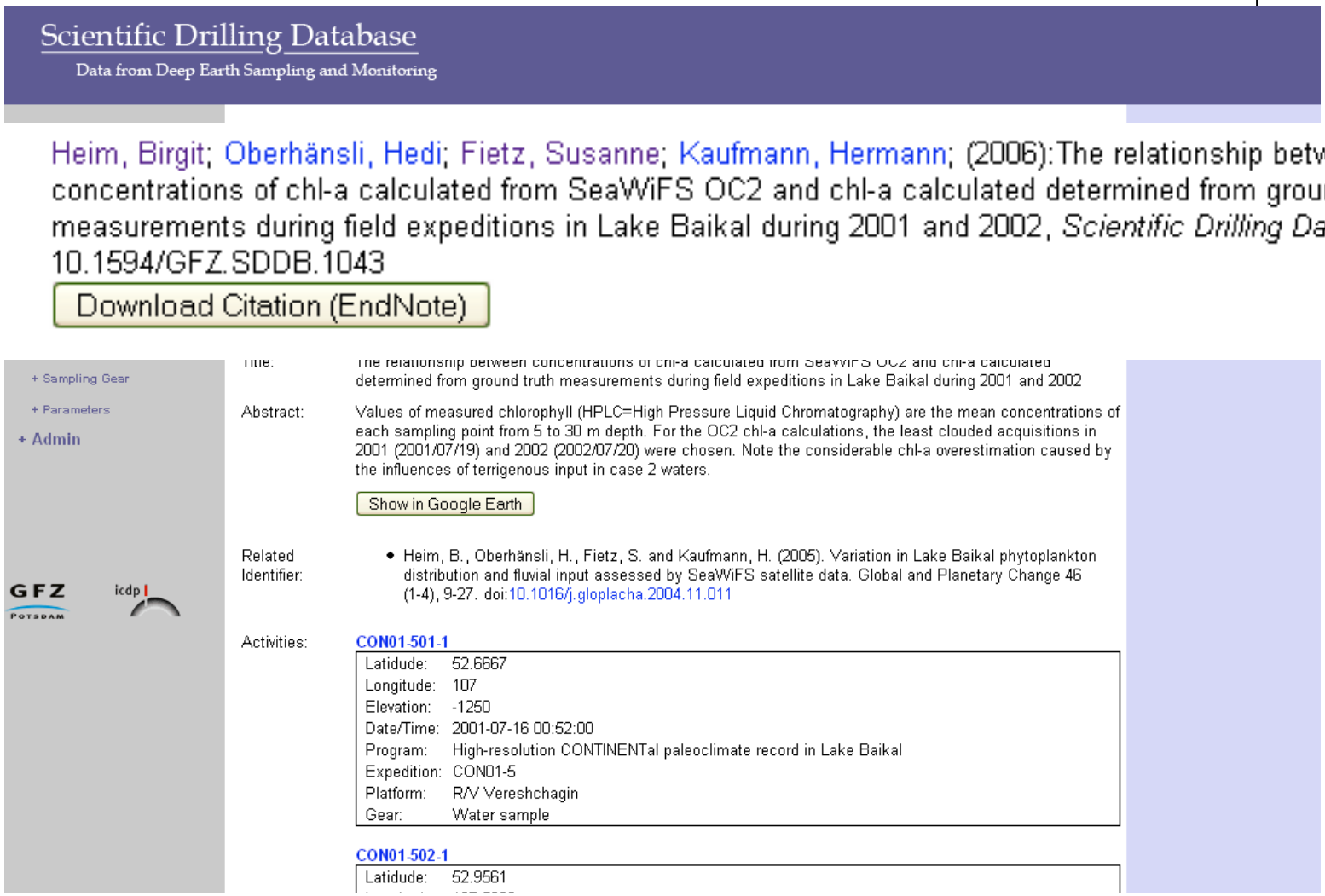

Fig. 2. The scattergram shows the relationship between concentrations of chl-*a* calculated from SeaWiFS OC2 and chl-*a* calculated determined from ground truth measurements during field expeditions in Lake Baikal during 2001 and 2002. Values of measured chlorophyll (HPLC) are the mean concentrations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

- TIBORDER catalogue of the German National Library of Science and Technology.
- doi:10.1594/GFZ.SDDB.1043 at the ICDP Scientific Drilling Database.

# TIBORDER / GBV Catalogue



# ICDP Scientific Drilling Database



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

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 Pressure Liquid Chromatography) are the mean concentrations of each sampling point from 5 to 30 m depth. For the OC2 chl-a calculations, the least clouded 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

Related Identifier:

- Heim, B., Oberhänsli, H., Fietz, S. and Kaufmann, H. (2005). Variation in Lake Baikal phytoplankton distribution and fluvial input assessed by SeaWiFS satellite data. Global and Planetary Change 46 (1-4), 9-27. doi:10.1016/j.gloplacha.2004.11.011

Activities:

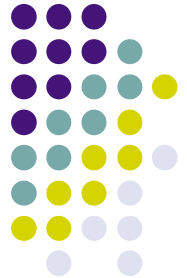
CON01-501-1

| | |
|---|---|
| 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

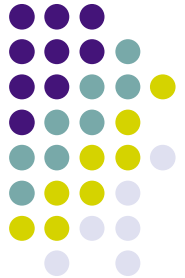
| | |
|---|---|
| Latidude: | 52.9561 |

# Data Syndication 



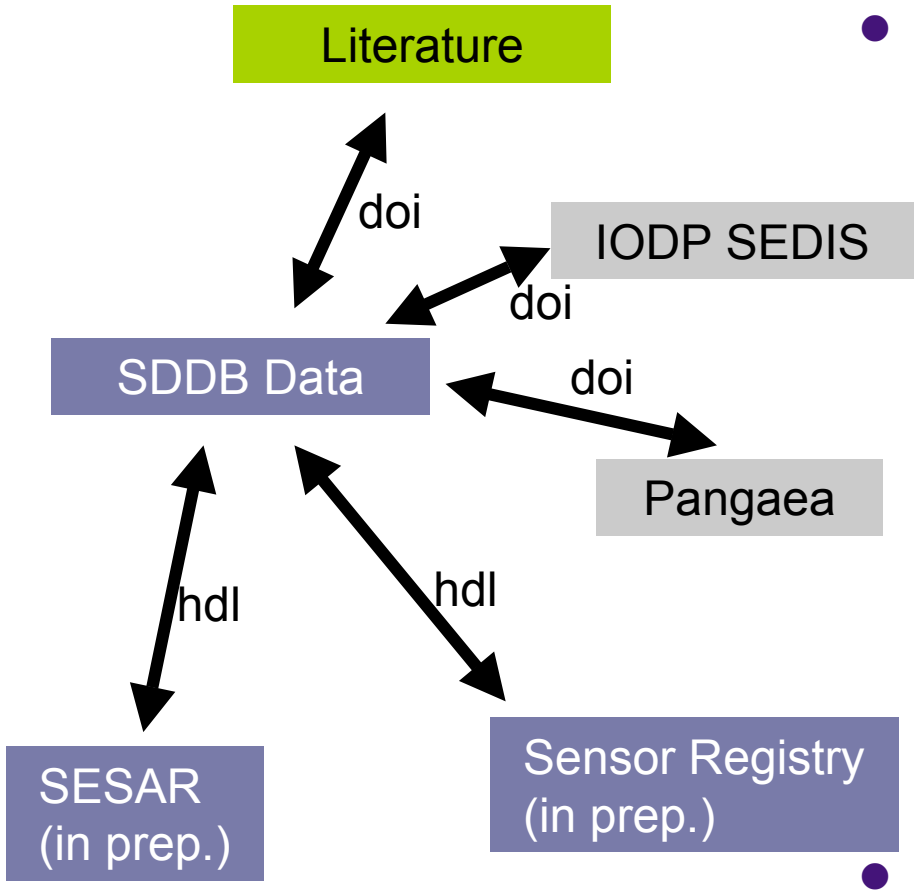
- Many available data remain underutilised because their existence is not known.
- Metadata can be harvested and indexed by data portals using OAI-PMH or RSS.
- Both OAI-PMH and RSS can be used to carry discipline specific metadata.
- Examples:
  - PANGAEA Data Portals (OAI-PMH)
  - Open Geospatial Consortium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

# DOI metadata



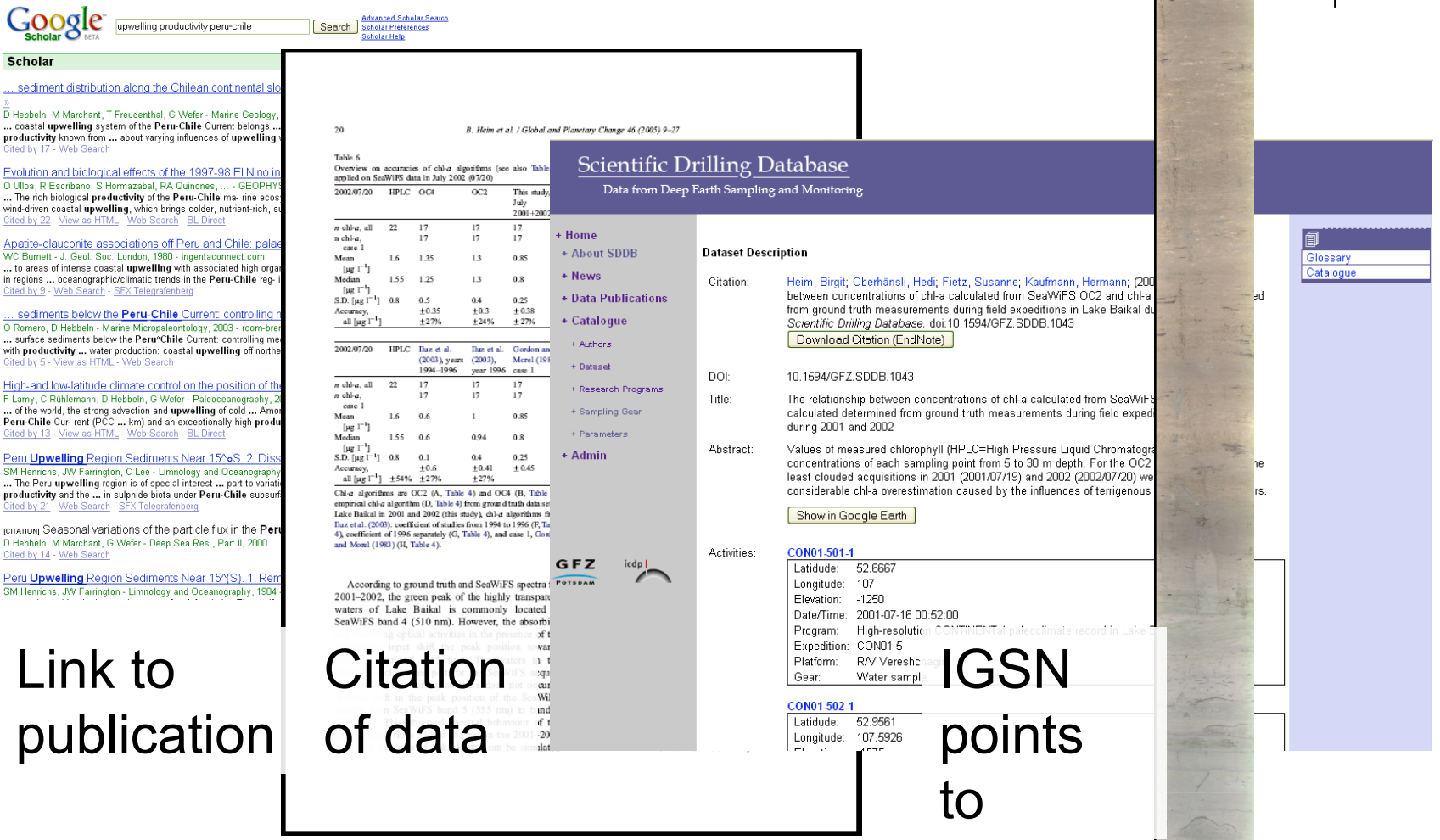
- The STD-DOI metadata are mainly Dublin Core elements, plus system specific elements.
- The metadata transmitted to the National Library via web service (HTTP/SOAP) and incorporated into the library catalogue.
- The metadata may contain references to other objects.
  - Element <RelatedIdentifier>
 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

# External Semantics




- 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

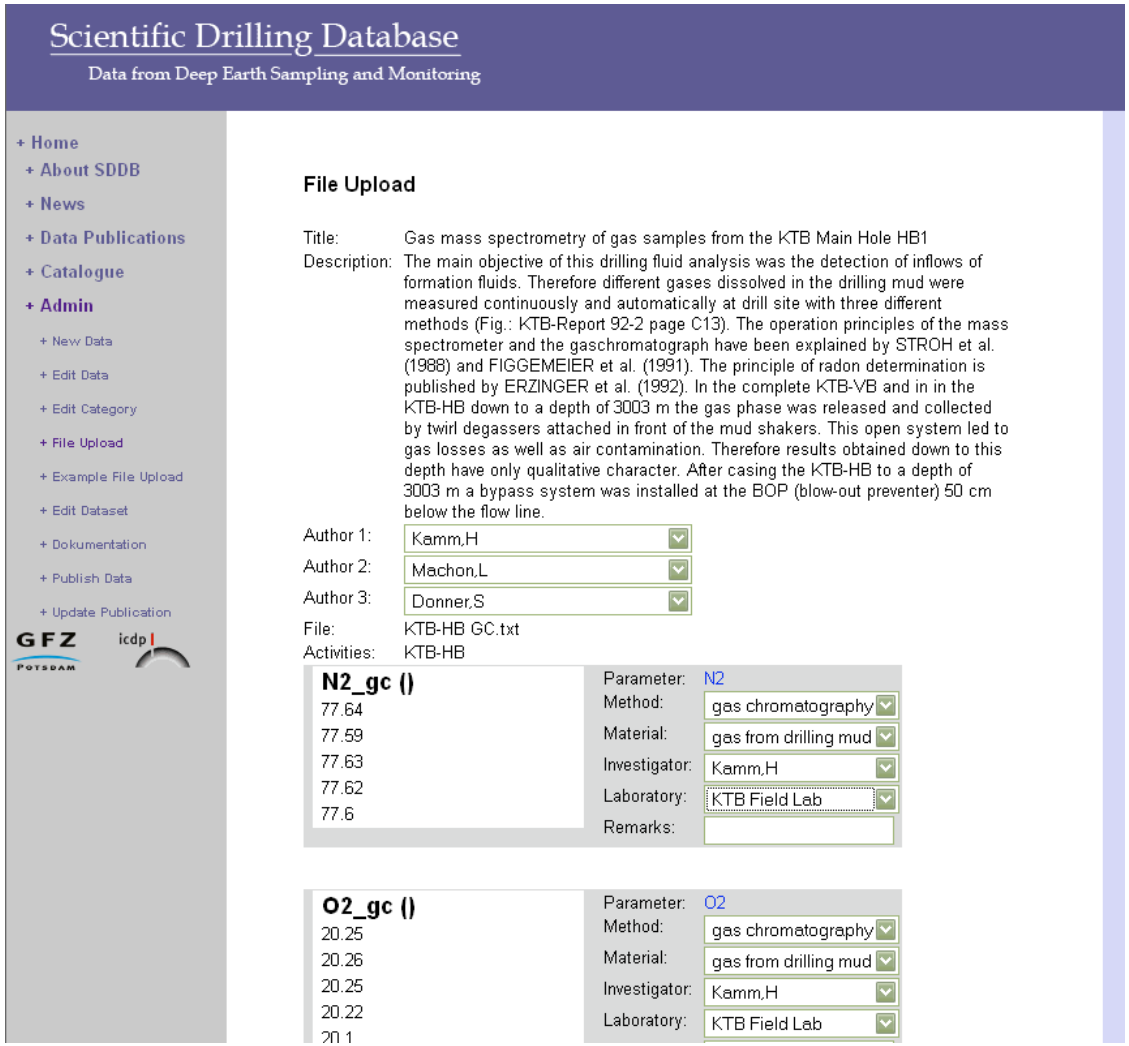



Link to publication

Citation of data

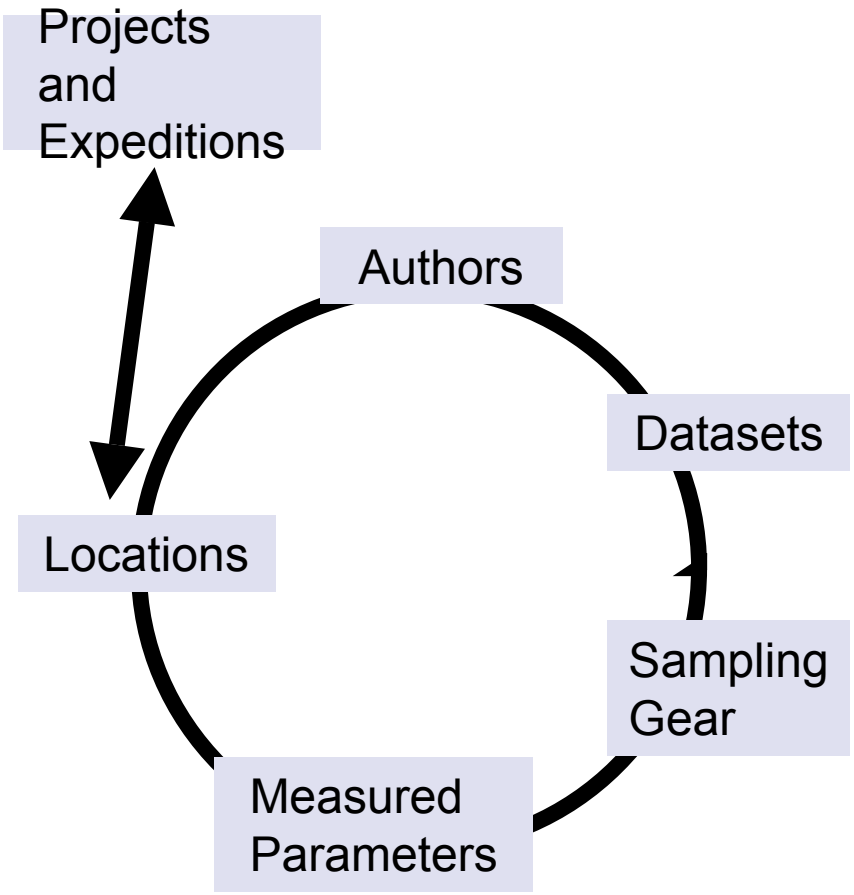
IGSN points to sample

# SDDB Metadata Management



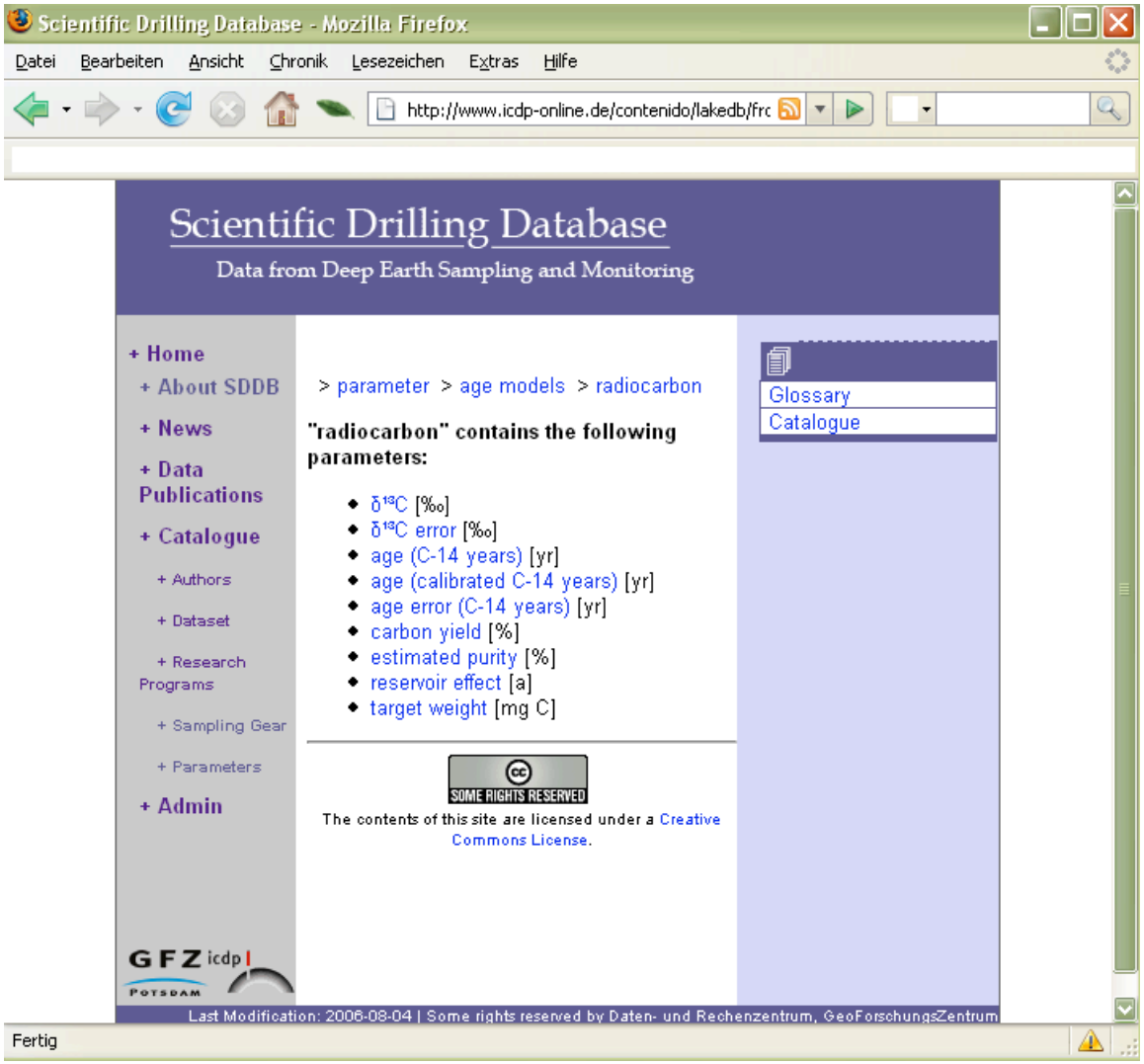

- Management of Metadata is kept simple by offering a data upload assistant.
- Data upload process is styled in analogy to eBay sales upload assistant.
- Most metadata are in the system already.

# Internal Semantics



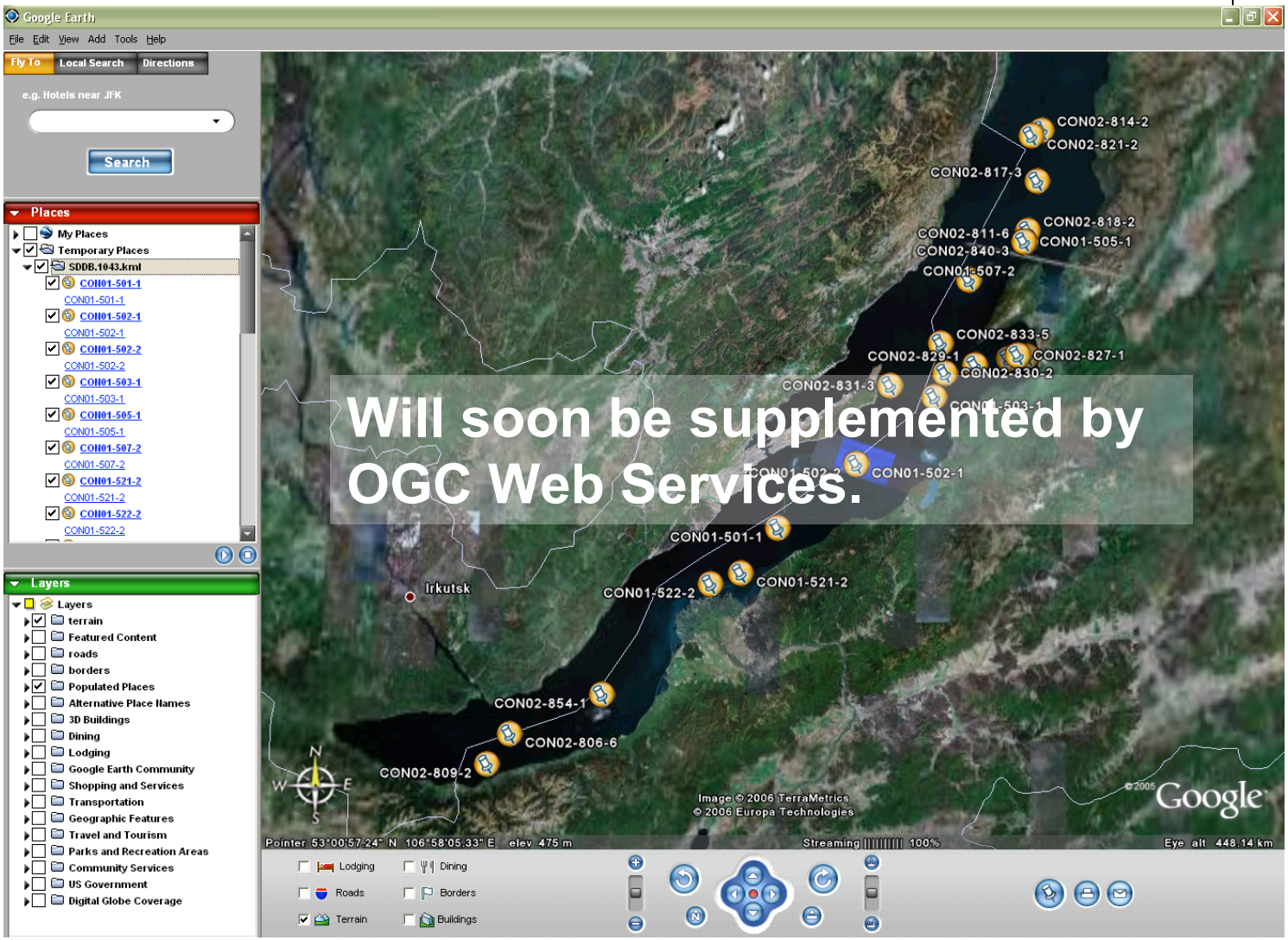

- SDDB metadata are held in a fully normalised relational database.
- SDDB metadata are fully browseable to allow iterative search.
- SDDB has (so far) no full-text search.

# Putting data into context

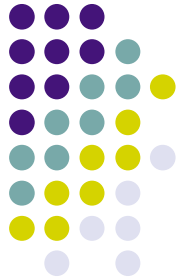



- Parameters are organised in a tree-structure.
- Homonyms are allowed – the parameter tree is ordered according to scientific context.

# The spatial dimension



# 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!