Earth Syst. Sci. Data Discuss.,
https://doi.org/10.5194/essd-2018-32-RC1, 2018

Earth System Science Data Discussions

Interactive comment on "OCTOPUS: An Open Cosmogenic Isotope and Luminescence Database" *by* Alexandru T. Codilean et al.

G. Balco (Referee)

balcs@bgc.org

Received and published: 19 May 2018

0. Summary:

This paper describes an online database mainly containing a large compilation of cosmogenic-nuclide data useful for erosion rate estimates. This is a data set that's extremely important in geomorphology for synoptic analysis of how erosion works on Earth generally, and this effort to develop a centralized repository of these data is potentially a very big improvement for this field.

With respect to the cosmogenic-nuclide database itself, this has the potential to make major progress (the luminescence data are interesting but it is not yet as clear how important they are for synoptic use...that seems like more of an if-you-build-it-they-will-

come project). I want to emphasize the potential progress here: this is good. However, some strange and unexplainable decisions seem to have been made by the developers with regard to accessing the database, and these issues keep the work potentially transformative instead of actually transformative. In effect, the authors have taken one big step forward, but then taken several little steps backward, which leaves the overall direction of net progress unclear. Why do this? The backward steps are enumerated in section 4 below, and I think the authors should not take them.

With respect to the paper and whether or not it adequately describes the product, there are three deficiencies that should be corrected before publication. One, the authors must give readers the info on how to access the WFS/WMS directly. Two, the paper must include a description of how data were selected for inclusion. A third minor point is that the paper could be improved by removing a handful of laudatory remarks, aimed at telling the reader how important the work is, that do not contribute to describing the database.

1. Historical details and disclaimer:

First, I describe the sequence of events leading to my reviewing this paper. Readers should consider this and decide whether it's relevant to assessing whether any conflicts exist, and whether or not the review is valuable. Initially, a couple of months ago, I received an email notice of the existence of the 'OCTOPUS' database that was widely circulated on listservs. I looked at the 'OCTOPUS' website and posted a blog entry about it at the following address:

https://cosmognosis.wordpress.com/2018/04/13/putting-the-o-in-octopus/

In this posting I pointed out that it was not possible to directly access any source data in the database without providing an email address and waiting to be supplied with a file to be downloaded. I was critical of this decision on the basis that prohibiting anonymous access was inconsistent with the stated "open" nature of the project, and argued that the overall project had great potential to improve how these data are used

in geomorphology, but this potential would remain unrealized without allowing open access to the data. An editor of this journal read the posting, and sent me an invitation to review the accompanying paper.

Subsequently, one of the authors of this paper, in a comment on the posting (also at the above link) responded to this criticism by describing me as wearing a "tinfoil hat." This refers to the belief, which is commonly associated in popular culture with clinically paranoid or otherwise mentally disturbed individuals, that an aluminum-foil head covering will protect one from mind-control and brainwashing technology attributed to government agencies or extraterrestrial beings. The implication is presumably that my criticism about database openness is a symptom of mental illness.

So if this is true, readers probably shouldn't pay too much attention to this review. But, oddly, the editor has not yet retracted the invitation to review, so hopefully people can look at the evidence and come to their own conclusions.

2. What's actually being reviewed here?

Presumably for a journal like ESSD the main thing a review should consider is just whether or not a paper accurately describes an accompanying data set, such that readers can have all the facts they need to determine whether the data are correct, relevant, or useful. In this case, however, the paper is not so much describing a data set as a means of accessing previously published data, so in addition to evaluating whether the paper is an adequate description of the product, I also have comments on the product itself and whether or not it meets the expectations laid out for it in the paper. So,

3. Does the paper accurately and completely describe the product?

Overall, the paper is fairly straightforward and does a good job of explaining what is happening. So for the most part this is fine, but there are (i) two serious omissions in the paper that must be corrected, and (ii) a few areas of unnecessary and irrelevant

material that should be removed.

3.A. The first serious omission is that the paper goes to great lengths to highlight that the core of the data management system is an OGC-compliant web map server/web feature server, but does not give any of the information needed for users to connect to these servers. I find this entirely inexplicable and I cannot understand the rationale for this decision.

For background, what is being described here is a piece of server software that responds to requests from a geographic information system to supply locations and attributes of spatially located features. Thus, GIS software, e.g., ARCGIS, QGIS, etc. running on a user's desktop can connect to the web map server to query, display, and analyse data stored on the server, rather than having to acquire and store a local copy of the data. Specifically, a web map server (usually WMS) serves raster data, and a web feature server (WFS) serves point or vector data.

WMS/WFS systems are fairly commonly used to provide access to geographic databases, and in a general sense this is extremely useful, because it relieves a user of data storage, versioning, and maintenance responsibilities, and ensures that whatever analysis one is conducting is always acting on the current and up-to-date version of the data. In this specific case, this feature of the "OCTOPUS" system is the one thing that has the greatest potential to transform how these data are used in synoptic geomorphology research, by, for example, making it so that all the cosmogenic-nuclide data are just as accessible for spatial analysis as digital elevation models, hydrologic data, imagery, etc., and by making it so the results of analyses can dynamically reflect the current state of a continually improving data set, rather than being static and rapidly obsolete. I emphasize that this is a potentially extremely valuable contribution. The text of the paper correctly highlights this. Inexplicably, however, neither the paper or the website explain how a user can connect to and use this capability. Figure 1 in the paper clearly shows this architecture as the dashed lines connecting the 'geoserver' block with the desktop GIS systems, etc., on the right side of the image. However, this

part of Figure 1, mysteriously, is shown in light gray, whereas the pipeline to the web site is shown in black. This is completely mystifying. What does this mean? Does the combination of the light gray with the absence of any information on how to access this capability mean that this architecture exists but we are not allowed to use it? Does it mean that it might exist in future? Does it mean that it is only accessible to certain people?

As discussed in more detail in the blog posting referred to above, about five minutes of looking around in the website source code reveals that, in fact, the web map server and web feature server aspects of the architecture do exist, and it's fairly straightforward to find their URLs. I tried connecting to these and found that they are, in fact, publicly accessible. This is really good. It's a major potential contribution to using the data in this database for synoptic analysis. In my view it's by far the most important contribution of the work described here. So why have the authors not highlighted this? This paper focuses on the website, which is nice-looking but a much less useful contribution from the perspective of data analysis, and completely fails to mention the existence of the WFS server. I simply cannot understand this. Are the authors concerned that their machine doesn't have the bandwidth to support an arbitrary number of WFS connections? Do they not want to supply data without collecting information about the users? What is the thinking here?

In short, by creating an open web feature server to serve these data, the authors have made a really useful contribution. It is a rather stunning and incomprehensible omission that neither the paper, the website, or anything else tell users how to utilize this service.

3.B. A less important, but still important, omission in the paper is that there is no, or very little, information given about how data were selected for inclusion. The repeated use of the term "global" and the extensive remarks about how this project will make obscure studies "discoverable" give the impression that the database is intended to contain all known measurements of cosmogenic nuclides in river sediments. I conducted an entirely non-random and self-serving test of this by looking for data from

three papers that describe studies near me in California or that easily came to hand: first, the only two papers that I've written that contain Be-10 measurements on fluvial sediment (Fisher Valley, Utah, 2005 in ESPL and the US Pacific coast ranges, 2013, in American Journal of Science), and then a paper about the Eel River in CA by Willenbring and others (Geology, 2013). Although these are all fairly old at this point and in reasonably credible journals, I did not find data from any of the three in the global compilation (or at least they were not displayed in the map interface). Of course, it is hard to keep up with rapidly growing data sets, and I am sure the developers are right now receiving innumerable emails from authors whose work is missing. In addition, it is possible that the omission of my papers, at least, is completely justified by the tinfoil hat issue discussed above.

Seriously, the point here is not to criticize the authors for failing to include my own obscure papers, but rather to point out that the authors must describe in the present paper how they selected data for inclusion. For example, simply ingesting the widely used spreadsheet originally put together by Eric Portenga would not have identified the three studies noted above, but a keyword search in Google Scholar would probably have found them. To summarize, this paper must include information about (i) how "studies" were selected to be included, and (ii) whether any editorial selection of studies took place, e.g., whether the authors decided that any work was of insufficient quality, or lacked sufficient documentation, to be included. Personally, I think it's important to include everything, even things that are suspected to be inaccurate or incorrect: if we are really doing science here we should be able to distinguish high- and low-quality studies by applying objective criteria to the data themselves. But the point is that there is no information in the paper to indicate whether the authors agree with me, or if they think the highly-curated approach of only including data they perceive to be high-quality is better. This is an important issue, and leaving it vague makes it hard for readers and users to assess the credibility of the entire project, because it remains unclear whether one is really working with all known data, or just the data that the present authors think are good. The paper should be amended to address this issue directly.

3.C. Laudatory remarks not related to factual description of the data set. The paper contains several remarks that, instead of describing the work, express opinions about how valuable the work is. The main examples include page 1, lines 14-21, and page 10, lines 15-23. These should be removed: the authors should give the reader the facts and let readers decide for themselves whether the product is useful or not. If something's really a contribution, people will recognize that without being told.

4. What about the product itself?

The summary here is that the product the authors are describing – the database and, hopefully, the ability to connect to it by various means to carry out synoptic research – is potentially an important step forward in the field. But the product also takes several steps backward, which limits its usefulness. The authors should undo the backward steps.

For some context, the existing state of the art for synoptic analysis of cosmogenic-nuclide erosion-rate data consists of several compilation spreadsheets, generally available as appendices or supplemental data to papers, that contain locations, Be-10 concentrations, and derived erosion rates. Anyone interested in this subject has, most likely, downloaded one or more of these spreadsheets, combined them, corrected some errors, added some new data, possibly deleted some old data, recalculated some or all of the erosion rates, etc.. When multiplied by the entire set of geomorphologists interested in this subject, the result is a proliferation of similar but mutually inconsistent spreadsheets, each of which share some information and some errors but not others. And then, in addition, the derived erosion rates in everyone's spreadsheet need to be recalculated each time we come up with a new way to compute cosmogenic-nuclide production rates. This system, of course, is very well designed to maximize redundancy, omissions, duplications, unnecessary work, and opportunities to make errors. It's terrible.

So if we can move from everyone having mutually redundant and inconsistent spread-

sheets to a single, easily accessible database that is generally believed to be correct and complete, we have already made major progress in improving synoptic understanding of these data. The present work takes a big step forward in this direction. However, it also takes unnecessary, and to me inexplicable, steps backward. Here I highlight these backward steps and argue that the authors should not take them. Go forward. Don't go backward.

4.A. Backward step 1: why download data? As just discussed, the enormous advantage of a system like "OCTOPUS" is that we replace multiple inconsistent and variably out-of-date spreadsheet copies of a data set with a single, internally consistent, and (hopefully) up-to-date data source. This potential advance is, basically, completely nullified if the data access and distribution model is for all users to download a local copy of the database. The whole point of constructing a central online database is so that you don't have to keep a local copy on your own machine, you don't have to worry about updates and versioning, and whatever calculation you are carrying out always acts on the most complete and up-to-date version of the data set. As also discussed above, the developers have already built the infrastructure to provide live connections to desktop analytical software, which is the whole point. Why, then, have they failed to take advantage of it by directing users to download a local copy of the data instead? In this model, we are just replacing a proliferation of mutually inconsistent spreadsheets with a proliferation of mutually inconsistent KML files. That does not seem like progress. At the risk of being redundant, I simply can't understand this decision, and I cannot strongly enough urge the developers to rethink it.

4.B. Backward step 2: trust but verify. I argued above that it is an enormous step forward to replace the inconsistent-spreadsheet system with a single online database that's generally agreed upon to be correct and complete. A really important part of this idea is the "generally agreed upon to be correct and complete." It's critical that users trust that this is true, or else there is no incentive to use the database and the potential advances never happen. Thus, one of the most important aspects of user

interface design for something like this is to make it as easy as possible for users to spot-check and verify that data are as they expect. The easiest way to do this is just to have complete, granular access to all of the source data at the individual-sample level via a web browser. If I'm an author who's generated data that are included in the database, I want to be able to quickly look at, for example, Be-10 concentrations associated with samples to verify that, for example, standardization mistakes haven't been made. If I can't do this, it's very hard to accept the database as an authoritative information source. As described in my blog posting, on initially looking at the web interface, I was truly amazed to discover that sample names/numbers displayed on the map interface are not, in fact, live links to a complete data report on each sample. It took a lot of clicking before I came to the realization that, really, there was nothing to click on. Thus, I cannot, to follow the above example, verify that Be-10 standardization was done correctly without going through the entire download procedure and sifting through the resulting file. The inability to directly look at individual sample data through the website is an enormous blow to the overall credibility of the project and, again, I cannot understand why this decision was made. The coding needed to have each sample number displayed in the pop-up boxes in the map interface link to a complete data report on each sample is trivial. To summarize, of course I'm not in a position to force the developers to adopt a particular user interface, but the omission of the simple and expected ability to click down through a series of links to a granular listing of the source data creates a serious obstacle to building credibility and trust in the project overall. The developers should seriously rethink this decision.

4.C. Backward step 3: why is data access hard and not easy? This section summarizes several points I've made already: the focus on a download model for data distribution; the inexplicable failure to highlight the most potentially transformative parts of this project, the WFS/WMS implementation; and the lack of a mechanism for granular examination of the source data through the website. All these decisions, basically, make it harder, rather than easier, to access the data. Frankly, this makes no sense. The real transformative value of this resource in synoptic analysis comes when it's avail-

able through multiple channels in multiple ways: examination of individual observations through the website; dynamic incorporation in GIS systems via the OGC server; perhaps additional interfaces to extract data sets associated with individual publications; data easily formatted for use in other online or desktop systems for computing erosion rates from cosmogenic-nuclide concentrations. Understandably, unlimited resources are not available to produce a special channel for everyone's workflow, but in this case the developers have already done the work (in the case of the WFS server) or could do the work with minimal effort (in the case of generating data reports on each sample from the web interface). Overall, what I see here is a philosophy of restricting, or bottlenecking, data distribution rather than encouraging it, and I very strongly encourage the authors to change this emphasis.

5. Technical details. Although this is not relevant to evaluating whether this paper is suitable for publication, I have some technical comments on the database architecture. Mainly, the documentation of the database schema that is given in the supplemental data seems to have some features that I think are going to cause trouble in future. As far as I can tell from this documentation, the database is structured as a single table in which each sample has a single record. The problem with this is that it creates a lot of redundancy: for example, information about each publication has to be redundantly included in each record for each sample associated with that publication. And in addition, it creates difficulty in situations where, for example, two Be-10 measurements are made on the same sample. Normally in relational database design one would deal with these problems by using multiple tables: for example, each record in a 'publications' table would uniquely contain information about a particular publication, and each sample record would contain only a pointer to the relevant publication record and not all the information about the publication. The overall idea is that each piece of information should be recorded in one and only one location in the database; this minimizes errors and makes error-correction and updating easier.

In the luminescence database, this appears to be worse, because each sample record

appears to redundantly contain information about stratigraphic sections and other site-specific data as well as publications. I don't expect the developers to run out and redesign everything to fix this, but I foresee some difficulties with management and extensibility in future.

6. Minor detail in the paper text.

Page 5, line 19. It's not clear to me why it's important that past calculations be reproducible, because if a calculation method is updated to a new version, the reason is that the old version was found to be incorrect or inaccurate in some way. So why would you want to reproduce something that you know is wrong? The whole point of a project like this is that the calculations are always internally consistent and up-to-date.

Interactive comment on Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2018-32, 2018.