Rescue of DSDP/ODP/IODP post-cruise data

Updated report by

Hannes Grobe (AWI), Evgeny Gurvich (AWI),
Stefanie Schumacher(AWI) & Michael Diepenbroek (MARUM)

(2012-04-02)

Index

Motivation and aim

The ocean drilling started in 1968 and since then has generated a huge amount of datasets, shared in many national and international journals. A summary of the engineering topics and of the first scientific results of each cruise are published in the Initial Reports since DSDP. First scientific publications related to a certain leg are also published in the „Initial Reports“ of the DSDP Program and the „Scientific Results“ of the ODP Project. The majority of publications (with related primary data) is produced later on and sometimes is published years after a leg. Those "post-cruise" publications are distributed in various journals related to marine geosciences of major publishers (a.o. Science, Nature, Elsevier, Springer, Wiley) and in smaller journals, e.g. of national societies. Nearly none of the related primary data are available in machine-readable form on the Internet. In very many cases no data are available at all. Thus for scientists working in the field of marine geology, it is nearly impossible to get an overview about the availability of research data.

Data from core documentation and scientific investigations on board were published through the Initial Reports from the ODP Project (Legs 100-210) and IODP (Leg 310 ff) and are available via the ODP JANUS Database. JANUS includes technical meta-information about the cores (length, sections, recovery etc.) and core images of the DSDP Legs (1-96). Logging data were acquired and archived by the Borehole Research Group (BRG) at Lamont-Doherty Earth Observatory.

The project initiated by IODP in 2007 aims at archiving former post-cruise data, printed in individual publications in an Open Access repository. The project started by screening previously published DSDP and ODP publications, extracting the data, reformatting the data according to international standards and making those 'data supplements' available through PANGAEA®[1], in particular for the SEDIS data portal. In addition, each data table and each supplement had to be long-term identified by a persistent identifier (Digital Object Identifier - DOI®[2]).

Work flow

2.1 Reference/Data search

Data search started on the journal-level, considering all scientific fields relevant to ocean drilling (sedimentology, palaeontology, geochemistry, petrology, geophysics). As a central source of ocean drilling related publications, the **Ocean Drilling Citation Database** of the GeoRef Information Service, operated by the AGI (American Geological Institute) was used. The georef search is able to provide a list of all DDP/ODP/IODP publications related to a certain journal. The systematic search for supplementary data started on the publisher level, going through all journals of the publisher. Each publication is accessed on the publisher's webpage and the search for data is started. Recent publications often have supplementary data, which are available online in the repository of the publisher. These files had to be downloaded. Each publications was browsed for data, either on the online version or on the pdf version. Papers with datasets in the printed and/or the online version were downloaded, harmonized and archived. All geo-referenced data of a publication are considered, even if the data are not directly related to a DSDP/ODP/IODP Project.

[1] http://www.pangaea.de/
[2] http://www.doi.org/

2.2 Extraction of data from publications

Published data tables are available in different technical qualities, depending on the publication year and state, in the printed paper or in a supplement archive. Younger publications often have online accessible supplementary data in excel, text or pdf format. Tables in the paper are always integrated in the pdf-format. Any geo-referenced data are converted to excel sheets. A conversion of excel and txt files to the import format is easy; the conversion of pdf-files might required some editing.

pdf-files are opened with Adobe Acrobat Professional, pages with tables are isolated and stored as MS Word document. The document opened in MS Word allows to copy the table into an MS Excel sheet. If this flow does not work, the table has to be extracted directly from the pdf file. The table is marked and copied into a plane editor file. Blanks are replaced by tab-stops, and the document can be copied into an Excel sheet. In a few cases, the tables are integrated as a grafic object (tiff or gif) in the pdf. In this case, data have to be retrodigitized in typing by hand.

The excel sheets are quality controlled, edited and compared with the original document. Line breaks and tab-stops in wrong order may confuse the orientation of lines and columns, numbers and names may contain misspellings from the OCR process. The final editing and review can be quite time consuming. In mean, the data of one publication need about 4 hours to be transferred from its original published format to the machine-readable standard form provided by the data archive.

2.3 Preparation of data for import

The prepared and corrected Excel sheets are prepared for import. Sample information, i.e. the standard *ODP sample designation* has to be added or completed. Metadata are defined in the database. References are completed with DOI, in older publications without DOI the pdf file/page on the publishers web site is linked. All data tables from a publication are imported, in principle one published table as one dataset. In case the table contains more than one Site or Hole, a dataset is defined by Site/Hole. The dataset titel mostly is equivalent to the table/appendix number and caption. Many datasets (childs) of one publication are merged into one parent set set which also includes the abstract of the publication (extracted from the original pdf file). The data set DOI, or, in case of many data sets, the parent DOI will become the official identifier of the supplement. Always a final control in comparison with the original publication is part of the quality control and internal review process.

Examples

1. Parent set with several child datasets:
   http://doi.pangaea.de/10.1594/PANGAEA.678472
   This publication contains three tables in the pdf file. Table 3 is split to the five sites. All tables needed a time consuming review after extracting from the pdf, because of the species names. The tables were extracted via MS Word document, therefore columns and rows were in a proper order.
2. Parent set with three child data sets:
   http://doi.pangaea.de/10.1594/PANGAEA.672082
   Here we have one table in the pdf file (Table 1) and two excel sheets as supplement

(Appendix A). Table 1 was extracted via copy-past, and only few editing was needed. The excel sheets were also in a good mode for the PANGAEA import.

3. A single data set referred to a publication:
   http://doi.pangaea.de/10.1594/PANGAEA.712516
   This publication has no data tables in the pdf file, but a supplement, which can be downloaded from the publisher's web page. The supplement pdf was converted and imported and has the same status as a parent set with all information included.
4. A single data set referred to a publication:
   http://doi.pangaea.de/10.1594/PANGAEA.706057
   The Table II in the pdf file is in an very bad mode. The table is inserted as a graphic, the scan was done with a low definition. We have used the table as a hardcopy from the printed journal and created an excel sheet via data-typist.
5. Parent set with two child datasets:
   http://doi.pangaea.de/10.1594/PANGAEA.710844
   The authors also give previous published data in the tables (EPSL). These data are imported with all data of the primary publications, and also parent sets were created (Init. Rep.): http://doi.pangaea.de/10.1594/PANGAEA.710841 and http://doi.pangaea.de/10.1594/PANGAEA.710824. Now the EPSL child data sets get the relevant DOI information of the Init. Rep. child datasets (Example: For Sr and Cl data see Gieskes (1974) dataset: doi:10.1594/PANGAEA.710820).
6. In a few cases data sets were published in the ODP/DSDP Reports AND in a journal. First priority is given to the journal and the data report is listed as additional reference: http://doi.pangaea.de/10.1594/PANGAEA.706226
7. In cooperation with Elsevier, available Supplementary Data in PANGAEA® are also visible on the splash page of a publication in Science Direct:
   http://dx.doi.org/10.1016/S0031-0182(01)00497-7

Statistics

The average time needed to process and archive the data sets related to one paper is 4 hours. In general one or more data entities (e.g. data tables in an articles – the childs) are comprised in a supplementary data set (the parents). In average a supplement contains about 3 child data sets.

In total about 4500 publications were scanned, more than half of them having data sets in tables, appendices, and supplementary materials (Table 1). These data were made available in machine readable form in PANGAEA leading to 8055 data sets which were comprised in 2473 data supplements (for comparison 4/2009: 2605 data sets comprised in 788 data supplements). A large part of the effort during the last 18 months was focused on supplementary data from the "Proceedings of the Ocean Drilling Program, Scientific Results" (see last rows in Table 1). All other journals have been scanned by PANGAEA staff for supplementary data non-regarding their direct relationship with ODP programs. This explains the relative low increase. In detail the numbers by program are:

- DSDP 623 data supplements with 2235 child data sets
- ODP 1794 data supplements with 5644 child data sets
- IODP 56 data supplements with 176 child data sets

Table 1. Overview of processed articles having supplementary data sets. Numbers given are by publisher and journal. “x”: exact numbers not yet known. The update status indicates when these journals have been scanned since the last report in 2009.

Publisher	Journal	Articles with data	update status
Springer	Bulletin of Volcanology	1	2011-12-06
Springer	Climate Dynamics	1	2011-12-06
Springer	Contributions to Mineralogy and Petrology	29	2011-12-19
Springer	Deep drilling in crystalline bedrock	0	2011-12-19
Springer	Developments in Paleoenvironmental Research	0	2011-12-19
Springer	Frontiers in Sedimentary Geology	1	2011-12-19
Springer	Int. Journal of Earth Sciences (Geol. Rundschau)	14	2011-12-20
Springer	Geo-Marine Letters	4	2011-12-19
Springer	Journal of Geophysics	0	2011-12-19
Springer	Marine Geophysical Research	0	2011-12-19
Springer	Mineralium Deposita	1	2011-12-19
Springer	NATO I	1	2011-12-19
Springer	Naturwissenschaften	1	2011-12-19
Springer	Scientific Drilling	0	2011-12-19
Springer	Monography	1	2011-12-19
Springer	**total**	**41**	
Elsevier	Marine Micropaleontology	171	2012-02-21
Elsevier	Palaeogeography, Palaeoclimatology, Palaeoecology	132	2012-03-29
Elsevier	Chemical Geology	60 + 18	2012-03-28 in prep
Elsevier	Deep Sea Research	2	
Elsevier	Revue de Micropaleontology	0	
Elsevier	Geochimica et Cosmochimica Acta	102	
Elsevier	Earth and Planetary Science Letters	219	
Elsevier	Global and Planetary Change	11	
Elsevier	Quaternary Science Review	22	
Elsevier	Marine and Petroleum Geology	3	
Elsevier	Earth-Science Reviews	1	
Elsevier	Cretaceous Research	16	
Elsevier	Quaternary Research	5	
Elsevier	Marine Chemistry	4	
Elsevier	Organic Geochemistry	38	
Elsevier	Sedimentary Geology	21	
Elsevier	Marine Geology	146	
Elsevier	**total**	**899 + x**	
GSA	Bulletin	32	
GSA	Geology	89	

GSA	Geosphere	1	
GSA	Special Paper	3	
GSA	**total**	**125**	
AGU	Paleoceoanography	x	
AGU	...	x	
AGU	**total**	?	
Nature	Nature + Nature Geoscience	**61**	2011-10-21
Science	Science	**43**	2011-12-06
DSDP	Initial Results, Part II, 1 to 96	x	
ODP	Scientific Results, 101 to 128	796	2012-03-21
ODP	Scientific Results, 129 to 198	103 + x	2012-03-21 in prep
ODP	Scientific Results, 199 to 210	199	2012-03-21
IODP	Scientific Prospectus, 301 to ?	x	