Earth Syst. Sci. Data Discuss., 8, C187–C192, 2015
www.earth-syst-sci-data-discuss.net/8/C187/2015/

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

J. Friddell (Referee)

julie.friddell@uwaterloo.ca

Received and published: 23 July 2015

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

We thank the reviewer for her insightful comments, 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.

The article describes the “dataset” (collection of datasets) clearly and succinctly and also describes additional datasets or resources that are available for accessing related data, which is very helpful. I have a number of suggestions for improvement of the article text:

1. The title does not represent the project appropriately. It should instead give the reader an indication of the scope, purpose, or outcomes of the data compilation/rescue project, rather than alluding exclusively to the IPY data issues.

→ Title has been changed to: “The IPY 2007-2008 data legacy – creating Open Data from IPY publications”

2. It may be helpful to have a very brief explanation of how the data were digitized (specific method(s) used).

→ As there are several ways - pdf extraction programs, special scanners, by hand, etc. - to digitize tables (in our case it was ABBYY FineReader 11), which are all valid and useful, we would like not to dwell on that. The important point is that the process is quality checked, so that no errors are inserted.

3. Page 449, introduction of GCMD: Does the GCMD hold only metadata, and not data? This should be briefly clarified.

→ The following sentence was inserted: “I.e., GCMD only contains data set descriptions, and links either directly to external sources (datasets) or to the data centers, where data are supposedly stored.”

4. Did the authors contact the authors of the original papers (and old books, if any those authors are still working) to discuss the project and the license terms? How were the licence terms chosen? It would be helpful to have some indication of whether this project was done in collaboration with some, or all, of the publication authors and if they are aware of this new availability of their data in PANGAEA.

→ The licence terms is an important issue 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 following short abstract was inserted 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.”

Page 451, Implementation: It is stated that this project will allow integration into existing data. Can this be further explained, or examples given?

→ We added a footnote under chapter 2 (Implementation): “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.”

5. Did the authors check to see if the data or metadata have been deposited elsewhere? It is possible that the datasets may already be archived in other repositories besides PANGAEA, or they may be sub-sets of larger datasets which are either already available in other repositories or are described by metadata in another repository. In this case, it is possible that a DOI has already been assigned by another repository, and the data producer is intending to deposit the data there. If this exploration has not been done, I acknowledge that it may be very difficult to make these links. Additionally, it is possible that the ideal goal of deposit of the full, original dataset may never be realized (or realized well into the future). However, it is important to keep in mind that duplication of datasets may be occurring. It would be worthwhile to mention this caveat, or the authors’ actions in preventing it, in the article.

→ 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.

6. Page 453, lines 14-15: At least one of the datasets reviewed from the first IPY contained data up to 1939. Perhaps the description in the article should be clarified to indicate that the books include data from the first IPY as well as time periods before and after.

→ As IPY publications always also included comparisons to earlier or later time-slices we would like to leave it as it is

7. Page 453, lines 22-24: Can the contribution of PANGAEA to the Biogeographic Atlas be explained briefly? It is not clear from the sentence what the Atlas is or how PANGAEA data contributed to it.

→ We added the following sentence:

“The Biogeographic Atlas of the Southern Ocean was published as a compilation of all benthos data available so far from the Southern Ocean floor. Due to the fact that many scientists from the international community archived their data in Pangaea, the repository could make a substantial contribution to this census as an IPY legacy.”

8. Page 454, lines 7-8: “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.” Can some of the 3-4 letter search terms be provided, to illustrate the example?

→ The following examples have been added to the text in brackets:

“try e.g. sipex snow toyota, tateyama ice sipex, or searching for the complete title of the article”

9. Figure 1: Not all datasets are available to be displayed as an HTML table as in the Figure (for example, http://doi.pangaea.de/10.1594/PANGAEA.837319 is only available as a .zip download). It appears that this occurs when a PANGAEA parent page links to more than 1 dataset/file, so the multiple datasets/files are made available as a .zip file. It should be explained that for “parents” which link to multiple datasets, clicking on the links at the bottom of the initial page will bring the reader to a second page in which the single-dataset information and access can be found. This will be very useful to users who are new to PANGAEA. Additionally, sometimes Events, Comments, Parameters, and other information is not present (for example, http://doi.pangaea.de/10.1594/PANGAEA.150150, the page for this article, does not have an Abstract), and the Size does not provide the number of data points on the

parent pages. These variations in presentation should be explained so that readers will know all details shown in Figure 1 are not available for all 450 + 94 datasets.

→ You are right, we should have clarified this. The beginning of chapter 3.1 has been changed to:

“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'. Meaning that, if an article contained several datasets, the datasets ('childs') were combined into a general parent with slightly reduced (general) metadata, and with links to the single datasets containing all metadata and data. The parents always have a clearly defined citation showing their status as a supplement to the related paper.”

10. Figure 2: It is difficult to see the green dots and orange lines on the maps. Perhaps there would be a better way to display the points and lines that would be easier to see.

→ We tried to improve the figure

11. This is a question for PANGAEA rather than for the article, but why is .tab the preferred extension/format for downloaded files? This is a less common format than .csv and some others. Perhaps the motivation for this format is explained on the PANGAEA website and could be referenced/linked in the article so that readers will be aware of the download format which is available from PANGAEA.

→ PANGAEA uses the tab format, because when you click on a txt weblink the txt opens in the browser, and does not show a download window. With the tab format, a download window appears. The .tab can be changed to .txt as soon as you saved it on your desktop (just rename the file and add .txt instead of .tab). However, you can also open the .tab file in Excel (so you don’t actually have to change it to .txt).

Grammar/Style

All general comments and specific suggestions have been implemented:

General comments:

1. Although most English speakers use “data” in the singular sense and it is increasingly accepted as accurate usage, “data” is technically a plural word (“datum” is the singular). To be technically correct, “data,” when indicating multiple data points, should be written as “data are” or “data were” instead of “data is” or “data was.”

2. Dominant usage is for “i.e.” to be followed by a comma.

3. This may be for ESSD, rather than the article authors: Could the links in the .pdf version of the final article open in a new tab or window? Currently, the reader loses the article itself when clicking on linked resources.

Specific suggestions:

p. 448 Lines 4-5: “However, despite of all the research done on land,. . .” Recommend to remove “of” after “despite”. Lines 15-16: “Both, the Arctic and the Antarctic were investigated in the articles,. . .” Remove the comma after “Both”

p. 449 Lines 2-3: “In other words, research on land, people, ocean, ice and atmosphere.” This is not a complete sentence. Line 15: “However, despite of all the data collected,. . .” Recommend to remove “of” after “despite” Lines 20-21: “. . .being comprised of documentations,. . .” Recommend to change to “documents”

p. 450 Line 21: “The so-called IPY Data and Information service IPY-DIS. . .” Need to capitalize “Service” Footnote: Is the text actually “unreadable”? Perhaps remove this word.

p. 452 Line 5: “. . .(check for typos, correctness of geocoding and units, precision of values etc.). . .” Recommend to precede “etc.” in a list with a comma, i.e., “. . .precision of values, etc.” There are several other locations where this usage exists, including Table 2. Lines 15-16: “. . .and are directly linked to the article (and author(s)!) they originate from.” Recommend the following: “. . .and are directly linked to the article and authors from which they originate.”

p. 453 Lines 3-4: “. . .chemistry (water chemistry, organic pollutants etc.), see Table 1 (PANGAEA parameter groups).” This is a run-on sentence. Would recommend “. . .organic pollutants, etc.). See Table 1 for PANGAEA parameter groups.” Lines 9-10: “Both, the Arctic and the Antarctic were investigated in the articles. . ." Recommend to remove the comma after “Both” Lines 18-19: “Due to the fact, that these datasets belong to continuous, polar research observations. . .” Recommend to remove the comma after “fact”

p. 454 Lines 8-9: “IPY datasets can also be searched for (and found). . .” I would recommend to write this as “IPY datasets can be discovered. . .” Line 14: “. . .thus also comprises the PANGAEA content with its IPY collection.” Would recommend to use “contains” or “includes” rather than “comprises”

p. 458 Table 2: “McMurdoSound” Recommend to add a space between “McMurdo” and “Sound”

p. 460 Figure 2 caption: “Arctic” and “Antarctic” should be capitalized.

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