Earth Syst. Sci. Data Discuss.,
https://doi.org/10.5194/essd-2018-32-AC4, 2018

Open Access Earth System Science Data Discussions

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

Alexandru T. Codilean et al.

codilean@uow.edu.au

Received and published: 4 July 2018

Dear Greg

Thank you for your detailed comments. We respond to them point by point below, and for clarity use your headings to identify the points to which the responses belong to.

1. Historical details and disclaimer.

Although this discussion has nothing to do with the current manuscript, I probably need to qualify my reference to tinfoil hats in my response to your blog entry. The latter was indeed critical of our decision to deliver download links via email, and we acknowledge that at first glance, our choice of requesting an email address to download data might

send the wrong message.

Below is a short explanation on why we decided to go down this path and why the information is stored rather than discarded.

Why request an email address?

The complete database is about 166 GB in size. Of this, the files belonging to the CRN collections, occupy 43 GB and the OSL/TL collection occupies 123 GB. This is the actual data and whatever system one creates, the goal of that system is to make those 166 GB of data available to users in a relatively easy way.

We could have made our lives very simple and upload the entire database to an FTP server and let users connect and download. One criticism to this option would have been that it is impossible to see what is what without a catalogue or a map.

The obvious next step would have been to visualise data on a map and allow users to select data to download using the map. This is what the web interface is aimed at achieving. This and nothing more – at this stage.

The next problem is the size of the data. The easiest way to allow users to download an arbitrary number of large files is to provide them with a list of links and allow them a certain amount of time to be able to comfortably download the data. One of the easiest ways to deliver that list along with some instructions is via email. This is the reason behind the web interface requiring an email address. Further, this is not a new thing – for example the new CRONUS Web Calculators (http://cronus.cosmogenicnuclides.rocks/2.0/) also require an email address and will deliver results to that address. Many data catalogues available on the web will require email addresses to manage data downloads.

The use of an email address is also making life on the server side easier. In the current version, the various data collections have an associated 270 zip files with sizes ranging between a few MB to 2.5 GB (totalling 166 GB). Users can select any of these files in

any configuration during a download session.

If for example we would have chosen to deliver everything to each user in one large zip file, each download session would have needed to produce a potentially very large file and store that in a folder accessible from via a web browser until the download had been completed. The latter would have required significant computing power (to cater for simultaneous downloads as compressing and moving large files needs resources) and substantial free disk space on the server – the latter does not come cheap.

By providing individual links to each zip file in a list sent via email, all that needs to happen on the server side is to generate a folder and place symbolic links to the zip files selected for download into that folder. This can be achieved quickly and at minimal disk space cost – even with many simultaneous download requests.

Why store user data?

When downloading data from the OCTOPUS web page, the user is asked to enter the following info: name, email, and intended use of data.

Except for a valid email address, none of the other bits of information are mandatory. In fact, there is no script in place to check whether the information actually makes sense. Once the user selects the studies to download and clicks the 'Request Download' button, a list of links is immediately sent to the email address provided. There is no verification of who the data requestor is or where that person is from.

All the information entered on the webpage is captured in a log file. OCTOPUS was made possible due to funding from the Australian National Data Service and additional support provided by the University of Wollongong. As with any funding agreement, we are required to provide reports to the funding body, and actual data on downloads and what the data is used for is extremely valuable in this respect. Also, such data will also be invaluable when applying for aditional funding.

Although not mandatory, providing some meaningful info when downloading the data

is supporting OCTOPUS. For us to collect this information is simply due diligence.

To put the minds of colleagues with paranoid tendencies at ease, we will add a few sentences to the manuscript to reiterate the above regarding collecting and storing user info.

2. What's actually being reviewed here?

You note that "the paper is not so much describing a data set as a means of accessing previously published data", and so the review should not be limited to the data itself but also to the platform used to distribute it. This is partially correct.

As mentioned numerous times on the ESSD discussion forum, the CRN collections are not merely compilations of previously published values. Rather, all denudation rates have been recalculated such that the resulting collections are internally consistent and fully reproducible. The recalculation of the data is not a trivial undertaking and it is probably the main reason why compilations of CRN data (and even much less compilations that offer reproducible calculation and spatial context) do not abound in the literature. To reduce OCTOPUS to merely a means of accessing previously published data trivializes the real contribution that OCTOPUS is, and the amount of work that went into producing the data collections.

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

3.A. WFS/WMS functionality

In hindsight, we concede that it was a mistake not to provide more details in the manuscript about the WFS and WMS capability. Our rationale was the following:

(i) Those who know what OGC stands for, and what WFS and WMS are, will know what to do. As you mention in the review, it took you 5 minutes of looking around to figure out what capabilities are available and it was straight forward to find the relevant URLs.

Accessing the WFS capability via QGIS, for example, is trivial if one knows where the WFS button in on the QGIS GUI. The same applies to ArcGIS and WMS (as the latter does not do WFS).

(ii) Those who do not know what WFS and WMS are, will likely not be interested. Let's be honest: we are as excited about OGC capabilities as you are, but most of the users will want nothing more than an Excel spreadsheet or a CSV file to import into MATLAB.

(iii) There is a web interface to use for downloading data, and most users will likely prefer a local copy of the data to use with their favourite desktop GIS application, rather than connect via WFS.

To rectify our naïve mistake, we will add a section to the manuscript providing details of how to access data via WFS. This will be important to those users who wish to download data but are not interested in also downloading the raster files.

When reading a paper that also includes a MATLAB script or an R script (for example) to undertake the calculations presented, I do not expect the paper to provide me with a step by step tutorial on how to use MATLAB or R. Likewise, our aim here is not to write a tutorial on GeoServer and WFS. Rather, we will include three examples of WFS requests that those who wish to download large chunks of the database (and are not interested in the raster data) can adapt to get the desired result. We will also describe in brief QGIS's WFS capability as a way of browsing the data with the entire attribute table available.

3.B. How data were selected for inclusion

Our aim was/is to compile all data – published and unpublished – that is publicly available. It is not our role to decide on data quality. There are two points that need to be mentioned, however:

(i) we had instances where a publication did not provide sufficient information for the

data to be recalculated (e.g., insufficient info to be able to confidently locate and delineate drainage basins) plus corresponding author refused to cooperate. In such cases, we had no choice but exclude those data from OCTOPUS.

(ii) at some point we had to stop harvesting data from the literature and start recalculating using CAIRN so that we can release version 1 of OCTOPUS.

Given the above, there are studies that were excluded from the current release of the databases and we apologise to authors (and the reviewer) for omitting some data. This was not an editorial decision except in cases where we had no choice due to lack of information (we briefly mention about this on page 8, lines 9-10).

Following the reviewer's advice, we will add a sentence to Section 4 (after line 5, page 7) to clarify that the aim was/is to include all data that is available.

3.C. Laudatory remarks not related to factual description of the data set.

Although laudatory, we do not think that the statements on page 1 (lines 14-21) and page 10 (lines 15-23) are exaggerations.

Take for example the studies that we had to exclude from OCTOPUS because we could not identify the drainage basins from where the data was collected. These data are virtually useless beyond the scope of the papers in which they were published as no further analysis can be conducted on them by third parties.

Another example are the unpublished TL data included in OCTOPUS. The overwhelming majority of these were produced by the UOW TL laboratory (now closed) and as part of the OCTOPUS project we have digitised and curated data existent only in laboratory note books that are now part of landfill.

We will reappraise all laudatory statements and will ensure that chest beating is kept to acceptable and safe levels.

4. What about the product itself?

4.A. Backward set 1: why download data?

This is a philosophical question that is probably beyond the scope of our manuscript. For people to stop wanting a local copy of the data requires a radical shift in the current *modus operandi*.

Some reasons come to mind why downloading data is necessary:

(i) The recalculation of basin-wide denudation rates can only be achieved offline. Thus, to recalculate the data one needs a local copy of CAIRN and a local copy of the data.

(ii) Although most desktop GIS and even programs like MATLAB and R include some (although, very limited) WMS/WFS capability, and so in theory, users would not strictly require a local copy of the data to run the same analyses that they are currently running with data being offline. However, there would likely be performance issues (data that is offline can be accessed faster and there is no need for a network connection) and so running analyses on remote data would not be the preferred option. Building an entire workflow that relies on data that is stored remotely also requires an exit strategy in case the online system goes down. For most, that exit strategy will consist of having a local copy of the data.

(iii) Most users of CRN and OSL/TL data will likely not have the appetite to add another layer of complexity to their workflows (see (ii) above) and will prefer their data neatly in a spreadsheet saved safely onto their hard drives.

4.B. Backward step 2: trust but verify.

This is another rather philosophical point and we disagree with the reviewer that the inability to directly look at individual sample data via the web interface (as opposed to only seeing a subset of the database entries) will cause credibility issues. After all the full dataset is available for download and inspection. In fact, using QGIS's WFS

capability, the data can be inspected without download.

Having said the above, however, we acknowledge that changes can be made easily to any web interface to allow more granular access to the data without actually downloading a local copy. This is an aspect we will be exploring in the near future as we are expanding OCTOPUS to other data collections – however, this discussion is beyond the scope of the current manuscript.

4.C Backward step 3: why is data access hard and not easy?

See (1) Historical details and disclaimer. Further, we think that the web interface – no matter how feature poor – along with the WMS/WFS capability allow that versatility in data access that the reviewer is talking about. The latter does not necessarily have to be exploited by the OCTOPUS web interface. Individual users may develop aditional web interfaces and workflows that leverage the WMS/WFS capabilities to provide alternative front ends to the OCTOPUS collections.

5. Technical details.

Relational databases have been around for many decades and we know that our 'flat' database design has some performance and data redundancy related costs. However, we do not think that these are too severe – given the relatively small number of data points in all collections (we are talking about thousands of measurements rather than millions). Further, the flat design is what most users will prefer in the end, and so it makes life easier.

6. Minor detail regarding reproducibility of past calculations.

Shouldn't reproducibility be the ultimate goal in science?

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