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.

While the lack of full availability of IPY data is certainly an issue, and the authors are to be praised for helping fill the gap, I don't think the issue is the lack of a central archive as such. Indeed I think the authors underestimate some of the complexity of interdisciplinary data stewardship and re-use. (They may want to consult Parsons, Godøy, et al. 2011 and Mokrane and Parsons 2014 which provide more context on why a central archive was not developed and some of the lessons learned). The point of this paper, however, seems to be primarily to describe their data rescue effort and the resultant collection. The title and abstract should be modified to reflect the true intent of the paper, and not suggest that they have solved the IPY data problem. I also have some questions and small concerns about their methods and assertions. Overall, though, this is a noteworthy effort, and I think the paper should be accepted, especially if it can address the following comments.

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.

P. 449:

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

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).

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/.

p. 450

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

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.

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/).

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."

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.

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

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)

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.

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.

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.

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.

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.

page 454

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

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