Earth Syst. Sci. Data Discuss., 8, C212–C212, 2015
www.earth-syst-sci-data-discuss.net/8/C212/2015/

Open Access Earth System Science Data Discussions

Interactive comment on "IPY 2007–2008 data legacy – a long story cut short" *by* A. Driemel et al.

M. Parsons (Referee)

parsons.mark@gmail.com

Received and published: 6 August 2015

The comment was uploaded in the form of a supplement:
http://www.earth-syst-sci-data-discuss.net/8/C212/2015/essdd-8-C212-2015-supplement.pdf

Interactive comment on Earth Syst. Sci. Data Discuss., 8, 447, 2015.

[Answers and comments of the authors in bold and blue]

We thank the reviewer for his invaluable comments (being an IPY insider), which have improved the manuscript hugely. Our responses to specific points are provided below. We hope that we have sufficiently addressed all of the issues that were raised.

This paper describes a very important and creative project of what might be called near-real time data rescue. The project produced valuable results according to sound data management practice that will enhance the legacy of IPY. I was a little confused by the abstract and title, though. There is quite a bit of discussion on how IPY failed to establish a central archive. Do the authors suggest to create such an archive? This first paragraph of the abstract seems to imply that, although I doubt that is the authors intent.

→ We are aware that our effort can only be a small contribution and it is not meant to be *the* central IPY archive. We rewrote the first paragraph of the abstract to hopefully clarify this:
"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 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 improve availability and visibility of IPY data, 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."

Also, we propose to change the title to: "The IPY 2007-2008 data legacy - creating Open Data from IPY publications"

I note also that I read and concur with Ms. Friddel's review. She seems to have similar concerns.

General Comments:
"Data" are plural in formal discourse.

→ Has been changed

P. 449: line 11: I'm not sure "postulates" is the right word. "Requires" may be better.

→ "postulates" has been changed to "requires"

P. 449: line 15-18: Not strictly correct. Parsons et al. actually point out that a full-time unit was never funded (although that was recommended in the IPY Framework document).

→ We changed the sentence to "However, despite all the data collected, and plans of a full-time, professional data unit (IPYDIS), in the end no central archive or portal was funded for IPY data (Parsons et al. 2011)."

After the last paragraph on page 449, it may be worth mentioning that there has been some success federating access to some of the archives listed working toward a common (Arctic) data portal. See http://nsidc.org/acadis/search/.

→ We inserted the following sentence:

"Fortunately, there has been certain success federating access to some of the archives listed by working toward a common Arctic data portal, see http://nsidc.org/acadis/search/."

p. 450 line 19: "Thwarts" is too strong a word. A good domain repository can facilitate interdisciplinary use. Perhaps, "hinders" is a better word.

→ "thwarts" was changed to "hinders"

p. 450 line 21: The IPYDIS project was never intended to be a central archive. It was a collaboration of data centers with a small coordination office and help desk. No central archiving authority or archive of last resort was ever supported.

→ As we understand it, it was meant to act like a portal (not an archive), and that is what we mean by "centralize and improve this situation". See also the next comment and our answer.

p. 450 footnote: Just for the record: As funding ended, ipydis.org was retired and a snapshot of the website was submitted to the IPY archive in Canada. Ironically, the site has been "under maintenance" for years further illustrating the authors point about the lack of continuity (http://sunsite.ualberta.ca/Projects/IPY/).

→ If the referee is ok with this, we would like to exchange the content of our footnote with the information given above by the referee (adding "Mark Parsons, personal communication") => please give us feedback if you consent to this

P. 451 line 1: I don't think you can say data are "mostly" in publications. I understand the point that vast amounts of data are inaccessible or not machine readable, but what makes it into publications are typically tailored subsets used to illustrate an argument. The fuller collection resides on an investigators hard drive or, if we're lucky, a proper data center. Indeed the collection of 705 IPY-related data sets in PANGAEA is an example. See my comment about methods and duplication below. I'm also not sure you can even say the majority of the knowledge has been reported in publications. I would just say something like "scientific knowledge from IPY is recorded in publications, but the data behind those publications largely remain concealed and not machine readable."

→ The sentence (The majority of IPY knowledge…) has been changed to:

"A large part of the IPY knowledge is recorded in publications, but the related data mostly are concealed in pdf-tables and thus not machine-readable and unavailable for further processing."

Line 6: I don't understand the clause "e.g. a data warehouse (which serves the IPY demand of inter-disciplinarity to create new knowledge)". Is it suggested that PANGAEA be the IPY data warehouse? See my earlier comments about the intent of the paper.

→ A data warehouse is a functionality which is also available in PANGAEA. To clarify this, we added a link to an explanation of this functionality in PANGAEA (http://wiki.pangaea.de/wiki/Data_warehouse)

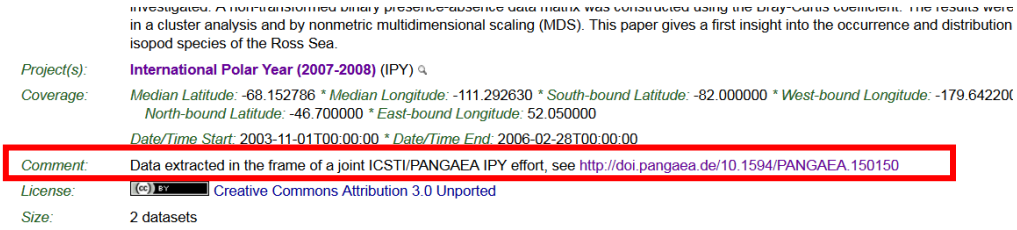
We also added an example to the bullet point as a footnote: "One example: searching for: "Chaetoceros socialis" +project:ipy, you get two hits in PANGAEA. If you remove "+project:ipy" you get 368 hits (including the two hits from IPY). Now you can click on "data warehouse" (upper right) and choose latitude/longitude and Chaetoceros socialis and you can download all abundance data on this species stored in PANGAEA."

Line 24: Is the PANGAEA Editor a human? If so, why not credit them explicitly?

→ The first author was the responsible PANGAEA Editor, so the credit is given by the first authorship.

Comments on the Data Collection and Methodology
Like Ms. Friddel, I found the data format a bit odd at first, but it opens right up in a basic text editor and seems to be quite machine-parsable.
I couldn't find a way to access the entire collection except through the individual links. it would be nice if the whole collection were available as a package suitable for analysis. At least the individual data sets could link back to the higher IPY collection (not just the IPY Web site)

→Datasets with comparable data can be combined by a search for "project:ipy" and then clicking on "Data warehouse" – see comment above (you need a login to access the Data warehouse). There is unfortunately no easy way to access all datasets at once. However, if somebody was interested in downloading all data listed in our article, we would find a way to provide him with the 1270 data tables.
To link back to the IPY collection, we added a comment to all parents containing a link to the IPY collection.


in a cluster analysis and by nonmetric multidimensional scaling (MDS). This paper gives a first insight into the occurrence and distribution isopod species of the Ross Sea.

Project(s): **International Polar Year (2007-2008)** (IPY)

Coverage: *Median Latitude:* -68.152786 * *Median Longitude:* -111.292630 * *South-bound Latitude:* -82.000000 * *West-bound Longitude:* -179.64220(*North-bound Latitude:* -46.700000 * *East-bound Longitude:* 52.050000

Date/Time Start: 2003-11-01T00:00:00 * *Date/Time End:* 2006-02-28T00:00:00

Comment: Data extracted in the frame of a joint ICSTI/PANGAEA IPY effort, see http://doi.pangaea.de/10.1594/PANGAEA.150150

License: Creative Commons Attribution 3.0 Unported

Size: 2 datasets

The authors should say more how they defined "IPY data" and the criteria for article selection, especially since they don't consider the 705 "related" data sets as part of IPY (For example, at one point Germany was considering all Polarstern data from 2007-8 to be Germany's contribution to IPY). This was an issue IPY data mangers struggled with too—what exactly are IPY data. See more discussion in Parsons, Godøy, et al. 2011.

→ We agree, the definition is rather difficult. However, we added the following information on how the IPY publication list was created in our case:

"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 as well as author and project names retrieved from IPYDIS data files. Bibliographic searches using Web of Science and Pascal databases were conducted with broad search criteria in order not to miss relevant articles."

When abstracting the data from the papers, did the authors see if the data were available somewhere else (even if not cited). For example, I suspect the example data set of Toyota et al. may be available at the Australian Antarctic Data Center, albeit under a different name and aggregated differently. This is actually quite an important issue, because we may be creating conflicting DOIs for the same data. Furthermore, since the data in an article are typically custom subsets, future data users would likely benefit more by having access to the full collection if it's available. A full data set is likely to better facilitate re-use than specific subsets. The authors should at least acknowledge this limitation and perhaps suggest some follow up on in the future.

→ We are aware of the fact, that some of the data could already be stored in a digital data center. However, in our experience, most data are hard to find (also for us, to check for duplicates), hard to obtain and not citable (only very few data centers assign dois for datasets). It is therefore neither feasible to check for the data, nor is it likely that duplicates occur very often. We are also aware that the data we extracted is only a subset, but that is the point of the whole IPY data discussion: people obviously did not want to or did not have the time to publish their original data, therefore our approach is the "next best thing to do" so to say. We added the following short abstract right before chapter 3.2:

"One drawback of this kind of data extraction is, that it is too time consuming to engage the authors of the paper in a proof-read process. E-Mail addresses often are outdated, or authors do not reply in time, and the whole process would not be feasible anymore. However, as the data have been published in an article and thereby also have been approved for public re-use (see. e.g. http://www.copdess.org/statement-of-commitment/,

we assumed that they had been quality checked by the authors before publication. Our approach also entails the drawback that publication related data are only subsets of the original research data. But in our opinion, digitising these subsets is better than having no data whatsoever."

Similarly, did the authors see if any of the data they rescued from IPY1 correspond with data rescued earlier by Kevin Wood and available at http://www.arctic.noaa.gov/aro/ipy-1/? This should be checked.

→ We cross-checked the data in our repository compared to http://www.arctic.noaa.gov/aro/ipy-1 and we found that although some of the datasets seem to be made up of the same (meteorological) data (example station Sodankylä, daily data)

1) the numbers are slightly different (e.g. daily pressure in hPa) which hints to a slightly different source, or to a statistical calculation (mean/median, etc.)

2) the http://www.arctic.noaa.gov/aro/ipy-1 does contain some data that is not in PANGAEA (e.g. pressure in mmHg, wind direction and speed at different heights (? Wind06, Wind12…),

3) PANGAEA contains a lot of data that is not stored in http://www.arctic.noaa.gov/aro/ipy-1 (soil data, hourly meteorological data, water temperature)

and it should also be stated that

4) http://www.arctic.noaa.gov/aro/ipy-1 does not provide citable DOIs for the data.

If you'd like to check the above mentioned statements, here is the link to PANGAEA data on the station Sodankylä:

http://www.pangaea.de/search?ie=UTF-8&q=event%3Alabel%3ASodankyl%C3%A4

Of course it is beyond the authors's control, but it is unfortunate that the full metadata for the data—the article— is usually not open access. This is clearly a longer-term issue to address in using this data rescue approach. This should be clarified in the text. Also the authors should clarify that the abstract listed for a data set is actually the abstract for the parent article not the data per se.

→ As you already said, that is not in our control, and a researcher should be aware of the restrictions of article access. We are here talking about the data and metadata in PANGAEA, and both are open to the public.

The authors need to describe more of their licensing approach. It is not clear that they have appropriate rights to assign a CC-BY license to the data sets without consulting both the original authors and the publishers of the papers. With the Elsevier journals they have a nice relationship established where there is a link to the data even if the article is pay-walled. That implies they have an agreement, but what about some of the other publishers and authors? Also it is not always reasonable to assume that the "authors" of a data collection are the same as the authors of the paper.

→ This is an important issues that has already been addressed in various commitments and press releases, see e.g.

http://www.copdess.org/statement-of-commitment/

http://www.nap.edu/openbook.php?isbn=0309088593

The authorship of article related data is of course a delicate issue, which we can unfortunately not solve individually in these kinds of data rescue activities. However, also for recent data submissions we mostly use the full authorship of the article for the related data, and I can say that around 99% of the scientists are completely fine with that.

page 454 line 6: please provide some examples of these portals and catalogs

→ A link to our wiki with more information was added:

http://wiki.pangaea.de/wiki/Portal

Suggested references:
Mokrane M and MA Parsons. 2014. Learning from the international polar year to

build the future of polar data management. Data Science Journal 13:PDA88-PDA93. http://dx.doi.org/10.2481/ dsj.IFPDA-15

Parsons MA, Ø Godøy, E LeDrew, TF de Bruin, B Danis, S Tomlinson, and D Carlson. 2011b. A conceptual framework for managing very diverse data for complex interdisciplinary science.
Journal of Information Science 37 (6): 555-569.
http://dx.doi.org/10.1177/0165551511412705