Publikation und Zitierbarkeit von Daten -

Ergebnisse und Anwendungen aus dem DFG-Projekt STD-DOI

Jens Klump *et al*.

GeoForschungsZentrum Potsdam

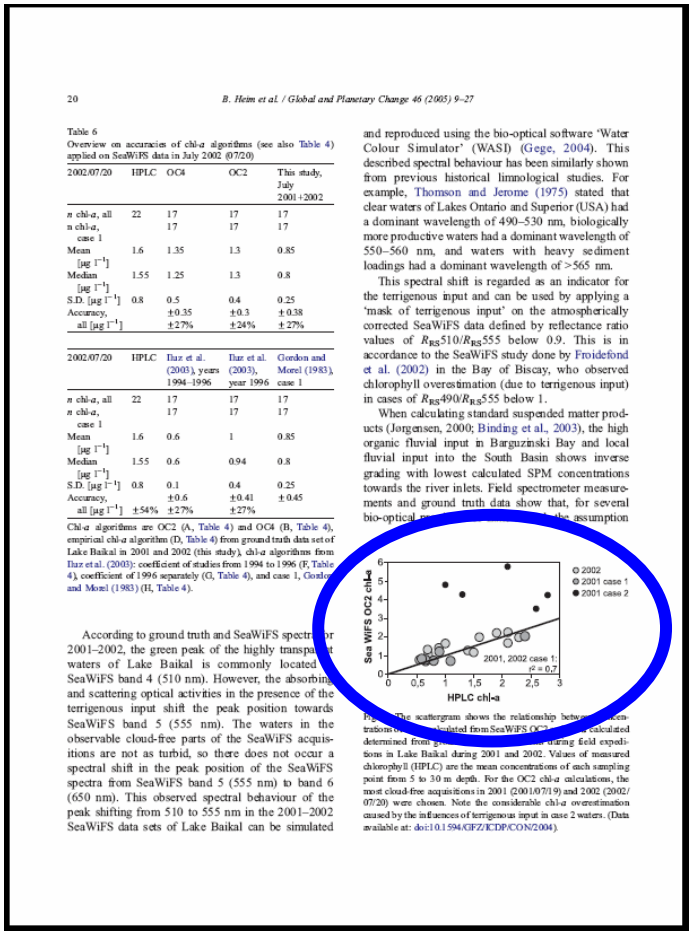
D-GRID Metadaten Workshop, 26.03.2007

Co-Autoren

- 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]

1. GFZ Potsdam (WDC-TERRA im Aufbau)
2. WDC-MARE, Univ. Bremen
3. TIB Hannover
4. WDC-MARE, AWI Bremerhaven
5. WDC-RSAT, DLR-DFD Oberpfaffenhofen
6. WDC-Climate, MPI-MET Hamburg

Datenveröffentlichungen heute



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

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

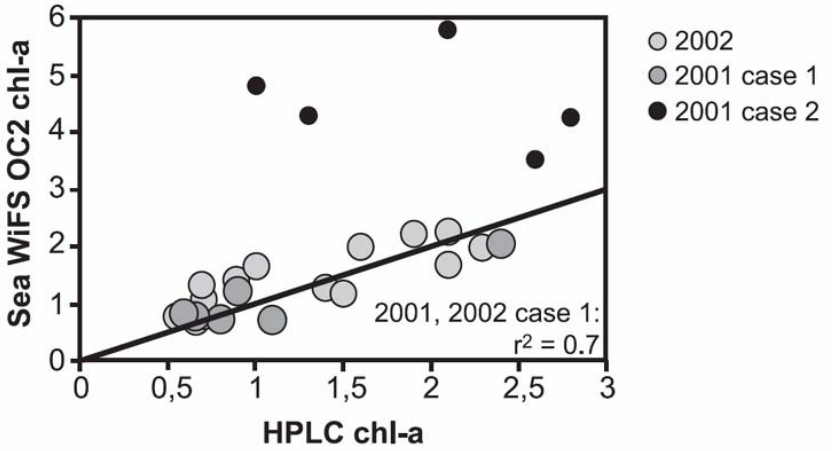


Fig. 2. The scattergram shows the relationship between concentrations of chl-*a* calculated from SeaWiFS OC2 and chl-*a* calculated determined from ground truth measurements during field expeditions in Lake Baikal during 2001 and 2002. Values of measured chlorophyll (HPLC) are the mean concentrations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

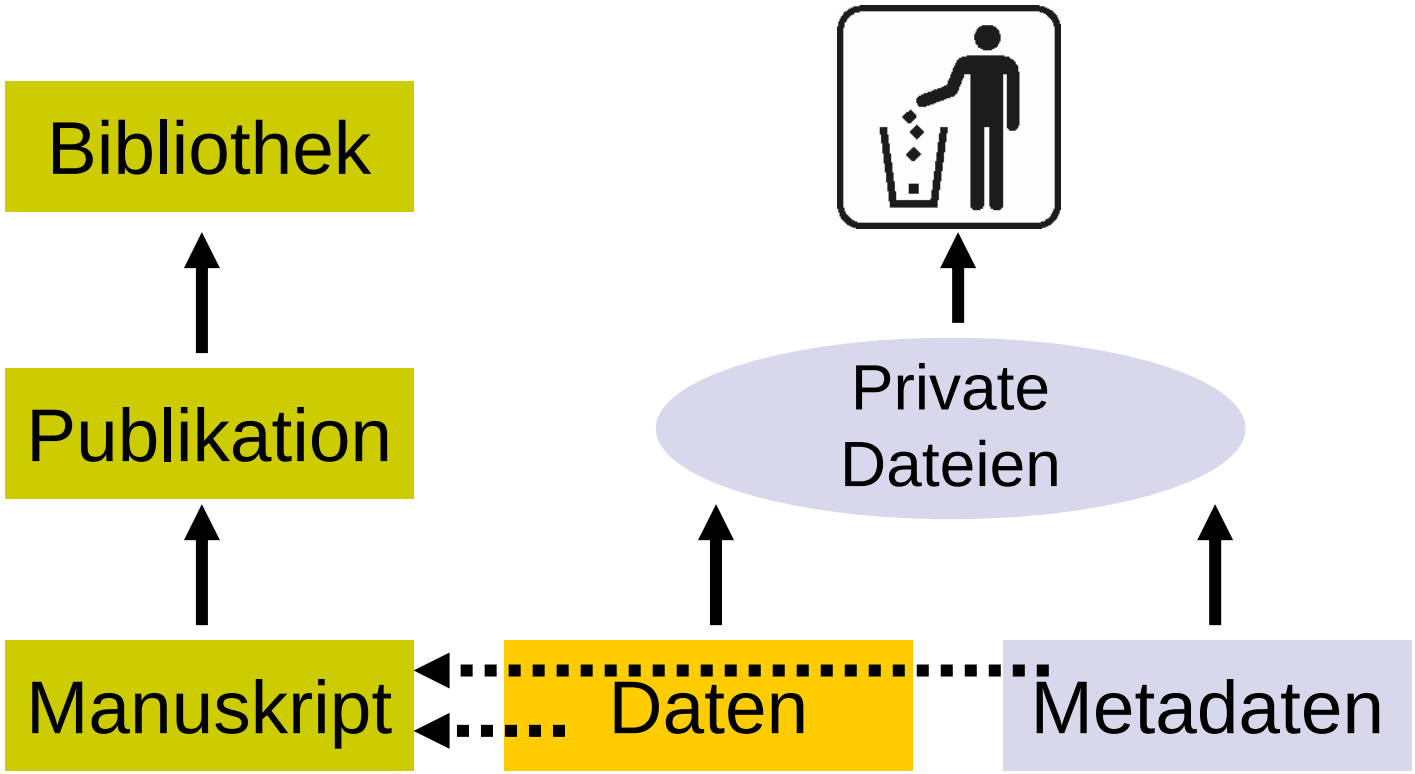
Umgang mit veröffentlichten Daten

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.

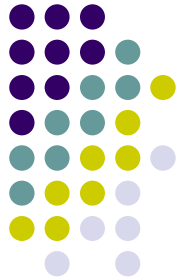
- Datenquelle wird nicht zitiert.
- Die Datenquelle muss aus dem Text abgeleitet werden.
- Keine Metadaten.
- Oft wird die Herkunft der Daten garnicht anerkannt.

Umgang mit Daten heute



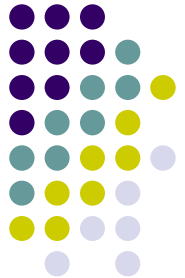
Nach Helly *et al*. (2003)

Warum Offener Zugang zu Daten?



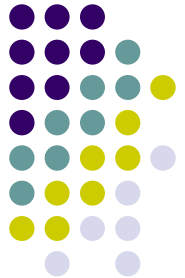
- Die Mengen der Forschungsdaten steigt exponentiell.
- Die meisten Daten werden unzureichend genutzt, weil sie nicht zugänglich sind.
- Die Folgen:
  - Unnötige Doppelarbeiten
  - Forschungsergebnisse sind kaum verifizierbar
- Aufrufe, Daten zugänglich zu machen, wurden zwar begrüßt, blieben aber folgenlos.

Warum Daten nicht zugänglich sind



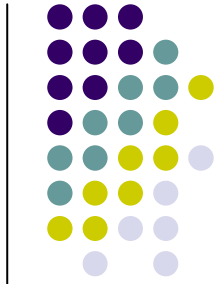
- Datenveröffentlichung scheitert heute noch an strukturellen Fehlern:
  - Zeitschriften räumen aus Kostengründen Daten keinen Platz in den Veröffentlichungen ein.
  - Autoren erhalten keine professionelle Anerkennung für Datenveröffentlichungen, weil Daten nicht zitiert werden (können).
  - Daten werden nicht zitiert, weil ihre Lokation (URL) meist nicht langlebig ist.

Besondere Situation am GFZ



- Am GFZ werden nicht nur abgeschlossene Datensätze aus Analysen und Modellen hergestellt, sondern auch Zeitreihen aus Mess-Systemen.
- Der Betrieb dieser Systeme ist sehr aufwändig, wird aber in der wissenschaftlichen Literatur kaum gewürdigt.
- Ziel: Anerkennung durch Veröffentlichung als wissenschaftliche Daten.

Notwendige Schritte



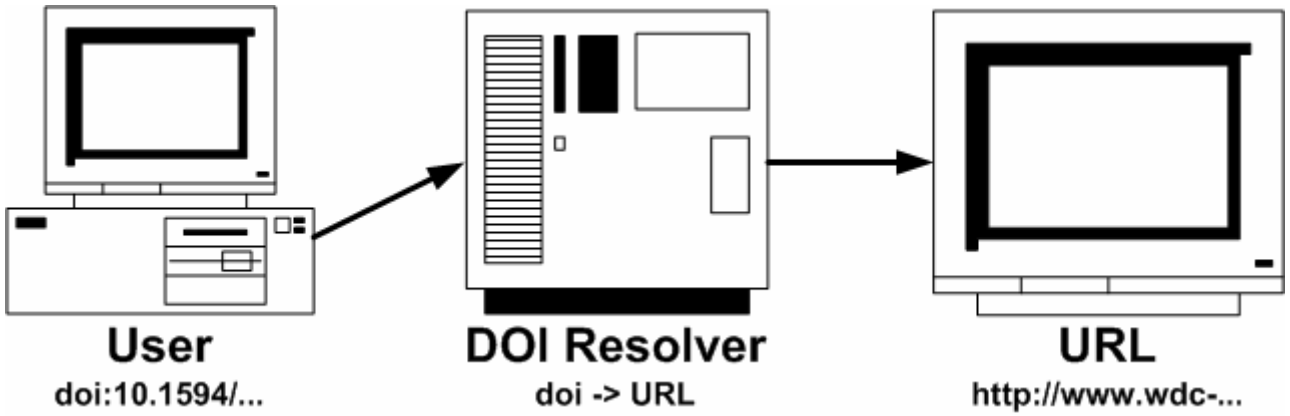
- Daten müssen zitierbar sein, um „wertvoll“ zu sein.
  - „Reputation“ ist die Währung der Wissenschaft.
- Autoren werden sich diese Arbeit nur machen, wenn es sich lohnt.
  - Datenpublikation ist arbeitsaufwändig.
- Daten müssen offen zugänglich sein.
  - Zugang gewährleisten über persistente Identifikatioren und Langzeitarchivierung.
- Urheberrecht muss gesichert sein.
  - Autor muss Kontrolle über die Nutzung der Daten haben.

DFG-Projekt STD-DOI

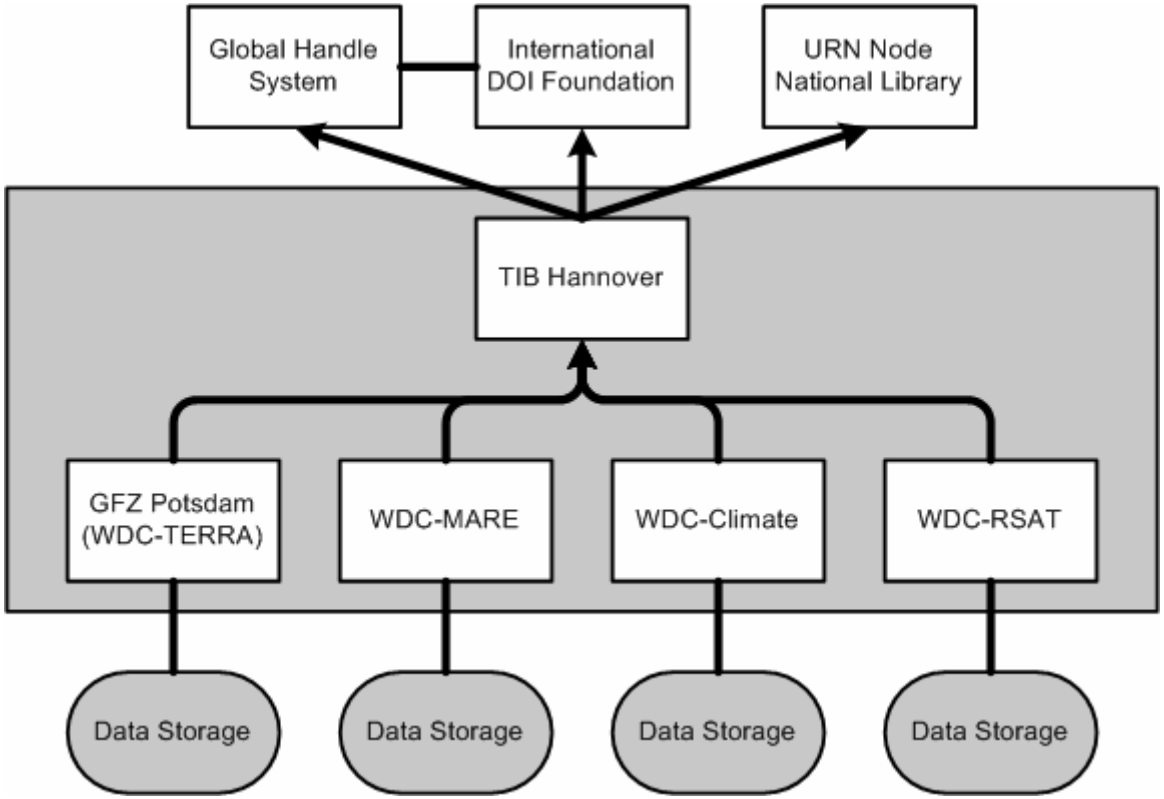
- Aufbau der Dienste zur Veröffentlichung von Daten.
- DOI Registrierungsagentur an der TIB Hannover.
- Bisher 6 DOI Registrierungsagenten.
- Projektpartner:
  - WDC-MARE (Bremen/Bremerhaven)
  - WDC Climate (Hamburg)
  - GFZ Potsdam
  - WDC-RSAT (Oberpfaffenhofen)
- Die Geokommission schlägt der DFG vor, die Veröffentlichung von Daten in die Förderrichtlinien aufzunehmen.

Was ist ein DOI?

- DOI = Digital Object Identifier. Ein persistenter, digitaler Identifikator eines Objekts.
- DOI = Name des Objekts, URL = Ort des Objekts.
- Der Ort eines Objekts kann sich ändern, der Name bleibt der gleiche, unabhängig vom Ort an dem sich das Objekt befindet.



STD-DOI Systemarchitektur



Beispiel für eine Datenveröffentlichung

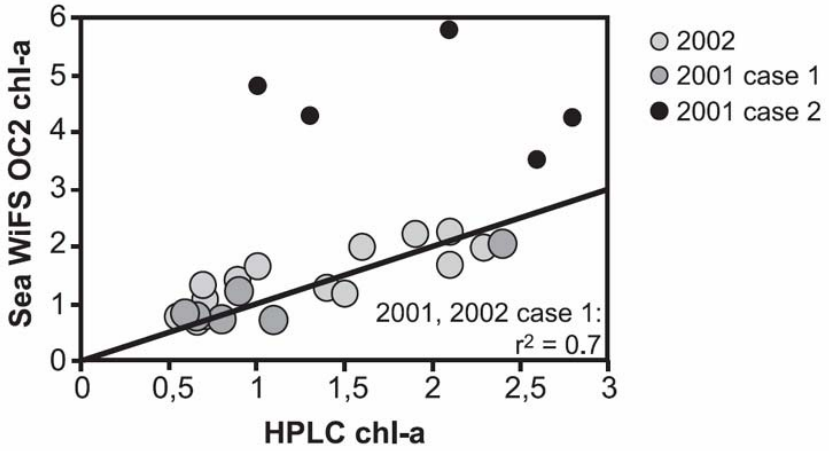
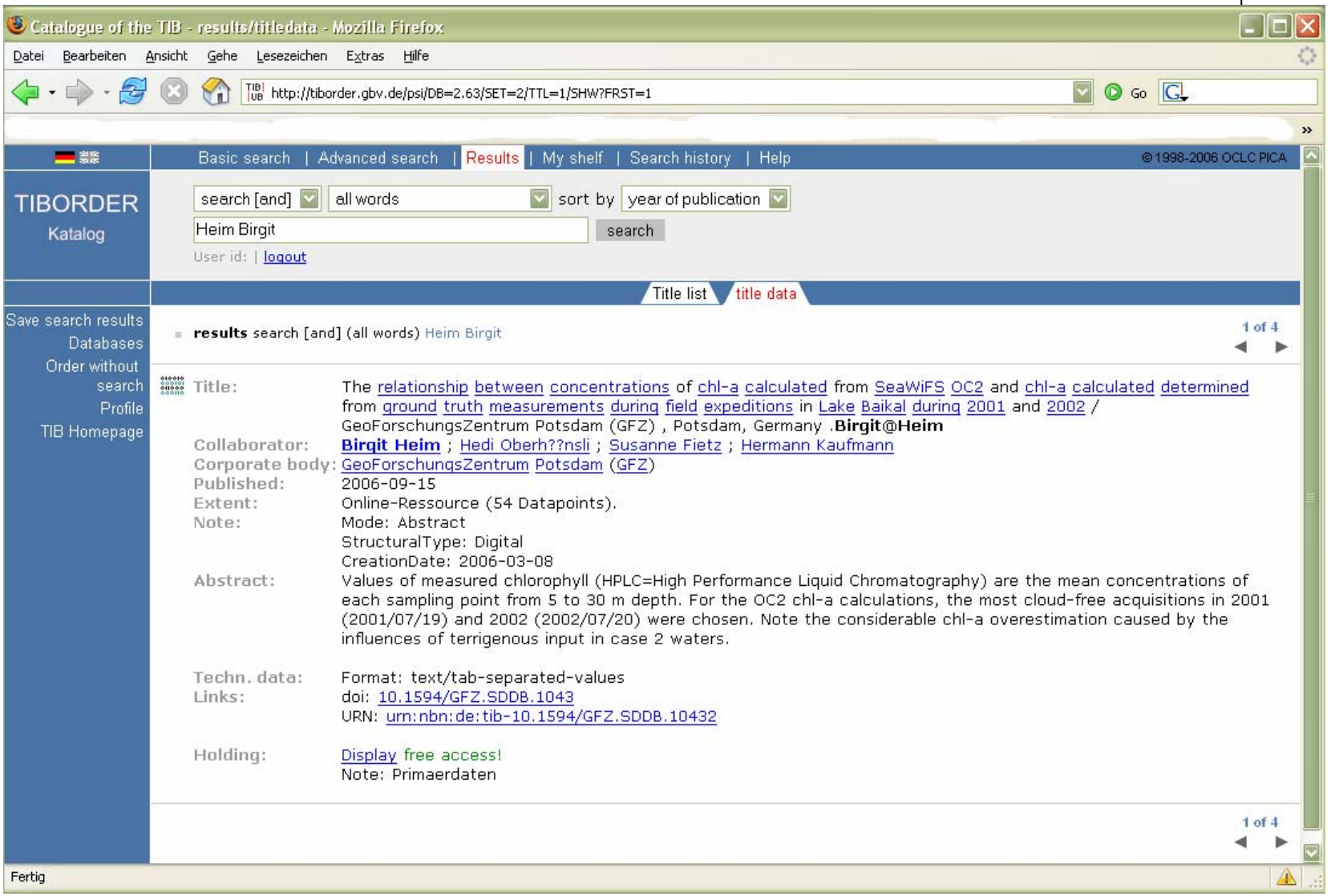


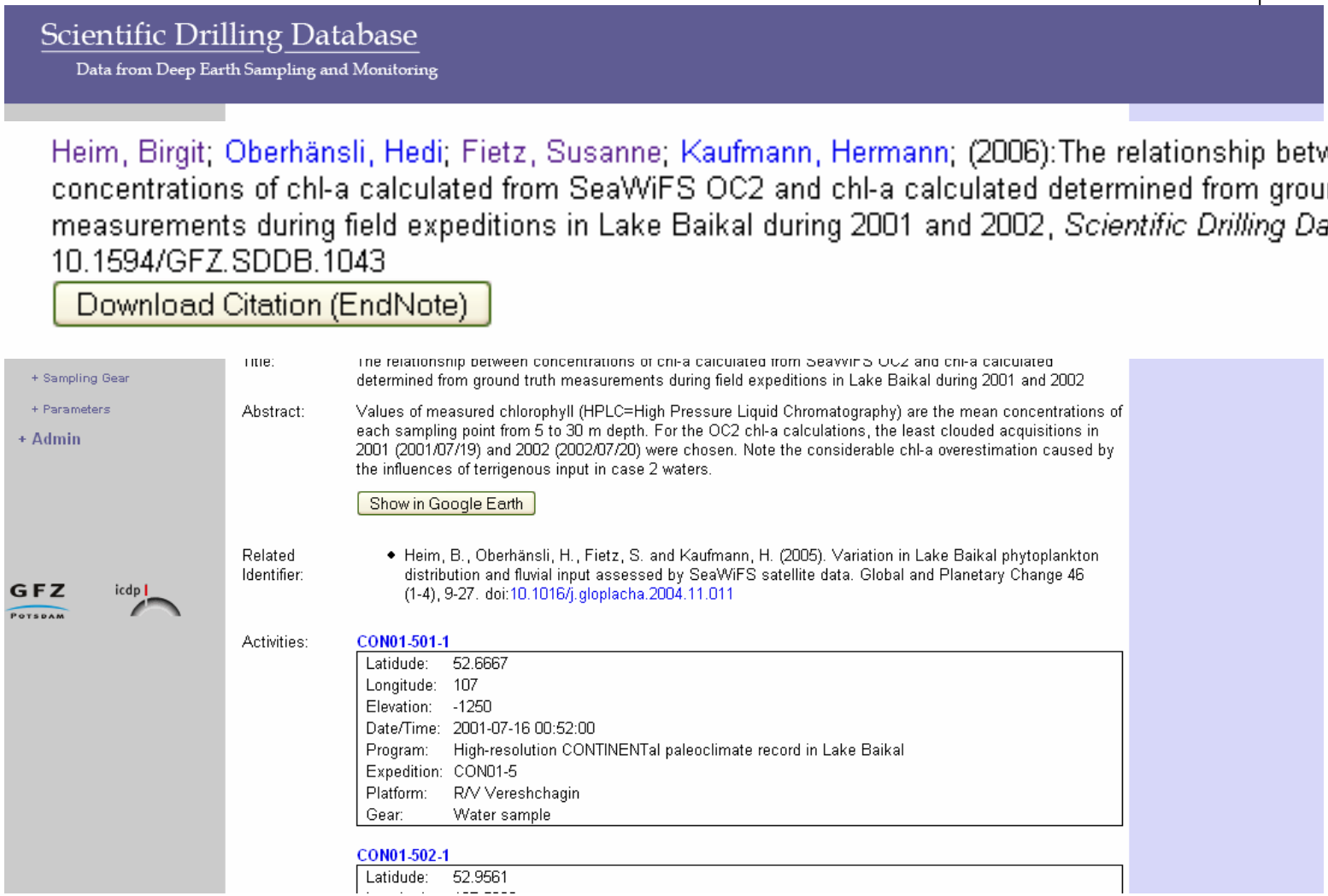
Fig. 2. The scattergram shows the relationship between concentrations of chl-*a* calculated from SeaWiFS OC2 and chl-*a* calculated determined from ground truth measurements during field expeditions in Lake Baikal during 2001 and 2002. Values of measured chlorophyll (HPLC) are the mean concentrations 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 Katalog.
- doi:10.1594/GFZ.SDDB.1043 verweist auf die ICDP Scientific Drilling Database.

TIBORDER / GBV Katalog



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:

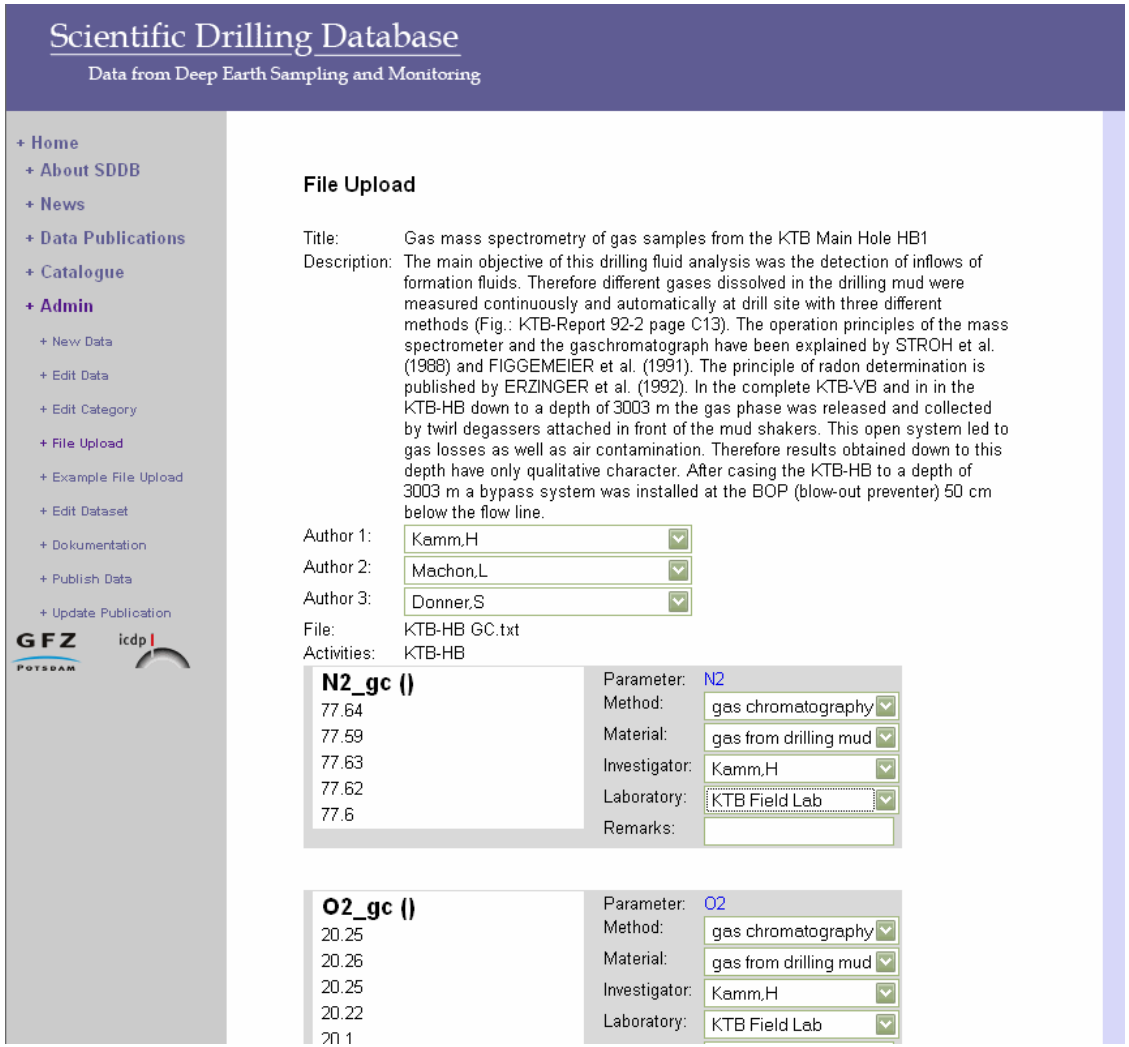
CON01-501-1

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

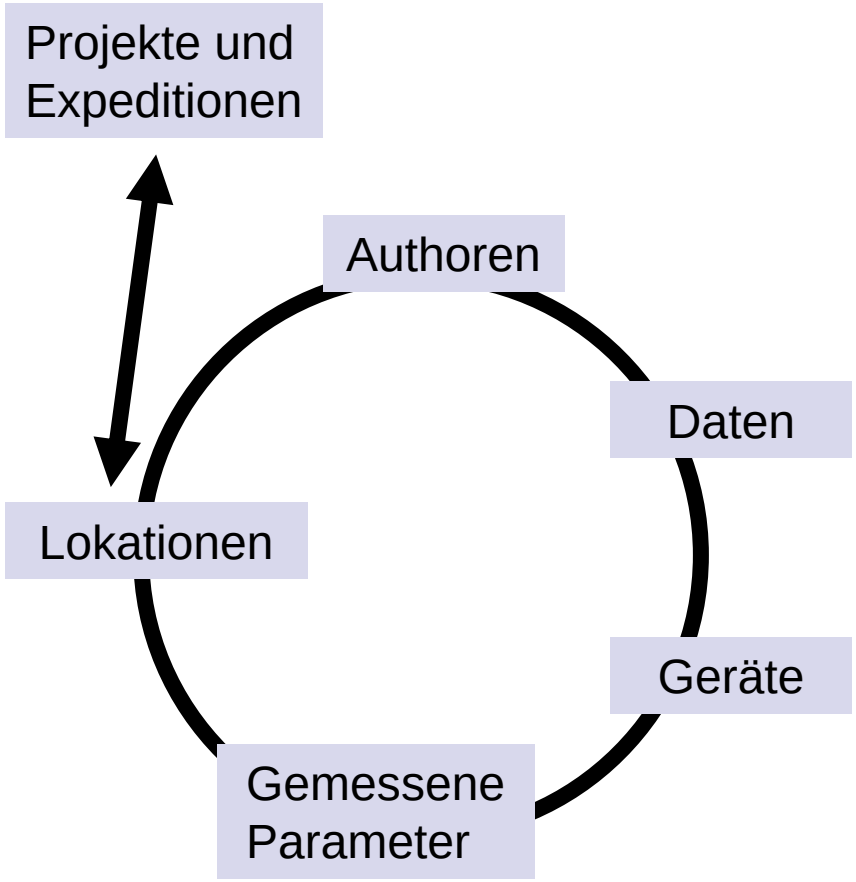
Latidude:	52.9561

SDDB Metadaten Management



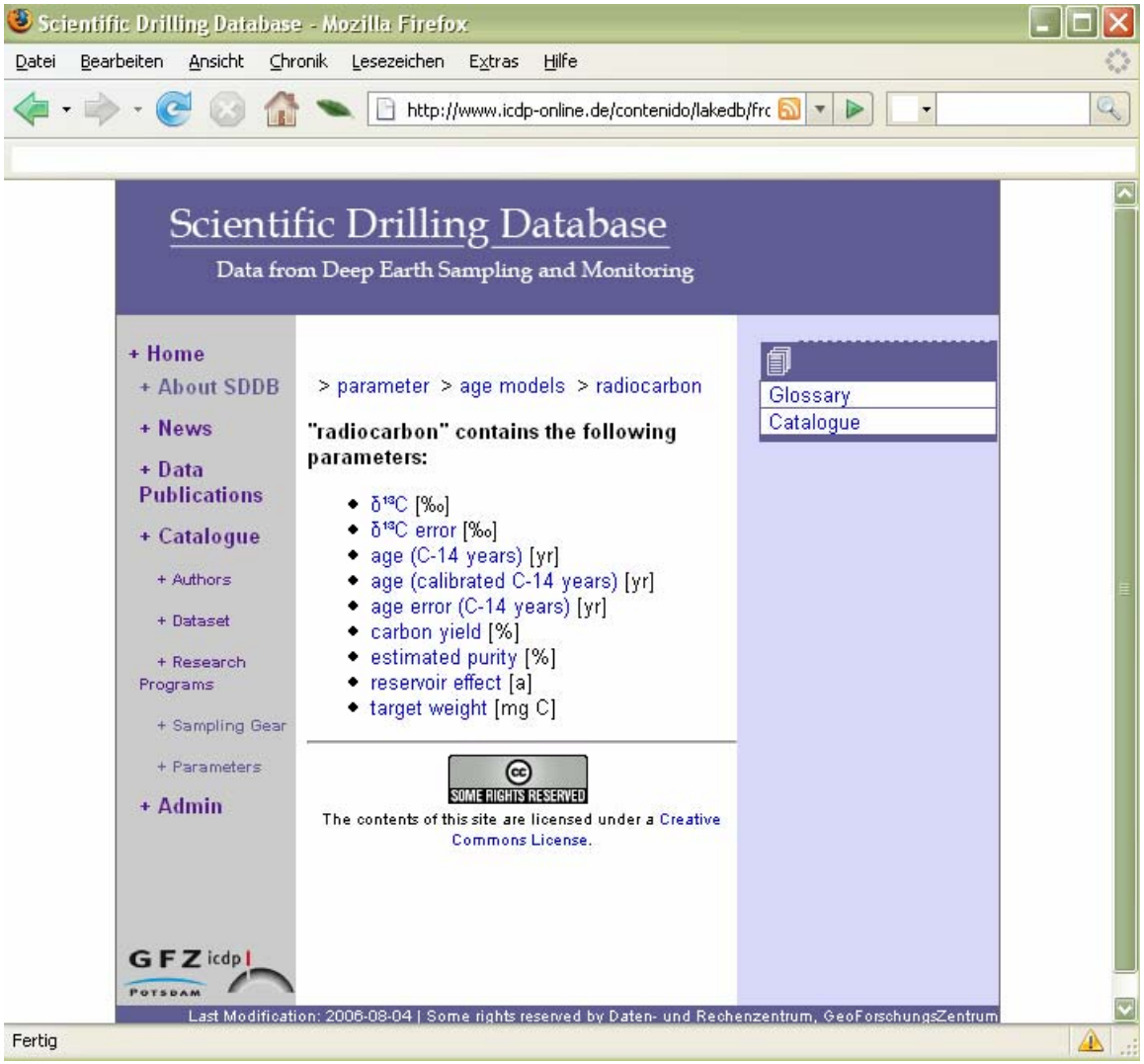
- Metadaten Management durch Autoren wird durch einen Daten-Upload Assistenten unterstützt.
- Der Assistent ist in Anlehnung an den eBay Upload-Assistenten gestaltet.

Interne Verknüpfungen



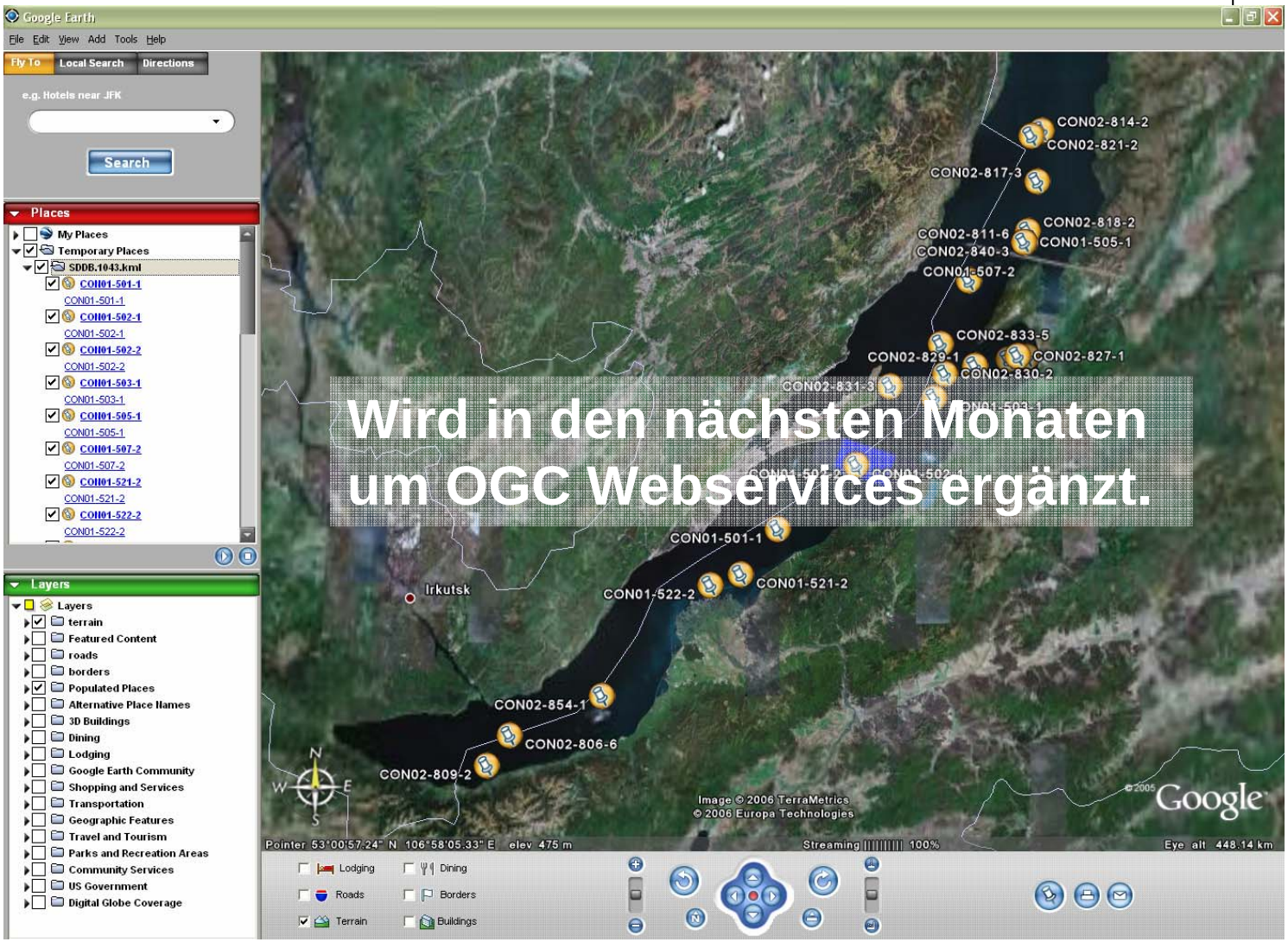
- SDDB Metadaten werden in einem normalisierten relationalen Datenbankschema abgelegt.
- SDDB Metadaten lassen sich durch Browsen iterativ durchsuchen.
- SDDB hat (bis jetzt) keine Volltextsuche.

Daten in den Kontext stellen



- Parameter sind in einer Baustruktur geordnet.
- Homonyme sind erlaubt – der Parameter-Baum ist nach wissenschaftlich em Kontext geordnet.

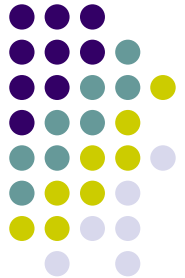
Die räumliche Dimension



STD-DOI Metadaten

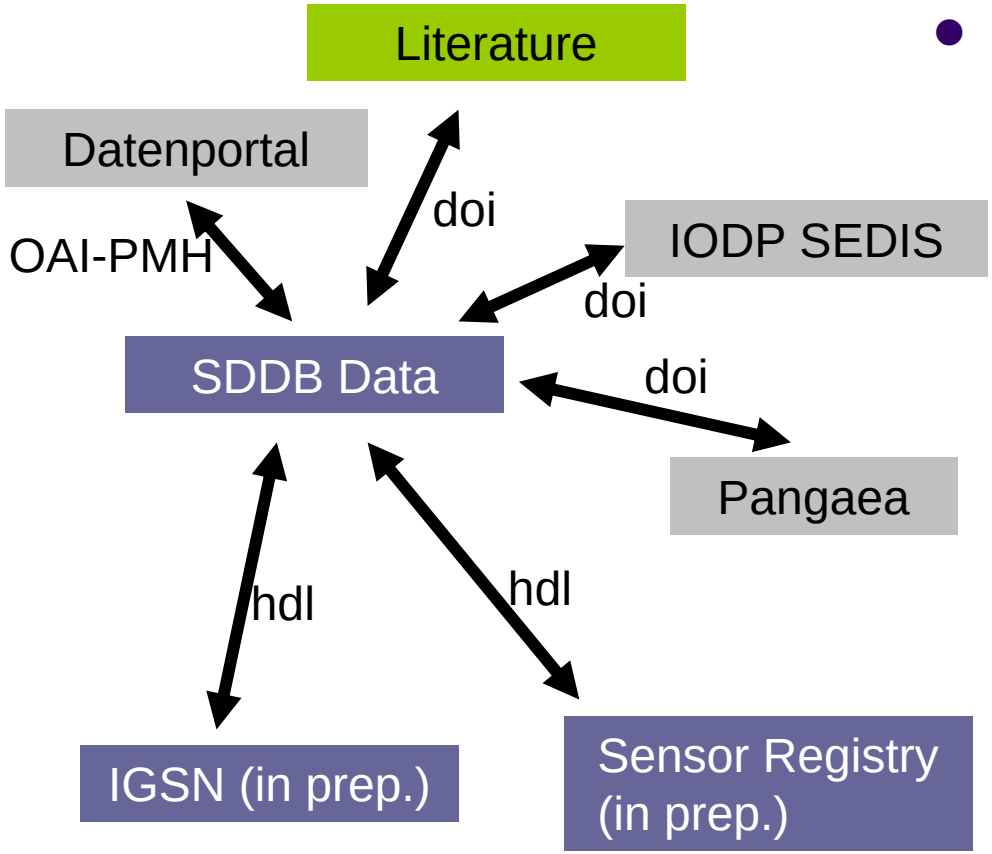
- Die STD-DOI Metadaten sind im wesentlichen Dublin Core-Elemente mit einigen Datenspezifischen Erweiterungen.
- Die Metadaten werden über Webservice (HTTP/SOAP) an die TIB-Hannover übertragen und dort in den Katalog eingefügt.
- Die STD-DOI Metadaten können Referenzen enthalten auf andere Objekte enthalten (semantische Vernetzung):
  - Element <RelatedIdentifier>
  - isCitedBy, isDuplicate, isNewVersionOf, …

Verweise auf andere Objekte - Semantische Vernetzung



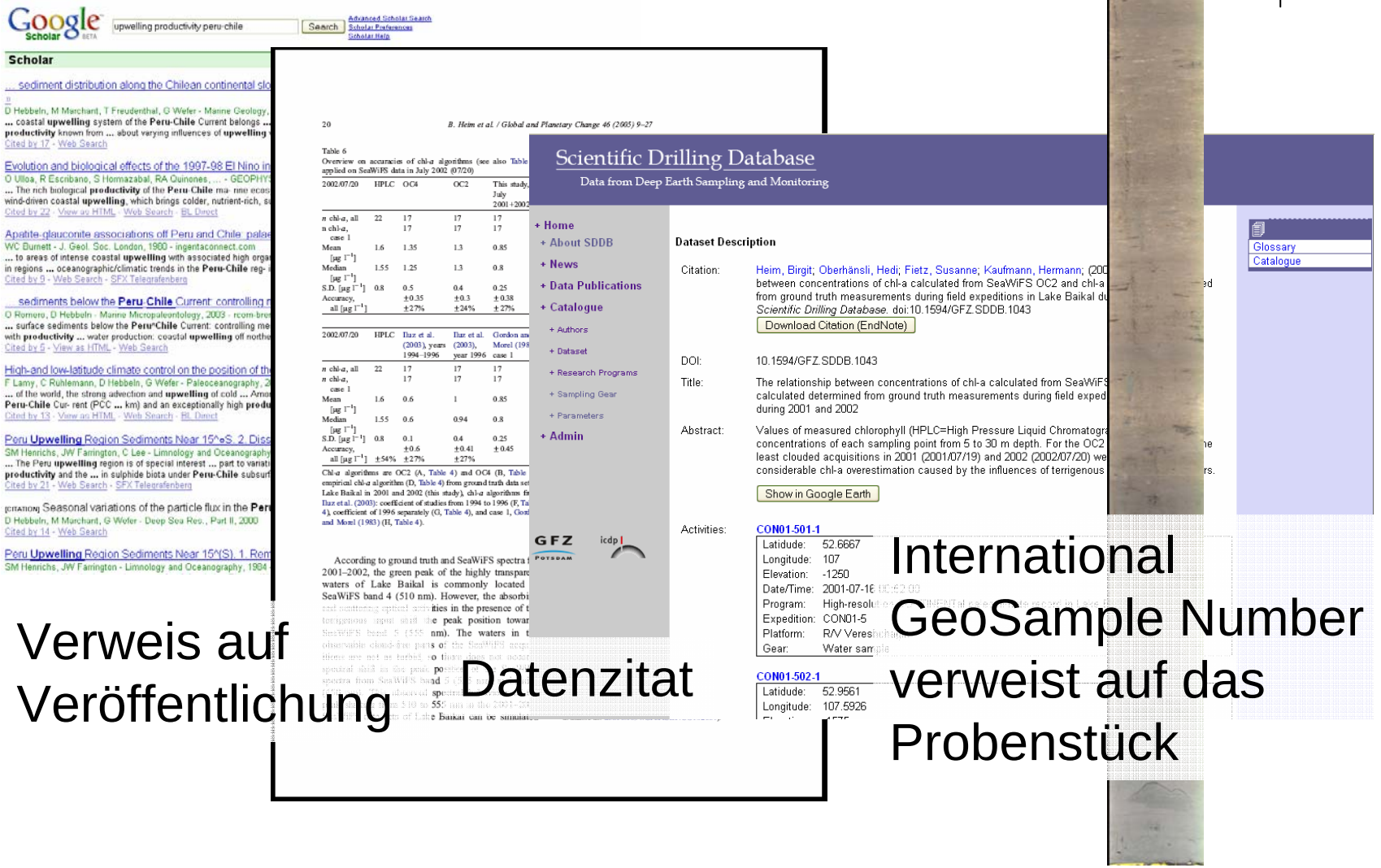
- Das Element <relatedIdentifier> kann genutzt werden, um auf andere Objekte zu verweisen:
  - Literatur, in der die Daten interpretiert werden
  - Proben, an denen die Daten gemessen wurden.
  - Andere Datensätze, die zum gleichen Set gehören, z.B. im gleichen Artikel interpretiert werden.
- Diese Verweise können von Maschinen ausgewertet werden (z.B. Datenportale) um Suchvorschläge anzubieten und damit die Entdeckung von Daten, Literatur und Probenstücken zu unterstützen. Auch weiterreichende semantische Mehrwertdienste sind denkbar.

Externe Verknüpfungen



- Das Element <relatedIdentifier> verweist auf externe Quellen:
  - Literatur, in der die Daten interpretiert werden.
  - Proben, an denen die Daten erhoben wurden.
  - Sensoren, mit denen die Daten gemessen wurden.
  - Andere Datensätze, die zur gleichen Sammlung gehören.

Wissen entdecken

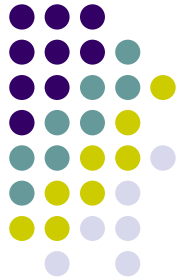


Verweis auf Veröffentlichung

Datenzitat

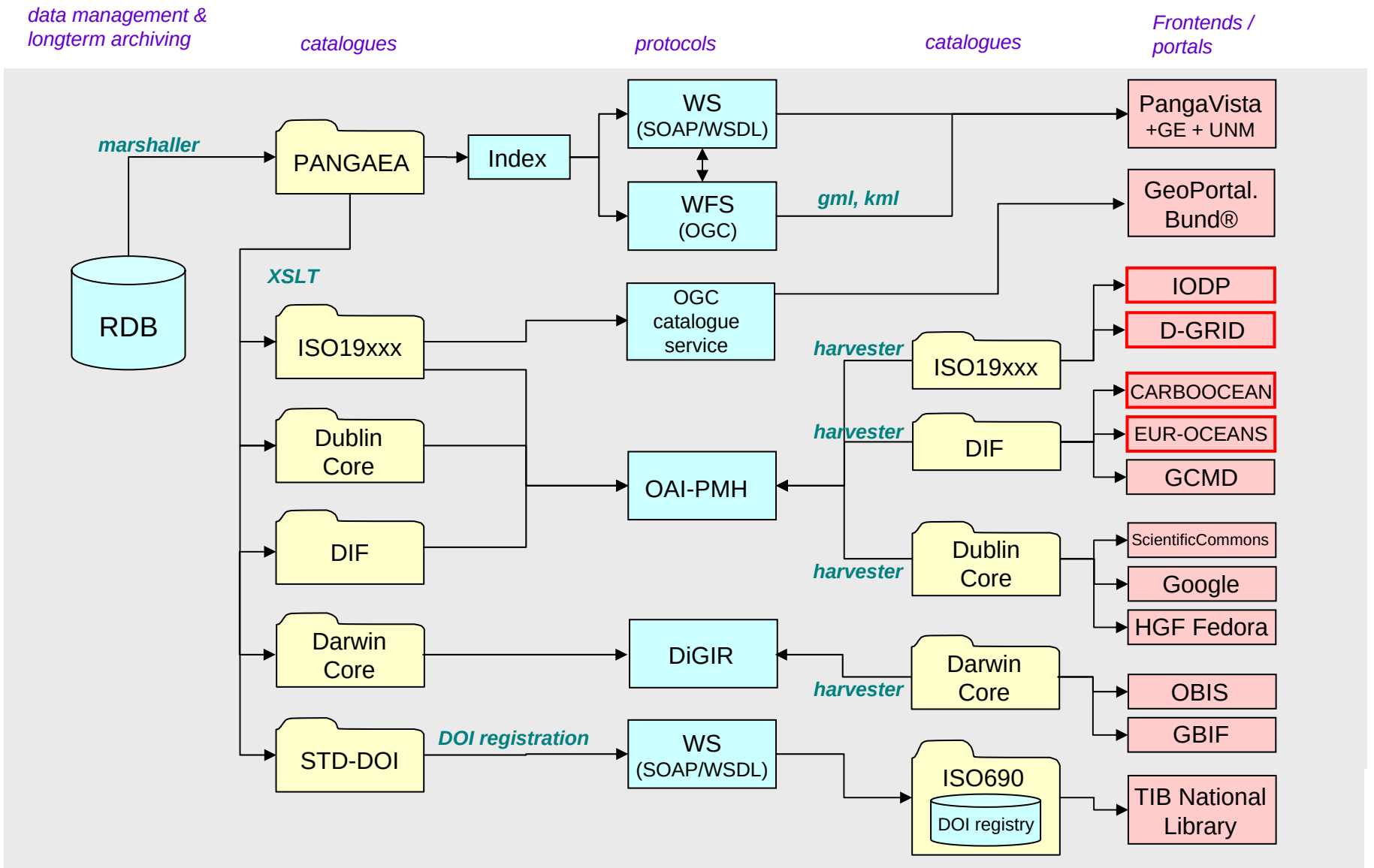
International GeoSample Number verweist auf das Probenstück

Data Syndication



- Datenveröffentlichungen können über Fachportale gesammelt und bekannt gemacht werden.
- Geeignete Protokolle sind OAI-PMH und RSS.
- Sowohl OAI-PMH, als auch RSS können mit Fachspezifischen Metadaten ausgestattet, oder wenigstens angereichert werden.
- Beispiele:
  - PANGAEA Datenportale (OAI-PMH)
  - Open Geospatial Consortium GeoRSS

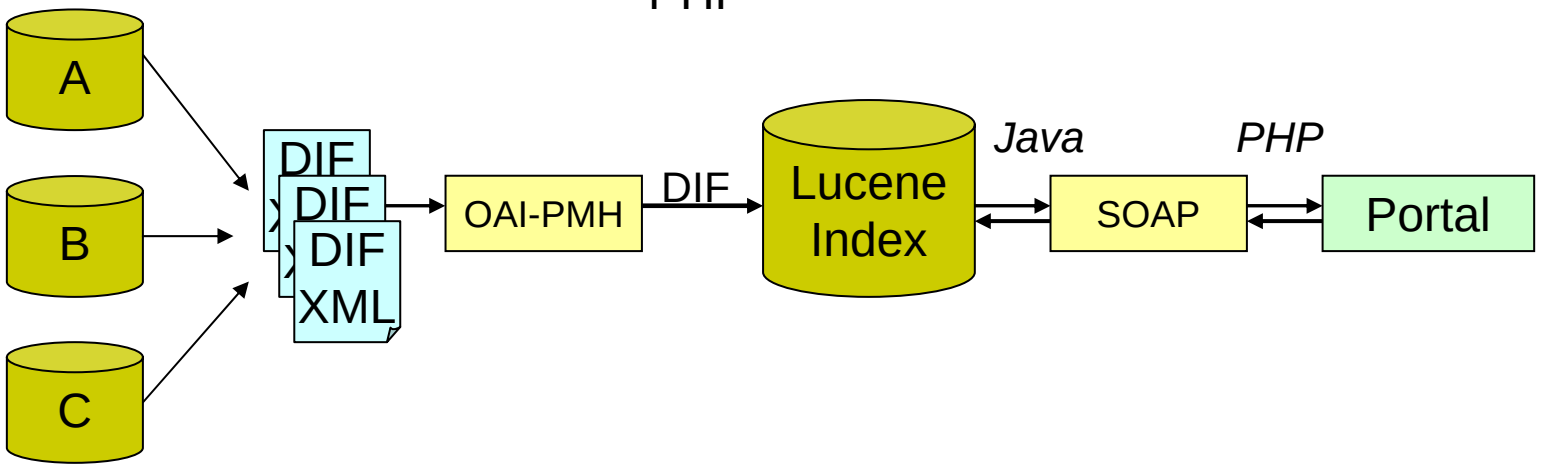
PANGAEA® – *standard interfaces for metadata (aus Huber et al. 2007)*



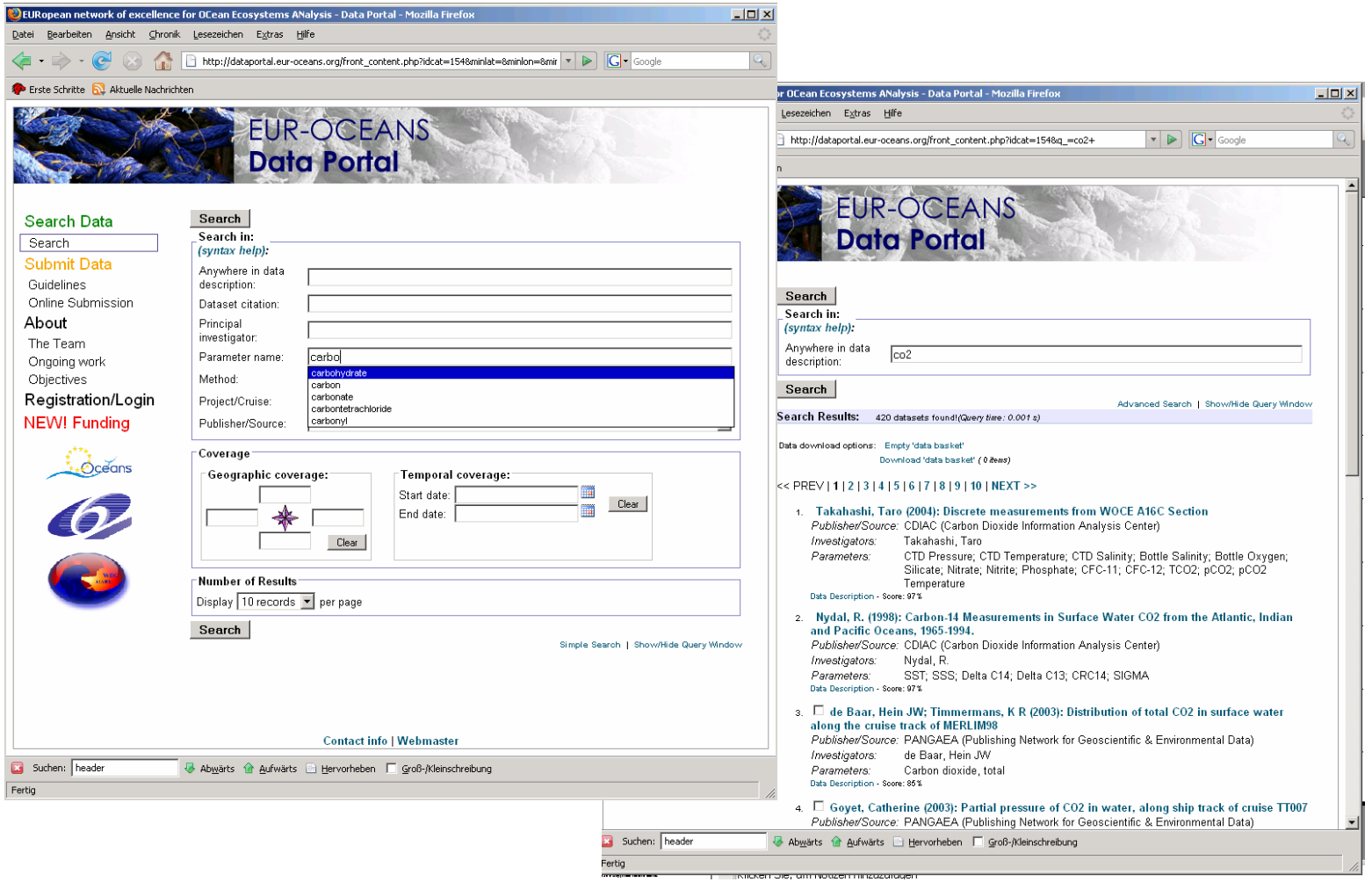
www.pangaea.de

Community portals-Technology

- OAI-PMH
- DIF / ISO
- Lucene
- SOAP
- PHP



Community portals-demo



Zusammenfassung

- Datenveröffentlichung verlangt folgende Voraussetzungen:
  - Persistente Identifikatoren
  - Langzeitarchive für Daten
  - Anreize für die Autoren zur Veröffentlichung der Daten
  - Geeignetes Lizenzmodell für wissenschaftliche Daten
- Datenveröffentlichungen können, und sollten auch, zitiert werden.
- Verweise auf andere Objekte können als semantische Vernetzung bei der Suche nach Daten, Literatur und Materialien helfen.
- Die Herausforderung ist nun, daraus einen Teil der wissenschaftlichen Kultur zu machen – die Technologie ist bereits vorhanden.

Weiterführende Informationen

- DFG-Projekt “Publikation und Zitierbarkeit von wissenschaftlichen Primärdaten”

  http://www.std-doi.de

- ICDP Scientific Drilling Database

  http://www.scientificdrilling.org

- Klump, J., et al. (2006): Data publication in the Open Access Initiative. Data Science Journal, **5**, 79-83. doi:10.2481/dsj.5.79

Vielen Dank!