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

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.

J. Friddell (Referee)

julie.friddell@uwaterloo.ca

Received and published: 23 July 2015

This is a very impressive project and a huge data rescue effort on the part of PANGAEA, WDS, and ICSTI. These organizations and the individuals who performed the work are to be commended on a substantial contribution to the body of IPY data available to the world. As stated in the article, the data are not new, but they are newly available in machine readable form, which is a huge advance and benefit for reuse of data from the IPY 2007-2008 program. The data cover a huge variety of fields and data types and represent a great deal of effort in collection/creation of the original datasets, thus this large effort to bring data into a uniform, machine-readable format should be useful to any researcher planning to do synthesis studies related to the polar regions or wanting to explore available data for the Arctic and Antarctic.

The data are accessible as outlined in the article (I have checked the links in the article

as well as a fair number of additional parents/datasets), but it is not feasible to check each of the 1270 plus additional 94 datasets that have been added to PANGAEA during this project. It is very difficult to evaluate completeness, accuracy, and other quality aspects of the various datasets without reading every article and book used in the project (450 + 94). From the datasets that I viewed, it appeared that there is sufficient information, appropriately organized, to facilitate reuse. For the remainder of this review, I will focus on the approach of the authors in the project and the article itself, rather than the datasets (and accompanying metadata) which seem to conform to PANGAEA standards, as far as I have checked.

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.

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

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

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.

5. Page 451, Implementation: It is stated that this project will allow integration into

existing data. Can this be further explained, or examples given?

6. 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.

7. 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.

8. 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.

9. 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?

10. 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.

11. 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.

12. 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.

Grammar/Style

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.

References Why are there page numbers after every reference? Will these be removed in the final published version?

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