Earth Syst. Sci. Data Discuss., 8, 447–460, 2015
www.earth-syst-sci-data-discuss.net/8/447/2015/
doi:10.5194/essdd-8-447-2015

Open Access Earth System Science Data Discussions

IPY 2007–2008 data legacy – a long story cut short

A. Driemel[1], H. Grobe[1], M. Diepenbroek[2], H. Grüttemeier[3], S. Schumacher[1], and R. Sieger[1]

[1]Alfred-Wegener-Institut Helmholtz-Zentrum für Polar- und Meeresforschung, Bremerhaven, Germany
[2]Marum – Center for Marine Environmental Sciences, Bremen, Germany
[3]Institut de l'information scientifique et technique-Centre National de la Recherche Scientifique, Vandoeuvre-lès-Nancy, France

Received: 21 May 2015 – Accepted: 01 June 2015 – Published: 22 June 2015

Correspondence to: H. Grobe (hannes.grobe@awi.de)

Published by Copernicus Publications.

Abstract

The International Polar Year 2007–2008 was a synchronized effort to simultaneously collect data from polar regions. Being the fourth in a row of IPYs, the demand for interdisciplinarity and new data products was high. However, despite of all the research done on land, people, ocean, ice and atmosphere and the large amount of data collected, no central archive or portal was created for IPY data. In order to address these issues, a concerted effort between PANGAEA – Data Publisher for Earth and Environmental Science, the ICSU World Data System (WDS), and the International Council for Scientific and Technical Information (ICSTI) was undertaken to extract data resulting from IPY publications for long-term preservation.

1380 IPY-related references were collected. Out of these, only 450 contained accessible data. All data was extracted, quality checked, annotated with metadata and uploaded to PANGAEA. The 450 articles dealt with a multitude of IPY topics – plankton biomass, water chemistry, ice thickness, whale sightings, Inuit health, alien species introductions by travelers or tundra biomass change – to mention just a few. Both, the Arctic and the Antarctic were investigated in the articles, and all realms (land, people, ocean, ice and atmosphere) and a wide range of countries were covered. The data compilation can now be found with the identifier doi:10.1594/PANGAEA.150150 and individually searched for using the PANGAEA search engine (www.pangaea.de) and adding "+project:ipy". With this effort, we hope to improve the visibility, accessibility and long-term storage of IPY data for future research and new data products.

1 Introduction

The International Polar Year 2007–2008 was a synchronized effort of over 60 nations and numerous organizations and institutes to simultaneously collect data from polar regions (Krupnik et al., 2011). During March 2007 until March 2009 a broad range of research topics were addressed, from glaciology to biology, from biochemistry to bio-

ESSDD
8, 447–460, 2015

physics, from oceanography to physiology, from atmospheric to social sciences and even human health. In other words, research on land, people, ocean, ice and atmosphere. The IPY 2007–2008 was the fourth in a series of international polar initiatives, the previous ones having been held in 1882–1883 (11 nations), 1932–1933 (46 nations), and the International Geophysical Year (IGY) of 1957–1958, which was inspired by the previous IPY (67 countries) (Bulkeley, 2010). A major outcome of the IGY was the creation of the World Data Centres (WDC) (Krupnik et al., 2011), which became the World Data System (WDS) in 2012.

The IPY 2007–2008 Data Policy (IPY, 2008) states, that "IPY generated data should be carefully and thoughtfully collected, used collaboratively, and adequately preserved." Also it postulates, that "IPY data, including operational data delivered in real time, are made available fully, freely, openly, and on the shortest feasible timescale." The IPY should help deepen the understanding of polar environmental change and its impact on society, which would require the creative use of a myriad of data from many disciplines (Parsons et al., 2011). However, despite of all the data collected, and plans of a full-time, professional data unit (IPYDIS), in the end no central archive or portal was created for IPY data (Parsons et al., 2011).

A project-specific subset of the Global Change Master Directory (http://gcmd.nasa.gov/portals/ipy/), is what comes closest to an IPY data portal. Here, a set of 642 metadata descriptions are listed for IPY (accessed 22 April 2015), being comprised of documentations, images, articles, contact links to scientists and institutes, or direct links to data centers/databases. For specific projects or countries, a multitude of smaller data centers can be found offering IPY related data: data on local/traditional community knowledge and observations, for example, is distributed via the Exchange for Local Observations and Knowledge in the Arctic (http://eloka-arctic.org). The US National Snow and Ice Data Center (http://nsidc.org/data) and the Advanced Cooperative Arctic Data and Information Service (www.aoncadis.org) as well as the Database of the Norwegian Polar Institute (http://metadata.data.npolar.no/datasets) and the DAMOCLES database of the Norwegian Meteorological Institute (http://damocles.met.no) offer data

ESSDD
8, 447–460, 2015

on Arctic IPY research. Norway also initiated the Norwegian permafrost database, NORPERM during IPY 2007–2008 (Juliussen et al., 2010). The New Zealand National Institute of Water and Atmospheric Research operates a Coastal and Marine Data Portal (www.os2020.org.nz) for selected projects, Australia stores data in the Australian Antarctic Data Centre (www1.data.antarctica.gov.au), and Canadian polar data can be obtained via the Polar Data Catalogue (www.polardata.ca).

So certainly some IPY data is available from all these distributed sources (and the upper mentioned ones are just some examples). Nevertheless, in the world of data centers and data providers, there are various things a scientist searching for data has to cope with. The problems most often encountered are: (1) project databases are not maintained after the project ends and data links thus get 404 errors, (2) the actual data is nowhere to be found, i.e. huge amounts of metadata are not backed by data, (3) the data has not been quality checked – abbreviations are not explained and/or units are missing, (4) the data is not geocoded, (5) data and metadata are stored in different files, and considerable effort has to be put into combining both to make sense of the data, (6) file based archives do not harmonize contents which significantly complicates integration of data from different sources, (7) the data is hidden in a map or figure, and the source data is not accessible, (8) data centers tend to be focused on a very specific discipline, which thwarts interdisciplinary work, (9) finding the data is hard and tedious, even with powerful search engines (Parsons et al., 2011), and lastly (10) the data is not freely available. The so-called IPY Data and Information service IPY-DIS (http://ipydis.org), had the intention to centralize and improve this situation (Parsons et al., 2011). However, this website is not supported any more.[1] This latter example nicely illustrates the importance of stable and permanent links and reliable long-term maintenance of databases.

As can be seen, IPY data mostly is fragmented, sometimes poorly managed, and often hidden or hard to find. The majority of the IPY knowledge has been reported

[1] Actually, it seemingly moved to another owner, as it presents an unreadable text in Japanese, which, according to an online translator tool, talks about “student loans”.

ESSDD
8, 447–460, 2015

in publications, and the related data mostly is concealed in pdf-tables and thus not machine-readable and unavailable for further processing.

2 Implementation

Our motivation to extract IPY data from publications was based on the following points:

- make data machine-readable and thus usable for the public
- allow the integration into existing data and thus the compilation of individual new data products with e.g. a data warehouse (which serves the IPY demand of interdisciplinarity to create new knowledge)
- improve the visibility of IPY data through portals, search engines and library catalogs
- allow Open Access to IPY results (Open Data, CC-BY licence)

In order to address these issues, a concerted effort between PANGAEA – Data Publisher for Earth and Environmental Science, the ICSU World Data System (WDS), and the International Council for Scientific and Technical Information (ICSTI) was undertaken to extract data resulting from IPY publications for long-term preservation. The data rescue started in 2013 and ended at the beginning of 2015. It was organized into iterative tasks – researching and identifying legacy data from the scientific literature, extracting (i.e. capturing or digitising) the numeric values, generating ISO 19115 (ISO, 2014) standard metadata, performing quality assurance/control processes on captured tables, and publishing data and its metadata with appropriate citations through PANGAEA (http://www.pangaea.de).

The process of researching and identifying legacy data in IPY publications began with the compilation of a list of 1380 references by ICSTI using keywords relevant to IPY projects. This bibliography served as a basis for the PANGAEA Editor to filter out journal

ESSDD
8, 447–460, 2015

articles containing extractable data – either from the articles themselves (in the form of tables) or from Supplements supplied with the publication. Extraction and digitisation of the data was performed such that numerical data or data tables were transformed into an ASCII format. Preparation of data included a technical quality control and editorial review (check for typos, correctness of geocoding and units, precision of values etc.) and annotation with metadata. Data and metadata were imported into the relational database of the PANGAEA system.

3 Outcome

3.1 Results of literature extraction

In total 450 of the 1380 articles collected by ICSTI fulfilled the criteria needed for PANGAEA. These 450 articles contained 1270 extractable datasets (i.e. data tables), which were assembled into 450 so-called parents. The parents have a clearly defined citation showing their status as a Supplement to the related paper. All 1270 datasets are now available to the public in open access (doi:10.1594/PANGAEA.150150), they are persistently identified with a citable DOI (which eliminates 404 errors) and are directly linked to the article (and author(s)!) they originate from. The data can be searched for with the PANGAEA search engine using standard search terms and adding "+project:ipy". In PANGAEA, there is no metadata without data, and no data without metadata. All data taken at a specific point in space and time are geo- and timecoded. When downloading several files of similar data, the tables can be directly compared. PANGAEA is laid out as a permanent facility, holding mandates of the ICSU World Data System as well as the World Meteorological Organization (WMO). Therefore, PANGAEA is guaranteeing the long-term availability and accessibility of the archived data and metadata in secure and machine-readable formats. For an example of an IPY dataset in PANGAEA see Fig. 1 (Toyota et al., 2011b).

ESSDD
8, 447–460, 2015

1400 different parameter/unit combinations were used to describe the data. Most parameters belonged to the field of marine taxa (abundances, biomass etc.) and chemistry (water chemistry, organic pollutants etc.), see Table 1 (PANGAEA parameter groups). General parameters included unspecific descriptive ones such as length, height, distance, type, area, number of XY, percentage, and so on. Often, a comment had to be added to these parameters (e.g. doi:10.1594/PANGAEA.836655) to describe out-of-the-way (i.e. topically exotic) data. We extracted articles on plankton biomass, water chemistry and ice thickness, but also the ones dealing with whale sightings, Inuit health, alien species introductions by travelers or tundra biomass change. Both, the Arctic and the Antarctic were investigated in the articles, and all realms (land, people, ocean, ice and atmosphere) and a wide range of countries were covered (see Fig. 2).

3.2 Other IPY related data in PANGAEA

As a voluntary contribution of PANGAEA to IPY 2007–2008, various old books on IPY 1882–1883 research were digitised, and extractable data was archived in PANGAEA (Krause et al., 2010). This additional effort resulted in 94 datasets for IPY 1 (http://www.pangaea.de/search?q=ipy-1), which can now be compared to and/or combined with recent IPY data. PANGAEA also contains 705 IPY 2007–2008 related singular datasets not connected to publications. Due to the fact, that these datasets belong to continuous, polar research observations and measurements (mostly from Germany), they are not explicitly labeled as IPY data. An overview of these datasets which are complementing the IPY data discussed above can be found in Table 2.

Last but not least, PANGAEA was one of the data sources used (Van de Putte et al., 2014) for a major outcome of the IPY 2007–2008: the Biogeographic Atlas of the Southern Ocean (http://atlas.biodiversity.aq).

ESSDD
8, 447–460, 2015

3.3 Availability and distribution of IPY related data in PANGAEA

All datasets related to the 450 publications are accessible via the DOI doi:10.1594/PANGAEA.150150. Apart from the standard search via the PANGAEA search engine, IPY data in PANGAEA is distributed and can be found in various ways: the content of PANGAEA is integrated in the Data Citation Index of Thomson Reuters. It is furthermore distributed via web services through portals, search engines and library catalogs. To give an example, the IPY dataset of Toyota et al. (2011b) can easily be found via Google with various 3–4 letter search terms. IPY datasets can also be searched for (and found) via the DataCite Metadata search (http://data.datacite.org/10.1594/PANGAEA.839311) and via the catalogue of the German National Library of Science and Technology (TIB). WorldCat – the world's largest network of library content and services – incorporates the content of all repositories, catalogs and databases which can be harvested following the OAI-PMH standard and thus also comprises the PANGAEA content with its IPY collection. Last but not least, Elsevier journals display a data-reference and a map as soon as article-related data is archived in PANGAEA (see e.g. doi:10.1016/j.dsr2.2010.12.002, Toyota et al., 2011a).

A provision of the PANGAEA content via the GCMD and in particular via the specific IPY and Arctic/Antarctic portals, respectively, is in progress.

4 Conclusions

Our effort can be described as a successful post-IPY data management. However, notwithstanding the overall positive effect of this work, it also shows the limitations of data management which is not timely synchronized with science activities. Large parts of existing data from IPY projects is still stored in publications only – the IPY publication database is still growing and now contains almost 4000 entries for IPY 2007–2008 (Goodwin et al., 2012, http://www.nisc.com/ipy/) – but no funds remain to extract the related data. Another problem encountered by the PANGAEA Editor performing the

ESSDD
8, 447–460, 2015

digitisation was that often only graphs and figures were included in the article, with no numerical data that could be extracted. Even more data is probably still hidden on personal computers and will never be available to the public. This lesson still has to be learned.

The Supplement related to this article is available online at doi:10.5194/essdd-8-447-2015-supplement.

Acknowledgements. We would like to thank Jamus Collier for reading and reviewing the manuscript, especially with regard to the English spelling.

References

Bulkeley, R.: The First three polar years – a general overview, in: The History of the International Polar Years (IPYs), edited by: Barr, S. and Ludecke, C., Springer, Berlin, 2010. 449

Goodwin, R., Tahirkheli, S., Lane, H., Duerr, R., Wallace, A., Dheerendra, P. T., and Krupnik, I.: The IPY Publications Database: Taking Action to Disseminate and Preserve IPY Knowledge, From knowledge to action, IPY 2012 Conference, Montreal, 22–27 April 2012, 2012. 454

International Polar Year 2007–2008 Data Policy, available at: http://www.ipy.org/index.php?/ipy/audience/C28/ (last access: 1 April 2015), 2006. 449

ISO 19115-1 Geographic information – Metadata, available at: https://www.iso.org/obp/ui/#iso:std:iso:19115:-1:ed-1:v1:en (last access: 1 April 2015), 2014. 451

Juliussen, H., Christiansen, H. H., Strand, G. S., Iversen, S., Midttømme, K., and Rønning, J. S.: NORPERM, the Norwegian Permafrost Database – a TSP NORWAY IPY legacy, Earth Syst. Sci. Data, 2, 235–246, doi:10.5194/essd-2-235-2010, 2010. 450

Krause, R., Grobe, H., and Sieger, R.: International polar year 1882–1883 – the digitized meteorological data legacy, WDC-MARE Reports (CD-ROM), 8, 1–25, doi:10.2312/wdc-mare.2010.8, 2010. 453

Krupnik, I., Allison, I., Bell, R., Cutler, P., Hik, D., Lopez-Martinez, J., Rachold, V., Sarukhanian, E., and Summerhayes, C. (Eds.): Understanding Earth's Polar Challenges: International

ESSDD
8, 447–460, 2015

Polar Year 2007–2008, University of the Arctic, Rovaniemi, Finland/CCI Press (Printed Version), Edmonton, Alberta, Canada and ICSU/WMO Joint Committee for International Polar Year 2007–2008, xvii + 371 pp., 2011. 448, 449

Parsons, M. A., de Bruin, T., Tomlinson, S., Campbell, H., Oystein, G., and LeClert, J.: Chapter 3.11 The state of polar data – the IPY experience, in: Understanding Earth's Polar Challenges: International Polar Year 2007–2008, edited by: Krupnik, I., Allison, I., Bell, R., Cutler, P., Hik, D., Lopez-Martinez, J., Rachold, V., Sarukhanian, E., and Summerhayes, C., University of the Arctic, Rovaniemi, Finland/CCI Press (Printed Version), Edmonton, Alberta, Canada and ICSU/WMO Joint Committee for International Polar Year 2007–2008, xvii + 371 pp., 2011. 449, 450

Toyota, T., Massom, R. A., Tateyama, K., Tamura, T., and Fraser, A.: Properties of snow overlying the sea ice off East Antarctica in late winter, 2007, Deep-Sea Res. Pt. II, 58, 1137–1148, doi:10.1016/j.dsr2.2010.12.002, 2011a. 454

Toyota, T., Massom, R. A., Tateyama, K., Tamura, T., and Fraser, A.: Sea ice and snow characteristics and heat fluxes observed during R/V Aurora Australis cruise to East Antarctica in September/October 2007 (Table 1), doi:10.1594/PANGAEA.839311, 2011b. 452, 454, 459

Van de Putte, A., Griffiths, H. J., Raymond, B., and Danis, B.: Chapter 2.1 data and mapping, in: Biogeographic Atlas of the Southern Ocean, edited by: de Broyer, C., Koubbi, P., Griffiths, H., Raymond, B., d'Udekem d'Acoz, C., Van de Putte, A., Danis, B., David, B., Grant, S., Gutt, J., Held, C., Hosie, G., Huettmann, F., Post, A., and Ropert-Coudert, Y., Scientific Committee on Antarctic Research, 510 pp., 2014. 453

Young, D. A., Wright, A. P., Roberts, J. L., Warner, R. C., Young, N. W., Greenbaum, J. S., Schroeder, D. M., Holt, J. W., Sugden, D. E., Blankenship, D. D., van Ommen, T. D., and Siegert, M. J.: Radar sounding observations including bedrock and surface elevation and ice thickness in the ASB region, Antarctica (Supplement 3), doi:10.1594/PANGAEA.838935, 2011. 460

ESSDD
8, 447–460, 2015

Table 1. Occurrence of different PANGAEA parameter groups used for classifying the IPY datasets.

Parameter group	Occurrence
Marine taxa, non-plankton	16 %
Marine taxa, plankton	14 %
General	13 %
Chemistry, organic	12 %
Chemistry, water	8 %
Terrestrial	8 %
Geology/Geochemistry/-physics	7 %
Physiology	6 %
Cryosphere	4 %
Isotopes	3 %
Hydrography	2 %
Atmospheric sciences	2 %
Oceanography	2 %
Fish (marine and freshwater)	2 %

ESSDD
8, 447–460, 2015

Table 2. Additional IPY 2007–2008 related datasets in PANGAEA belonging to ongoing polar reasearch activities (access date: 7 May 2015).

IPY project	Basis/campaign	No. of datasets	Search term (always adding "-supplement")
Lena 2007	Samoylov Station	53	"@ref29480"
ANDRILL Southern McMurdoSound	McMurdo Station	20	campaign:"SMS"
BIPOMAC	POLARSTERN ANT-XXIII/9	172	campaign:ANT-XXIII/9
DAMOCLES	POLARSTERN ARK-XXII/2	206	campaign:ARK-XXII/2
ANDEEP SCACE/SYSTCO	POLARSTERN ANT-XXIV/2	108	campaign:ANT-XXIV/2
CASO, CLIVAR, BIAC etc	POLARSTERN ANT-XXIV/3	145	campaign:ANT-XXIV/3
	POLARSTERN ARK-XXIII/3	1	campaign:ARK-XXIII/3 +rabe
Total		705	

Please note, that the datasets mentioned here can consist of core documentations, single radiosonde flights and other lower-level data types.

ESSDD
8, 447–460, 2015

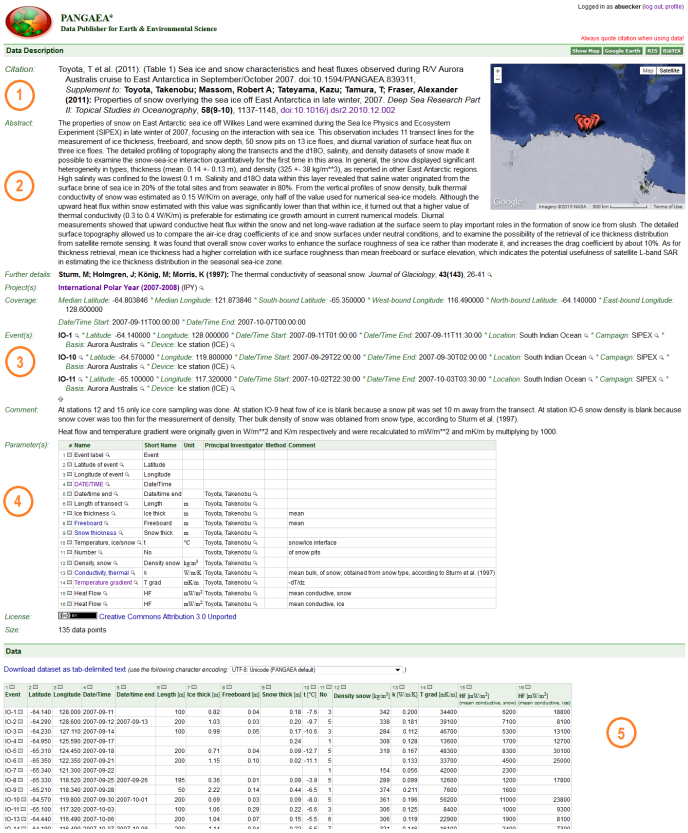

Figure 1. Example of an IPY dataset in PANGAEA, metadata (upper part, 1 = citation, 2 = abstract, 3 = georeference, 4 = parameters, units, comments) is followed by the actual data (lower part, 5) (Toyota et al., 2011b).

ESSDD
8, 447–460, 2015

IPY data legacy

A. Driemel et al.

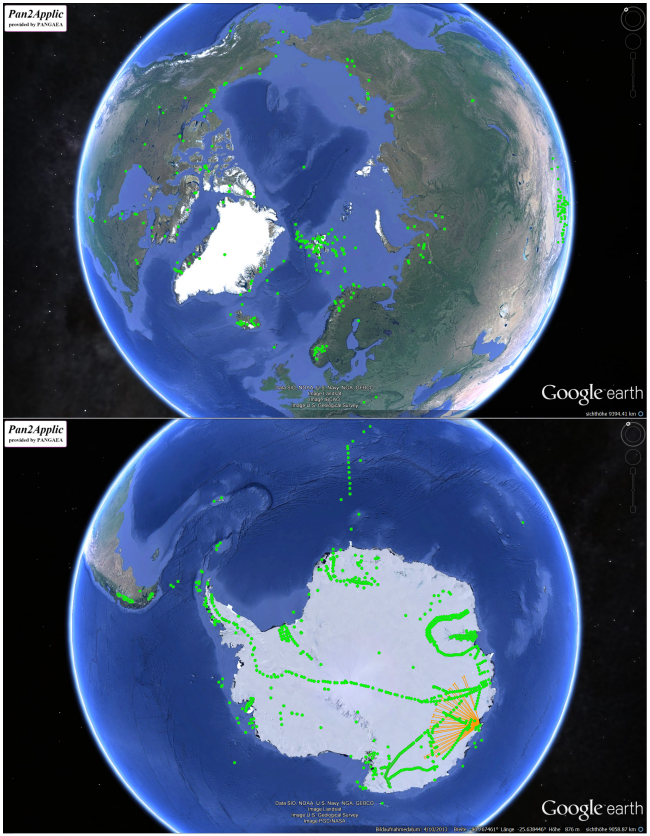

Figure 2. Overview of arctic and antarctic sample sites of the 450 IPY articles now stored in PANGAEA. Green are singular events, the orange lines belong to a radar sounding dataset of Young et al. (2011).

ESSDD
8, 447–460, 2015