Geochronology Discuss.,
https://doi.org/10.5194/gchron-2020-6-AC1, 2020

Interactive comment on "Technical note: A prototype transparent-middle-layer data management and analysis infrastructure for cosmogenic-nuclide exposure dating" *by* Greg Balco

Greg Balco

balcs@bgc.org

Received and published: 15 April 2020

This responds to the review comments by Sebastian Kreutzer. To summarize, this review is overall very supportive of the paper, but makes a number of minor suggestions to correct or clarify specific parts of the paper (review points 1, 3-5, 6, and 9), and also opens a discussion of several aspects of geochronology data management that are broadly related to the subject matter of the paper (points 2, 7,8, 10-14).

I very much thank Dr. Kreutzer for carefully reviewing the paper, and I am very happy

(perhaps too happy...the reviewer and everyone else may regret this by the time they get to the end of this response) to continue discussion on the topics he brings up.

I'll deal with the minor clarifications and corrections first. These include (i) a request to clarify the definition of "middle-layer" (review comment 1); (ii) a request for more discussion of how important changes to the middle-layer calculations are to the results of exposure-dating studies (comment 6); and (iii) several minor grammatical or style issues (comments 3,4,5). I agree with all these comments and will incorporate them into a revised manuscript. Finally, review comment 9 is a helpful suggestion for a minor change to the ICE-D website, but is unrelated to the text of the paper.

The remainder of the review does not propose changes to the paper, but instead includes a number of comments and questions about the ICE-D software infrastructure and highlights some broader implications of those questions for geochronology data management generally.

An important disclaimer for the following discussion of these broader comments is that (as noted in the 'competing interests' section of the manuscript) I am an editor of Geochronology. In this response, I am writing as a an author, not an editor, and my opinions on broader issues involving data management in geochronology should not be taken as any indication of journal editorial policy on these issues.

I think a good way of describing the context of the broader comments is that the review recognizes (and I agree) that there are a number of problems with how data management is currently handled in the field of geochronology, but this paper is only intended to address one of them. I agree completely with the points made in the review that (i) broader discussion of additional issues related to geochronology data management is valuable, but (ii) these issues are not pertinent enough to the main point of the paper to require specific revisions to the text.

I also appreciate and agree with the point at the beginning of the review that a peer-reviewed journal article might be the worst possible way to document a software sys-

tem. The whole point of a journal article is to be an archive of information that is carefully checked and reviewed, and does not change after publication – but any useful software must evolve continuously as errors are corrected and functions are added or improved. If a software description in a journal article does not become obsolete rapidly, then the software is probably not very useful. This is the reason that the paper focuses on a conceptual description of the key elements of the software infrastructure and not on detailed documentation, and I am glad that the review recognizes this. Regardless, in the following paragraphs I'll try to answer most of the questions about details of the software implementation and the project philosophy, and give some broader perspective on why certain design decisions were made. I'll start with what I think are the more important or interesting issues and then descend into the minor points.

First, review comments 7, 8, 11, and 12 are questions about exactly how certain aspects of the software infrastructure works, and how one can use certain functions. The answer to most of these questions is going to be something like "Well, yes, but that's not really finished yet," and reflects the fact that the ICE-D software infrastructure is nothing like a complete, professionally developed, production system. It's a prototype that has some design features that I think are important and relevant to geochronology data management. As the review recognizes, the purpose of the paper is to describe these conceptual aspects and not to document the software in detail. In fact, throughout the paper I have been quite careful to describe it as a 'prototype' as often as possible, and not give the impression that any new aspects of the software are being presented as complete production tools.

One common part of the answer to all these questions has to do with the security of the various APIs and online interfaces. All contributors to the ICE-D project so far are Earth scientists with some self-taught knowledge of computational science, but no professional training or experience in software development. Clearly this is good enough to produce something that works OK most of the time, but it is unlikely to be good enough to produce software that is secure against malicious misuse. Hacking or

misuse is a serious risk for this project not only because it could destroy or corrupt data, but also because the system uses cloud computing services whose cost scales with usage, so a malicious takeover of ICE-D servers could potentially be very expensive. Given these circumstances, it seems unwise to permit public viewing of the source code. As noted in the 'access to data, etc.' section of the paper, there has been no security audit of this code by a competent professional, so I have no way to know whether the source code reveals vulnerabilities that could be used to hijack or misuse project resources. Clearly this is a problem that needs to be fixed, but it will have to wait until the project has enough resources to support a serious security audit. Until that time, really the only option is to restrict access to the source code to trusted viewers.

So, on to review comment 7. The source code for the various elements is in Google Cloud Source repositories, but is not publicly viewable, for the reasons discussed in the previous paragraph. The database is a standard MySQL server and can be accessed using any suitable client, but the server is firewalled to enable only specific IP addresses, and anonymous login is not permitted. I am happy to provide access to anyone who wants to use the database for research purposes, and at the time of this writing have done so for about 25 colleagues, collaborators, and students.

On the other hand, the online exposure age calculator web service API is publicly accessible, but is not very well documented, with the exception of some brief documentation here:

https://cosmognosis.wordpress.com/2014/09/26/a-web-service-implementation-of-the-online-exposure-age-calculator/

As discussed in the paper, the whole point of the project is that you should be able to plug into the infrastructure at any point – raw data, calculated ages, etc. – and all these capabilities do exist in prototype form and are used by the ICE-D web server. That is, all the arrows on Figure 2 have some real existence. However, public access at all levels is not currently feasible for security reasons, and documentation is very weak.

These are just facts imposed by the limited resources available to the project. Building a working prototype is one thing, but documenting all the APIs in detail and making them robust against misuse is a significant task that will require additional resources.

Review comments 8, 11, and 12: most of the same applies. I completely agree with the position that all the software should, in principle, be developed as public open source projects using normal version control and bug tracking systems. However, as discussed above, as a practical matter I don't think it's possible to meet ideal expectations right now.

Review comment 11 (b) does bring up one interesting side issue. Seen from the perspective of the overall project design, careful versioning of middle-layer code to facilitate reconstructing past calculations is fundamentally not consistent with the basic concept of storing only observational data and performing all calculations dynamically. If the middle-layer code is updated, then by definition the results of calculations using the new code are better than the results of calculations using the old code – so if one accepts this model in its entirety, there would never be any reason to want to reproduce obsolete calculations. Why would you want to reproduce calculations you know to be wrong? That is off topic from the perspective of this review and response, but is something to think about.

One final point that is important to keep in mind for the discussion of openness is that a poorly documented and only partially open system may not be ideal from the broadest possible perspective of open data and open science, but may still be extremely useful in facilitating scientific progress in a particular field. The ICE-D project has mainly developed up to this point as a collaboration infrastructure used by a relatively small number of cosmogenic-nuclide geochemists. It is not funded by any public agency. Although of course it would be great if it were also more broadly useful as a means of public access to data, and we have used a fairly open development model so far, there is no public access mandate, or any inherent expectation of usefulness outside a fairly restricted group of researchers. The immediate goal of the project is simply to facilitate

synoptic research using exposure-age data by creating a modern data management infrastructure. From this perspective, poor openness and poor documentation are not ideal, but also not necessarily critical disadvantages. If there are only, say, 20 people using the system, it is no problem to create 20 separate logins to the MySQL server, distribute the source code to 20 people, and personally explain how to use the APIs 20 times. Scalability is not needed. If the system enables new and useful Earth science research for those 20 people, this is a success. This is a very important broader point for this discussion: incremental progress is valuable in itself, and it is not necessary to develop a system that does everything for everyone in order to make major progress in computational infrastructure in a field. To return to the overall theme of this response, the point of this paper is to propose that the transparent-middle-layer infrastructure model contributes to solving one important problem in geochronology data management. The additional problems of public access, open-source software development, and good documentation are important, but they are different. Solving one problem is not as good as solving all problems, but it is a lot better than solving zero problems.

Points 10 and 14 in the review then address another much broader aspect of the ICE-D project. Essentially, these comments ask what the rules are for how data are incorporated into the databases, and what the rules are for how data can be used. Before discussing this subject, I want to reiterate the point that these issues are not particularly related to the main point of the paper that focuses on the transparent-middle-layer concept, so this discussion is largely unrelated to the paper itself. That being clear, however, for me this subject is one of the most interesting aspects of the overall project.

Basically, there are no rules.

The databases have been built and managed by a group of collaborators who have expertise in the field and interest in using the database (they are named in the acknowledgements section of the paper and on the websites). There is no formal review process, and members of this group correct errors in the data as they are discovered, using their best judgement to make sure that the observational data are correct and up

to date. There is no formal review or validation procedure for data ingest. The premise of the project is that its user group is a collaborative group of researchers working in the same field who all have a personal interest in ensuring that the data are complete and correct, and this approach leads to a "trust-but-verify" philosophy that requires a degree of trust on the part of users that the data set is a correct representation of past work, but allows users the opportunity to verify (or not) that trust by offering complete, granular, public viewability of the data.

There are also no rules for how data in the databases can be used. This is by design. First of all, the majority of data are drawn from published work and are indexed to publications, so are subject to whatever prior restrictions apply to the publications anyway, and additional restrictions would not be meaningful. In the ICE-D:ANTARCTICA database, about 80% of the data have been published elsewhere. The other 20% are unpublished data, and there are no rules as to how these data can be used. Unquestionably, this is unusual, because nearly all scientific data management projects are extremely concerned about who gets credit for what, and how people who generate data maintain "control" of data they have generated. Most of these projects have a lot of rules. In contrast, this has not been a concern in developing the ICE-D project. In the example of ICE-D:ANTARCTICA, a number of researchers, myself included, have incorporated large amounts of unpublished data in the database, even in the absence of any rules about who can do what with these data, because the perceived value of being able to use the transparent-middle-layer features of the database to collaborate with others, interpret their data, and compare it with other data sets, outweighs any perceived risks. Many other data management projects, mainly those focused on data archiving rather than on analysis tools, have assumed that researchers will be more likely to contribute to community data sets if there are more rules. That assumption might be wrong: rules themselves, by adding complexity and codifying an attitude of mutual distrust within a research community, may be a disincentive to participation. It may be much more effective to forget about the rules, which are mostly unenforceable anyway, and focus on highlighting the advantages, rather than the risks, of engage-

ment with the project, making the incentives to participate strong enough to outweigh any disincentives.

Before this discussion travels too far from the main point, remember, again, that this subject has very little to do with the paper under review. The paper is about the application of a transparent-middle-layer software infrastructure to geochronology data management, not about the philosophy of who owns which data. For readers further interested in this subject, there is a much better articulated discussion of some of these aspects of the ICE-D project in a recent proposal that can be viewed here:

https://cosmognosis.wordpress.com/2020/02/25/computational-infrastructure-for-cosmogenic-nuclide-geochemistry/

Moving on to less significant points, review comment 13 is a question about the website user interface. The answer has to do with technical aspects of how the online exposure age calculators are typically used and, unfortunately, probably makes very little sense unless one is famiilar with them already. The text data string formatted for input to the online exposure age calculator is actually redundant on most of the pages served by the ICE-D web server, because the exposure age results that are commonly shown on the same pages are already the result of sending that data string to the online exposure age calculator. In other words, for most of the pathways of viewing data by site or sample, you don't need to cut and paste that text input string anywhere, because the web server already did that for you and is showing you the results. However, any user who is interested in computing exposure ages using an alternative production rate calibration would need the formatted input data. Thus, even though it's redundant, it appears on those pages to facilitate that workflow. Again, as noted above, the website has mostly been developed under the premise that the likely user group is already familiar with the exposure age calclator. Regarding the CREp formatting, at the moment CREp does not accept direct input of formatted text (or anything else) in an HTTP request, but instead requires uploading an Excel spreadsheet. Thus, going directly from the ICE-D web server to CREp is a coding pain in the neck that requires temporary file

writes and other overhead, and no one has invested the time needed to fix this problem yet.

And then, finally, review comment 2 calls attention to the "FAIR data principles." As background for readers who may not have heard of this, "FAIR" is an acronym intended to formalize and reinforce the common-sense approach that scientific (or any other) data should be "Findable, Accessible, Interoperable, and Reusable." In recent years this acronym has become popular in data management circles and associated with the Wilkinson et al. citation, although I am not sure if it originates there. From the perspective of this paper, certainly the transparent-middle-layer concept for managing and working with geochronology data has the potential to contribute to these goals: a centralized repository of observational data facilitates findability and accessibility; server implementations of both the database and the middle-layer calculations facilitate interoperability; and dynamic calculation of derived parameters from source data can enormously improve reuseability. On the other hand, there are several reasons that I did not mention this citation in the paper. The first one just relates to my earlier comment that there are lots of problems with data management in geochronology. This paper is only offering a solution to one of them, and makes no attempt to solve most of them; at no point does the paper claim that the ICE-D infrastructure is completely "FAIR." The second is perhaps more an issue of the philosophy of how citations should be used in scientific papers, but it seems to me that ascribing these basic concepts of information management, which probably date back centuries, to a 2016 paper is probably inappropriate. The third is getting fairly far off topic with regard to the present paper, and relates to the fact that in addition to establishing the acronym, the Wilkinson paper also articulates a set of "FAIR Principles" that include many prescriptive recommendations which go well beyond the basic ideas encapsulated in the acronym and seek to enforce a certain philosophy towards accessibility, metadata, and licensing. Although I agree with many of these recommendations in principle, my intention is to avoid endorsing them in the aggregate. In fact, I would argue that many of them are largely irrelevant in the context of a research tool primarily designed for use by

researchers in one field, and also that many of them may act to suppress, rather than advance, progress in data management by focusing on a theoretical ideal outcome rather than highlighting the value of incremental progress that may not be anywhere near the ideal, but is better than the status quo. To torture the common aphorism, I think judging improvements in data management against the "FAIR principles" often makes the perfect the enemy of the good, and geochronology data management is so far from perfect that we should be very happy about any incremental progress toward the merely good. This is a purely philosophical disagreement, and, again, I emphasize that my personal opinion of the "FAIR principles" is wildly off topic with respect to this review response – which also highlights that it is not very relevant to the paper and probably should not be brought up in the text. Following the disclaimer above, it is also important to make clear that my personal opinion of the "FAIR Principles" as expressed in this response should not be taken as any indication of journal editorial policy on this subject.

That's it. Again, I want to thank Dr. Kreutzer for his interest in this paper and his interest in discussing many of the issues raised in his review. However, although I agree that discussion of most of these issues is important in the broader context of geochronology data management, in this review process we should also keep in mind that the paper itself is intended to focus narrowly on only one issue.

Interactive comment on Geochronology Discuss., https://doi.org/10.5194/gchron-2020-6, 2020.