Rescue of DSDP/ODP/IODP post-cruise data

Report by

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

(2009-09-10)

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

General requirements

The data rescue procedures require significant scientific expertise. Data extraction from publications and preprocessing of data can partly be done by students, whereas the final processing of data sets, annotation with metadata (including an abstract), matching of parameters, verification of methods, and import into PANGAEA® has to be done by scientists who have been working in this field.

The IODP-MI funding for this project was complemented by about the same amount of funding through the PANGAEA® group. This contributed to the significant progress made in the project.

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

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[3], 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 publication was browsed for data, either on the HTML or the pdf version. Papers with included or related datasets were downloaded, harmonized and archived. All georeferenced data of a publication are considered, even if the data are only indirectly related to a DSDP/ODP/IODP Project.

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 georeferenced 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 require 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 graphical object (tiff or gif) in the pdf. In these cases, data had to be retro digitized (transcription 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. On an average, 4 hours were needed to transfer the data related to a single publication from the original formats to the machine-readable standard provided by the data archive.

Preparation of data for import

The processed 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

[3] http://www.agiweb.org/georef/

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 related to a publication are imported. In general a table in the publication corresponds to one data set (granularity). In case the table contains more than one Site or Hole, a dataset is defined by Site/Hole. The dataset title mostly is equivalent to the table/appendix number and caption. Many data sets (childs) of one publication are merged into one parent set which also includes the abstract of the publication (compiled 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 now also visible on the splash page of a publication in Science Direct: http://dx.doi.org/10.1016/S0031-0182(01)00497-7 Currently, for technical reasons only licensed users can use this functionality. This will be changed within the next months.

Statistics

The average time needed to process and archive the data sets related to one paper is 4 hours. The data sets (the childs) are comprised in a supplementary data set (the parents). In average a supplement contains 3 child data sets.

In total 2857 publications were scanned, 1021 of them having data sets in tables, appendices, and supplements. So, 1021 supplementary data sets have been stored in PANGAEA®, comprising 3716 data sets (Table 1).

Publisher	Journal	GeoRef	Data	in %
Springer	Bulletin of Volcanology	1	1	100
Springer	Climate Dynamics	5	1	20
Springer	Contribution of Mineralogy and Petrology	37	21	57
Springer	Deep drilling in crystalline bedrock	1	0	0
Springer	Developments in Paleoenvironmental Research	2	0	0
Springer	Frontiers in Sedimentary Geology	22	1	5
Springer	Int. Journal of Earth Science (Geol. Rundschau)	46	12	26
Springer	Geo-Marine Letters	10	1	10
Springer	Journal of Geophysics	3	0	0
Springer	Marine Geophysical Research	6	0	0

Springer	Mineralium Deposita	2	1	50
Springer	NATO I	5	1	20
Springer	Naturwissenschaften	3	1	33
Springer	Scientific Drilling	13	0	0
Springer	Monography	40	1	3
Springer	**total**	**196**	**41**	**22**
Elsevier	Marine Micropaleontology	214	143	67
Elsevier	Palaeogeography, Palaeoclimatology, Palaeoecology	244	109	45
Elsevier	Chemical Geology	107	60	56
Elsevier	Deep Sea Research	15	2	13
Elsevier	Revue de Micropaleontology	0	0	
Elsevier	Geochimica et Cosmochimica Acta	278	102	37
Elsevier	Earth and Planetary Science Letters	443	219	49
Elsevier	Marine Geology *	361	220	61
Elsevier	**total**	**1662**	**855**	**47**
GSA	Bulletin	238	32	13
GSA	Geology	478	89	19
GSA	Geosphere	10	1	10
GSA	Special Paper	273	3	1
GSA	**total**	**999**	**125**	**11**
	all journals total	**2857**	**1021**	**36**

Table 1 Overview of processed journals (8/2009)
* in progress, data are extracted, but ~65% needs still to be imported

Perspective

Compared to the overall efforts for data management in 40 years of ocean drilling, the current efforts for post cruise data rescue are rather moderate. Nevertheless, the results at the end of the 2 years period suggest that it is feasible to get in a reasonable time very near to a complete archive of post cruise data.

The IODP-MI initiative is enhanced by the self funded contributions from the PANGAEA® group. Temporarily 4 curators have been working on this project. It is very likely that these activities will be continued. In fact, before 2007 more than 3000 post cuise data sets have already been archived in PANGAEA®, however, not yet as supplementary data sets to DSDP/ODP related publications. This data inventory needs to be checked and adjusted to IODP rules and standards. The efforts needed are comparatively low.

A further enhancement is possible by including the WDC for Paleoclimatology in Boulder as a new data provider for SEDIS (to be agreed with IODP-MI). The WDC for Paleoclimatology is already part of the WDC network (via OAI-PMH) and could be incorporated with little effort into the SEDIS network. This WDC holds between 100 and 1000 post cruise data sets.

Again, also this inventory needs to be scanned and adjusted to IODP rules and standards, in particular data sets have to be related to leg/site/hole information according to the IODP metadata guide.