Geophysical Research Abstracts,
Vol. 11, EGU2009-11628-2, 2009
EGU General Assembly 2009

The PANGAEA® Data Warehouse

U. Schindler (1), M. Diepenbroek (1), H. Grobe (2), and R. Sieger (2)

(1) MARUM - University of Bremen, PANGAEA, Bremen, Germany (uschindler@pangaea.de, mdiepenbroek@pangaea.de), (2) Foundation Alfred Wegener Institute for Polar and Marine Research (AWI), PANGAEA, Bremerhaven, Germany (hannes.grobe@awi.de, rainer.sieger@awi.de)

PANGAEA® - Publishing Network for Geoscientific & Environmental Data (www.pangaea.de) currently provides published data entities for download that can be cited like publications using authors, year, title and a Digital Object Identifier (DOI). A lot of scientists, especially modellers, need compilations of various data sets for analyzing. On the other hand, data producers want to be cited for their work, which is nearly impossible with huge compilations containing thousands of distinct data sets, especially when data compilation is done manually outside of the PANGAEA data library.

We will present our recently introduced, AJAX-based web interface based on data warehouse technologies that can used for highly efficient retrievals and compilations of time slices or surface data matrixes on any measurement parameters out of the whole data continuum. The user is able to first select the original datasets using the current PANGAEA full text search engine (based on ISO 19115 metadata). After that he can select the configuration of measurement parameters and methods for the data compilation using an innovative scoring algorithm based on the original search query. The data matrix can be downloaded along with an ISO 19115 compatible metadata description about the compilation, referencing all original datasets. Additionally, each data point can be traced back to the original, citable data set.

The PANGAEA® Data Warehouse

Uwe Schindler[1], Michael Diepenbroek[1], Hannes Grobe[2], Rainer Sieger[2]

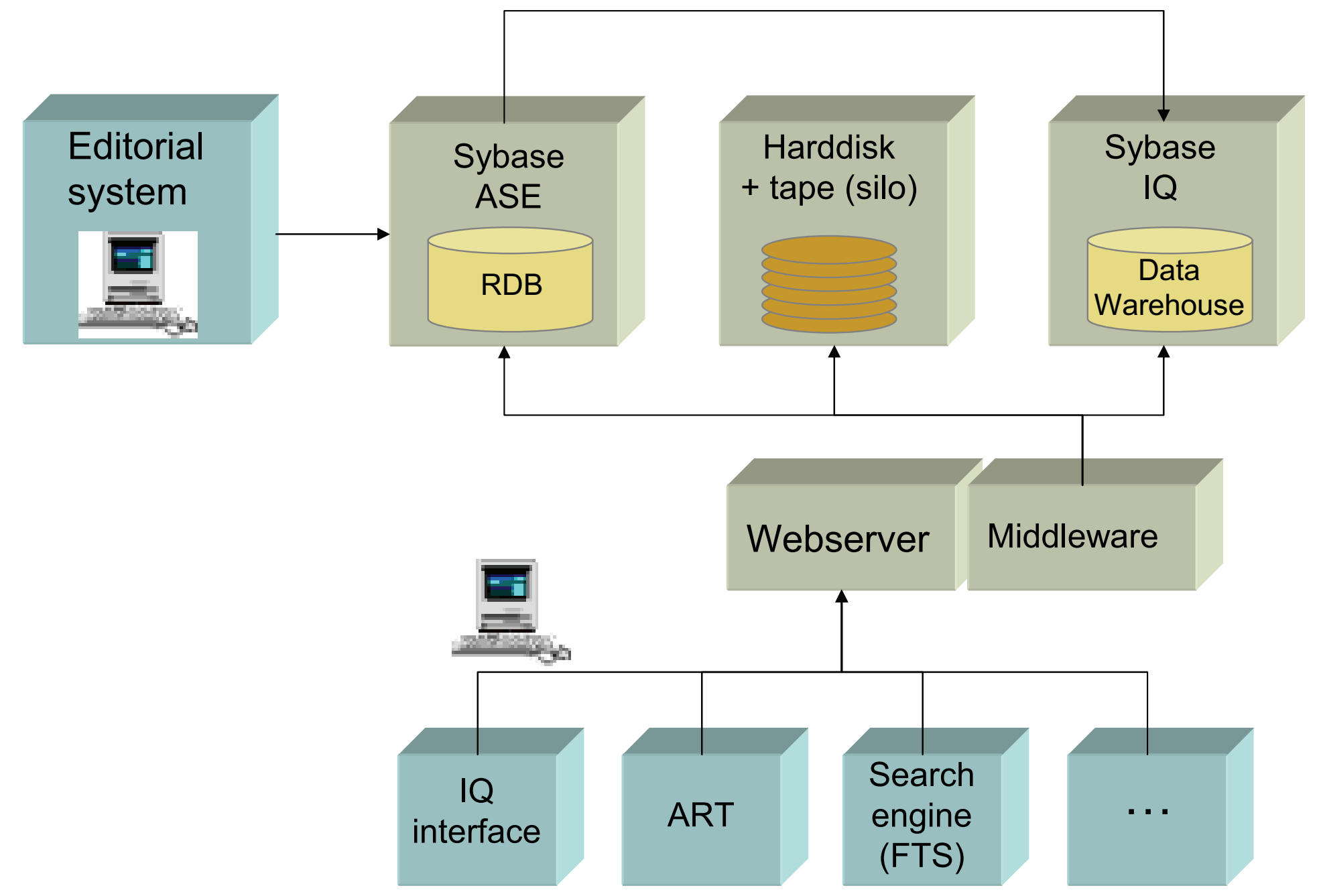
Abstract

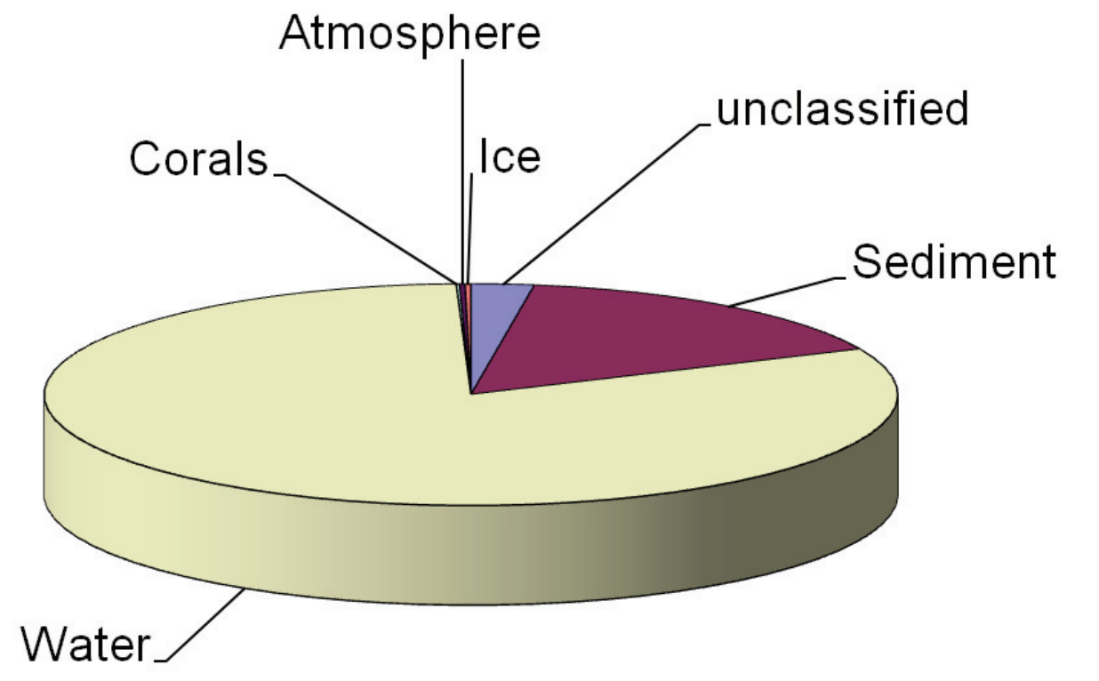
PANGAEA® - Publishing Network for Geoscientific & Environmental Data (www.pangaea.de) currently provides published data entities for download that can be cited like publications using authors, year, title and a Digital Object Identifier (DOI). A lot of scientists, especially modellers, need compilations of various data sets for analyzing. On the other hand, data producers want to be cited for their work, which is nearly impossible with huge compilations containing thousands of distinct data sets, especially when data compilation is done manually outside of the PANGAEA data library.

We present our recently introduced, AJAX-based web interface based on data warehouse technologies that can used for highly efficient retrievals and compilations of time slices or surface data matrixes on any measurement parameters out of the whole data continuum. The user is able to first select the original datasets using the current PANGAEA full text search engine (based on ISO 19115 metadata). After that he can select the configuration of measurement parameters and methods for the data compilation using an innovative scoring algorithm based on the original search query. The data matrix can be downloaded along with an ISO 19115 compatible metadata description about the compilation, referencing all original datasets. Additionally, each data point can be traced back to the original, citable data set.

About PANGAEA®

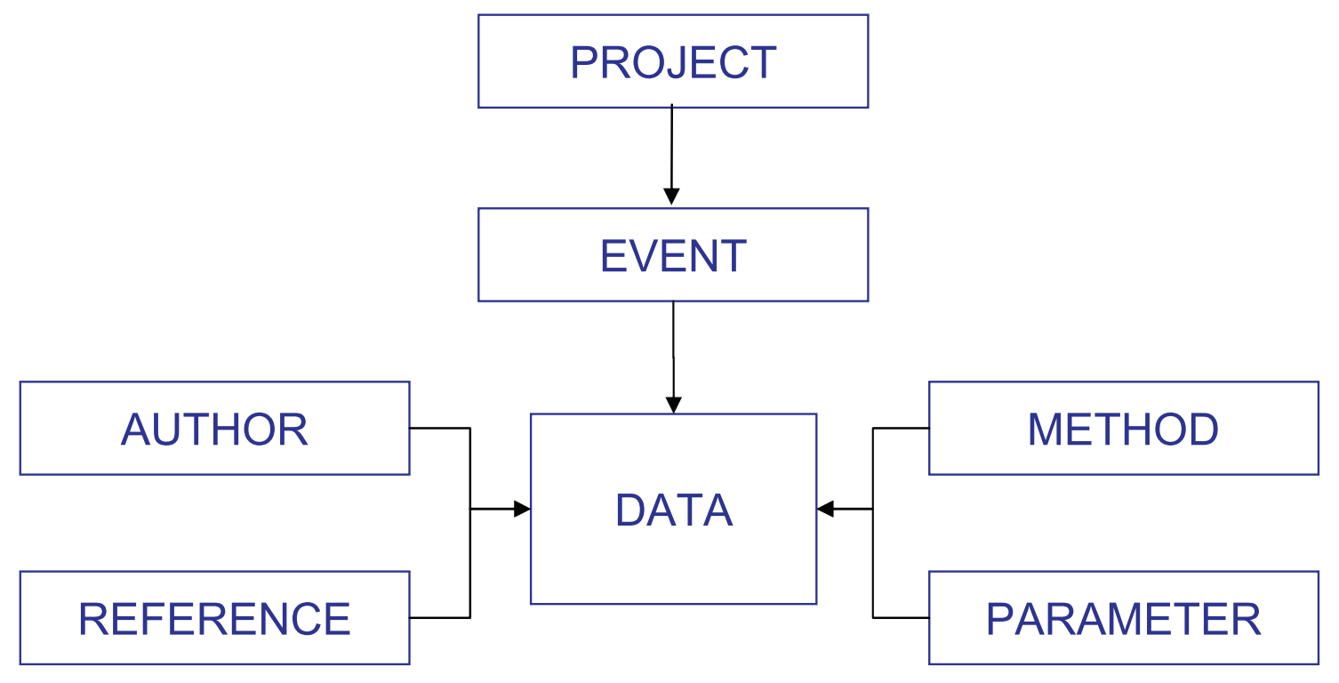
PANGAEA® (www.pangaea.de) is operated by the Alfred Wegener Institute for Polar and Marine Research (AWI), Bremerhaven, member of the Helmholtz Association of National Research Centres, funded by the Federal Ministry of Education and Research and the Center for Marine Environmental Sciences (marum) at the University of Bremen with support of the Department of Geosciences. Around 20 people are associated with WDC-MARE. The budget amounts approximately 1,2 Mio Euro per year for personnel, hard-, and software. Third party funds are about 70% of the total budget.The basic technical structure corresponds to a three tiered client/server architecture with a number of clients and middleware components controlling the information flow and quality. On the server side a relational database management system (RDBMS) is used for information storage. To ensure fast data access the data are mirrored in a data warehouse (SYBASE IQ) which is also used as interface to the german GRID community. All interfaces to the information system's metadata, datasets and data warehouse are based on web services including a simple map supported search engine.





Database status (2/2009):

578 931 data sets
4 669 156 961 data items



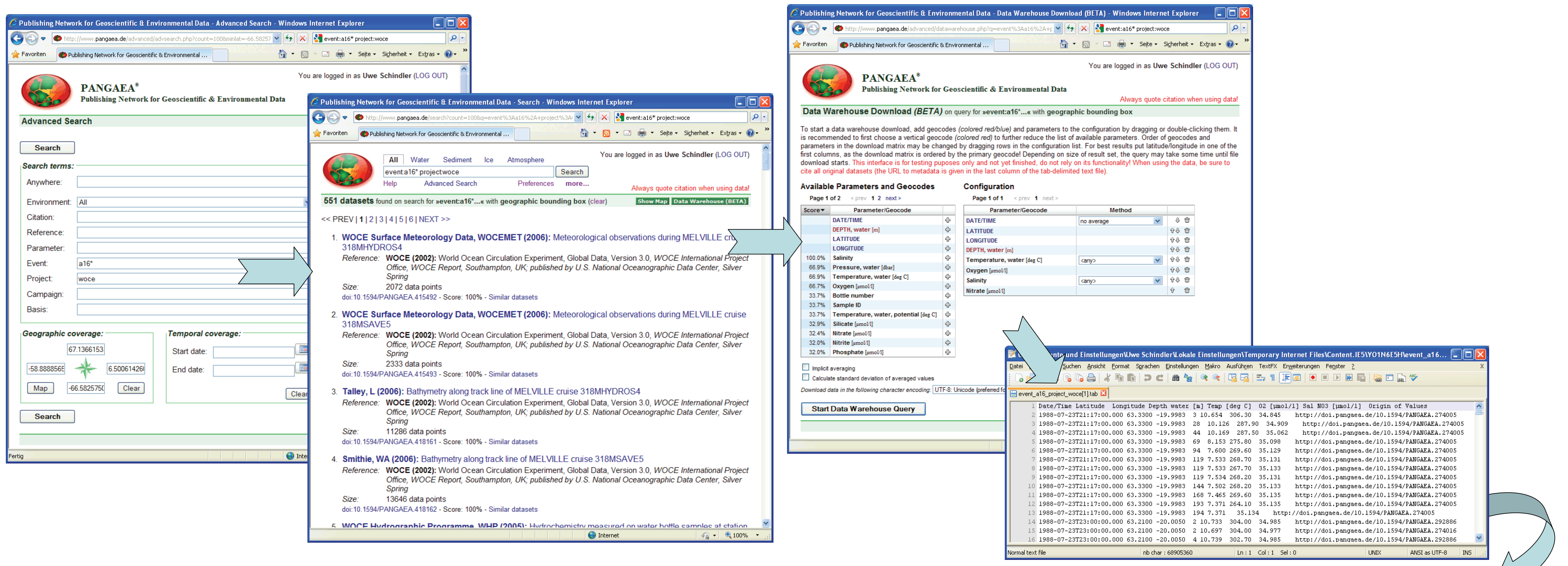
Data for nearly 50 000 different measurement types, e.g.:
Sediment and Ice Profiles, Seismic Profiles, Mineral distribution, Geological Maps, Atmospheric Profiles, Radiation Data, Distributed Samples, Oceanographic Profiles, Time Series, Plankton and Fish

With its comprehensive user interfaces and the built in editorial system for import, and maintainance of information **PANGAEA®** is a highly efficient system for scientific data management and data publication.
The challenge of managing the heterogenous and dynamic data of environmental and geosciences was met in PANGAEA® through a flexible data model which reflects the information processing steps in the earth science fields and can handle any related analytical data. Metadata are consistently stored in a relational database and can be served as ISO19115, DIF, FGDC, or DC compliant catalogues.
The data description of any data set includes the principle investigators (PI) name and email for contact, parameters, used methods as well as spatial and temporal coverage.
Except a few data from ongoing projects which are under moratorium all data are freely available and can be used by referencing the related publication or the data set citation.

Sample Workflow: WOCE A16 Profile

To start a data warehouse query, you must initially do a conventional full-text search engine query on the PANGAEA dataset metadata. Just enter some specific keywords for the data you are searching. In the example, we searched for data of the WOCE A16 profile. Optionally you can also add other constraints like a specific bounding box from where the data should come from or a date range for time series.
You get a list of citable datasets entities. In the past, you had to download all these single datasets as separate files and combine them using separate tools.
With the new data warehouse interface of PANGAEA, you can download the data as a single file containing only the relevant data points. From the result list, you can click on "Data Warehouse (BETA)" (currently only available for logged-in users) to open the new interface.

The result set of datasets is analyzed for measurement parameters, methods and used geocodes (corrdinates like latitude, longitude, date, time, elevation) and all of these are displayed in a scored list. An AJAX based GUI helps you to choose the interesting parameters and combine them in a "configuration" sheet for the resulting data file. Parameters and geocodes can be combined as columns in the resulting data matrix, methods can be choosen together with aggregation functions.
With one further click you get the tab-delimited result matrix as a single file.
As a lot of different datasets were involved when creating the result, each data point is referenced back to the citable dataset, the measurement value was taken from, helping to correctly reference the original data.
The result file can be visualized with common tools like Ocean Data View (ODV, http://odv.awi.de).



Publishing Network for Geoscientific & Environmental Data - Advanced Search - Windows Internet Explorer

http://www.pangaea.de/advanced/advsearch.php?count=10&minlat=-66.58257 — event:a16* project:woce

You are logged in as Uwe Schindler (LOG OUT)

PANGAEA® Publishing Network for Geoscientific & Environmental Data

Advanced Search

Search terms:
Anywhere:
Environment: All
Citation:
Reference:
Parameter:
Event: a16*
Project: woce
Campaign:
Basis:

Geographic coverage: 67.1366153 / -58.8888565 / 6.50061426 / -66.5825750

Temporal coverage: Start date: End date:

Publishing Network for Geoscientific & Environmental Data - Search - Windows Internet Explorer

http://www.pangaea.de/search?count=10&q=event%3Aa16%2A+project%3A — event:a16* project:woce

All | Water | Sediment | Ice | Atmosphere — event:a16* project:woce — Search

You are logged in as Uwe Schindler (LOG OUT)

Always quote citation when using data!

551 datasets found on search for »event:a16*...« with geographic bounding box (clear) — Show Map — Data Warehouse (BETA)

<< PREV | 1 | 2 | 3 | 4 | 5 | 6 | NEXT >>

1. **WOCE Surface Meteorology Data, WOCEMET (2006):** Meteorological observations during MELVILLE cruise 318MHYDROS4
   Reference: **WOCE (2002):** World Ocean Circulation Experiment, Global Data, Version 3.0, *WOCE International Project Office, WOCE Report, Southampton, UK; published by U.S. National Oceanographic Data Center, Silver Spring*
   Size: 2072 data points
   doi:10.1594/PANGAEA.415492 - Score: 100% - Similar datasets
2. **WOCE Surface Meteorology Data, WOCEMET (2006):** Meteorological observations during MELVILLE cruise 318MSAVE5
   Reference: **WOCE (2002):** World Ocean Circulation Experiment, Global Data, Version 3.0, *WOCE International Project Office, WOCE Report, Southampton, UK; published by U.S. National Oceanographic Data Center, Silver Spring*
   Size: 2333 data points
   doi:10.1594/PANGAEA.415493 - Score: 100% - Similar datasets
3. **Talley, L (2006):** Bathymetry along track line of MELVILLE cruise 318MHYDROS4
   Reference: **WOCE (2002):** World Ocean Circulation Experiment, Global Data, Version 3.0, *WOCE International Project Office, WOCE Report, Southampton, UK; published by U.S. National Oceanographic Data Center, Silver Spring*
   Size: 11286 data points
   doi:10.1594/PANGAEA.418161 - Score: 100% - Similar datasets
4. **Smithie, WA (2006):** Bathymetry along track line of MELVILLE cruise 318MSAVE5
   Reference: **WOCE (2002):** World Ocean Circulation Experiment, Global Data, Version 3.0, *WOCE International Project Office, WOCE Report, Southampton, UK; published by U.S. National Oceanographic Data Center, Silver Spring*
   Size: 13646 data points
   doi:10.1594/PANGAEA.418162 - Score: 100% - Similar datasets
5. **WOCE Hydrographic Programme, WHP (2005):** Hydrochemistry measured on water bottle samples at station

Publishing Network for Geoscientific & Environmental Data - Data Warehouse Download (BETA) - Windows Internet Explorer

http://www.pangaea.de/advanced/datawarehouse.php?q=event%3Aa16%2A+ — event:a16* project:woce

You are logged in as Uwe Schindler (LOG OUT)

Always quote citation when using data!

Data Warehouse Download *(BETA)* on query for »event:a16*...« with geographic bounding box

To start a data warehouse download, add geocodes *(colored red/blue)* and parameters to the configuration by dragging or double-clicking them. It is recommended to first choose a vertical geocode *(colored red)* to further reduce the list of available parameters. Order of geocodes and parameters in the download matrix may be changed by dragging rows in the configuration list. For best results put latitude/longitude in one of the first columns, as the download matrix is ordered by the primary geocode! Depending on size of result set, the query may take some time until file download starts. This interface is for testing pupposes only and not yet finished, do not rely on its functionality! When using the data, be sure to cite all original datasets (the URL to metadata is given in the last column of the tab-delimited text file).

Available Parameters and Geocodes — Page 1 of 2 < prev 1 2 next >

Score	Parameter/Geocode
	DATE/TIME
	DEPTH, water [m]
	LATITUDE
	LONGITUDE
100.0%	Salinity
66.9%	Pressure, water [dbar]
66.9%	Temperature, water [deg C]
66.7%	Oxygen [µmol/l]
33.7%	Bottle number
33.7%	Sample ID
33.7%	Temperature, water, potential [deg C]
32.9%	Silicate [µmol/l]
32.4%	Nitrate [µmol/l]
32.0%	Nitrite [µmol/l]
32.0%	Phosphate [µmol/l]

Configuration — Page 1 of 1 < prev 1 next >

Parameter/Geocode	Method
DATE/TIME	no average
LATITUDE	
LONGITUDE	
DEPTH, water [m]	
Temperature, water [deg C]	<any>
Oxygen [µmol/l]	
Salinity	<any>
Nitrate [µmol/l]	

□ Implicit averaging
□ Calculate standard deviation of averaged values

Download data in the following character encoding: UTF-8: Unicode (preferred fo

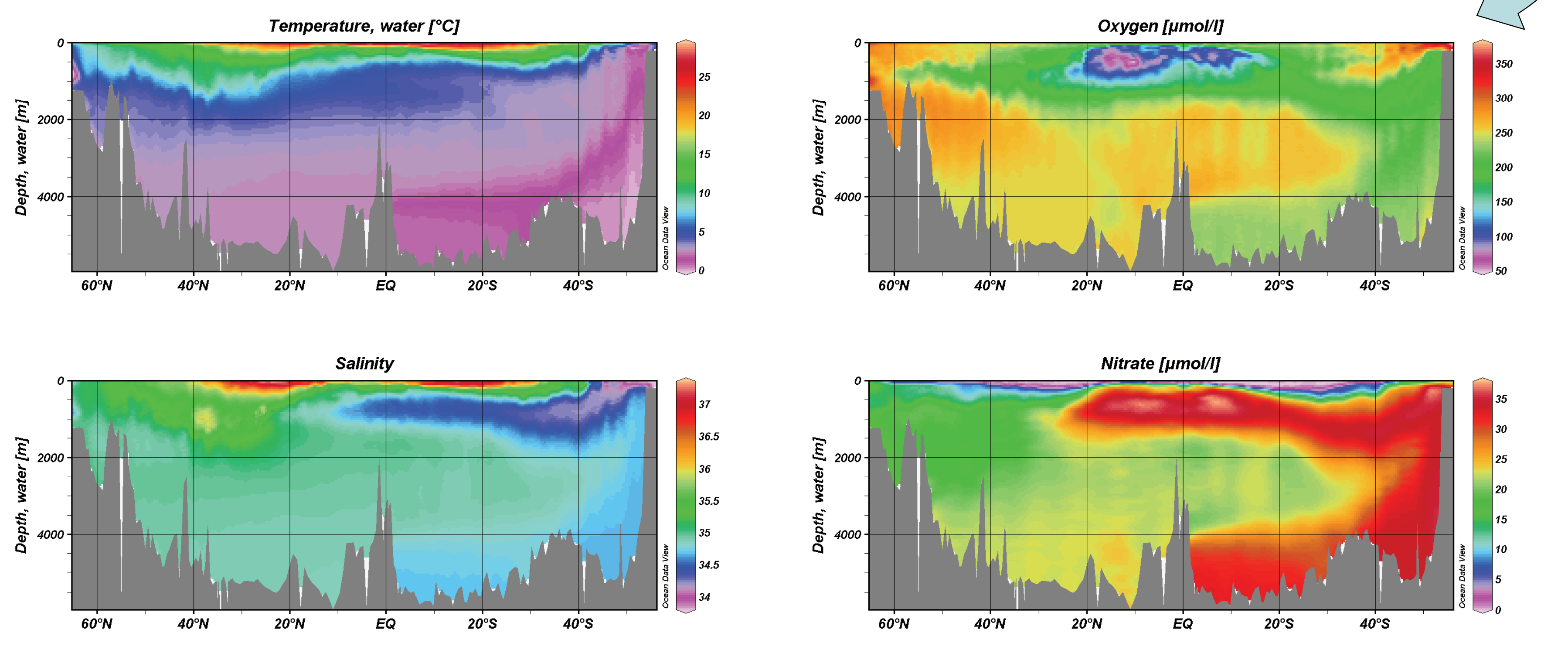
Start Data Warehouse Query

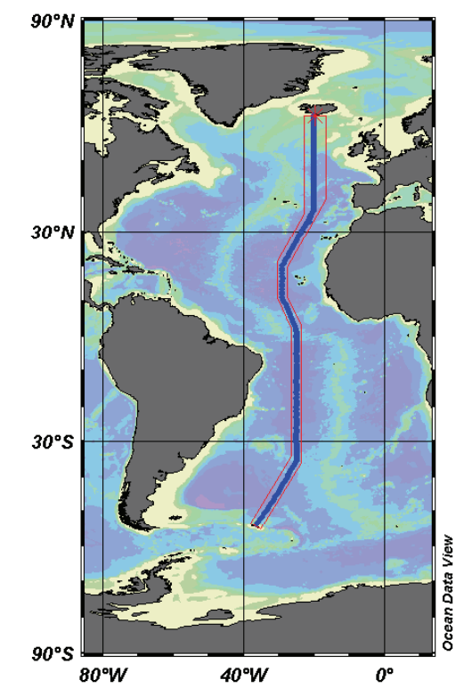
…und Einstellungen\Uwe Schindler\Lokale Einstellungen\Temporary Internet Files\Content.IE5\YO1N6E5H\event_a16…

event_a16_project_woce[1].tab

```
Date/Time Latitude  Longitude Depth water [m] Temp [deg C]  O2 [µmol/l] Sal NO3 [µmol/l]  Origin of Values
1988-07-23T21:17:00.000 63.3300 -19.9983  3 10.654  306.30  34.845   http://doi.pangaea.de/10.1594/PANGAEA.274005
1988-07-23T21:17:00.000 63.3300 -19.9983  28  10.126  287.90  34.909    http://doi.pangaea.de/10.1594/PANGAEA.274005
1988-07-23T21:17:00.000 63.3300 -19.9983  44  10.169  287.50  35.062    http://doi.pangaea.de/10.1594/PANGAEA.274005
1988-07-23T21:17:00.000 63.3300 -19.9983  69  8.153 275.80  35.098   http://doi.pangaea.de/10.1594/PANGAEA.274005
1988-07-23T21:17:00.000 63.3300 -19.9983  94  7.600 269.60  35.129   http://doi.pangaea.de/10.1594/PANGAEA.274005
1988-07-23T21:17:00.000 63.3300 -19.9983  119 7.533 268.70  35.131   http://doi.pangaea.de/10.1594/PANGAEA.274005
1988-07-23T21:17:00.000 63.3300 -19.9983  119 7.533 267.70  35.133   http://doi.pangaea.de/10.1594/PANGAEA.274005
1988-07-23T21:17:00.000 63.3300 -19.9983  119 7.534 268.20  35.131   http://doi.pangaea.de/10.1594/PANGAEA.274005
1988-07-23T21:17:00.000 63.3300 -19.9983  144 7.502 268.20  35.133   http://doi.pangaea.de/10.1594/PANGAEA.274005
1988-07-23T21:17:00.000 63.3300 -19.9983  168 7.465 269.60  35.135   http://doi.pangaea.de/10.1594/PANGAEA.274005
1988-07-23T21:17:00.000 63.3300 -19.9983  193 7.371 264.10  35.135   http://doi.pangaea.de/10.1594/PANGAEA.274005
1988-07-23T21:17:00.000 63.3300 -19.9983  194 7.371   35.134    http://doi.pangaea.de/10.1594/PANGAEA.274005
1988-07-23T23:00:00.000 63.2100 -20.0050  2 10.733  304.00  34.985   http://doi.pangaea.de/10.1594/PANGAEA.292886
1988-07-23T23:00:00.000 63.2100 -20.0050  2 10.697  304.00  34.977   http://doi.pangaea.de/10.1594/PANGAEA.274016
1988-07-23T23:00:00.000 63.2100 -20.0050  4 10.739  302.70  34.985   http://doi.pangaea.de/10.1594/PANGAEA.292886
```

Normal text file — nb char : 68905360 — Ln : 1 Col : 1 Sel : 0 — UNIX — ANSI as UTF-8 — INS





Data visualization with ODV (http://odv.awi.de)

Author contacts:

[1] MARUM - University of Bremen, PANGAEA, Bremen, Germany (uschindler@pangaea.de, mdiepenbroek@pangaea.de)

[2] Foundation Alfred Wegener Institute for Polar and Marine Research (AWI), PANGAEA, Bremerhaven, Germany (hgrobe@pangaea.de, rsieger@pangaea.de)