Geochronology Discuss.,
https://doi.org/10.5194/gchron-2020-6-RC1, 2020

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

Sebastian Kreutzer (Referee)

sebastian.kreutzer@aber.ac.uk

Received and published: 31 March 2020

Contribution summary

The contribution presents and describes a prototype cyber-infrastructure to manage and analyse cosmogenic-nuclide dates termed *ICE-D*. Elements are databases, a computation environment, and an HTML-based graphical user interface accessible with any state-of-the-art web browser. To date, the project is organised into three sub-projects,

pooling datasets of related geographical origin. The system relies on a modular design concept, the author called this 'middle-layer'. Raw measurement data are separated from model assumptions and calculation tools used to determine exposure ages. The system does not contain age values, but measurement parameters to calculate ages on the fly.

Recommendation

I suggest a publication of this manuscript in GeoChron after a discussion of, what I believe there are, minor points.

Justification

What the author summarised in a brief technical note is quite a piece of software design and software development work. The manuscript is clearly written, logical structured, and it was my pleasure to read it. However, I have to admit that it is always hard to review such a manuscript. Software tools are moving targets, and what happens tomorrow with such a project, whether it gets accepted or rejected by the community, depends on various aspects; some of them are out of the author's control. The good news is that the paper did not bother the reader with a lengthy tool description, but focusses on some design aspects. The presented conceptional ideas are clearly of relevance beyond the cosmogenic-nuclide dating community. To me, the essential question the manuscript raises is *"How meaningful are data repositories storing final (exposure) age results?"*. The author tries to avoid a broad, potentially offensive discussion, but attempts an answer to that question by concluding that *"[published] [...] dataset [...] are now obsolete."*. Since I do agree with that reasoning, and in general with the presented concept, this review allows me to pick on some details, I highlighted

while reading through the manuscript. Some of my comments the author may consider being more of major than minor nature, but I do not expect the author to rewrite the manuscript or go back to the workbench to refactor the software. What I want to start here is an open discussion, and some points may find their way into the revised manuscript.

Detailed comments

1. At first, I was struggling with the term 'middle-layer'. Finally, I thought ' yes', why not give the child a name. The author has chosen 'middle-layer' to express that particular calculation variable may evolve over time, and this may lead to different exposure ages. However, I would like to see a brief definition, somehow between line 16 and 24 on the first page, where the author says that "... therefore we decided to call 'middle-layer'...." or something similar. This would make it clear. Because what is presented here is not really new, it is something people did already in the past, and people do today when they align these ages (or at least they should do it).
2. At the end of section 1, I felt reminded to arguments already discussed in Wilkinson et al. (2016) who laid out the *The FAIR Guiding Principles for scientific data management and stewardship.* I think it would be good to add a few lines referring to their article. Because if data repositories and tools follow these guidelines, we would not have this problem.
3. Page 2, line 10: "...that a sample ... "
4. Page 2, line 12: "rocks" (?, plural instead of singular)
5. Page 2, line 17: I would prefer to see the chemical symbols instead of the names (the final decision is up to the author).

6. Page 3, first two lines: As written above, I am following your reasoning, however, as not being a member of the cosmogenic-nuclide community, I would like to see some numbers here, exemplifying the real effect. To which extent do exposure ages really change if specific parameters change? It does not need to be long, but it will make the manuscript stronger.
7. Page 4, your section 3: Maybe I've overlooked this aspect, but in this section, you talk about transparency and application interfaces etc. Where can I find the source code of this 'middle-layer' calculation (is it all part of the online calculator) in and how can I access the database without screencasting the webpage? If this is written already somewhere else, it should be repeated here. What I did expect from reading the manuscript, but before trying the webpage, was that the system comes with a dedicated API that allows other people to access the data the way they want it. This can even include the 'middle-layer', e.g., the API would enable using the middle layer calculator as a layer in between ... if wanted, if not, access should be possible directly to the database. After trying the webpage(s), I got the impression that this is not yet implemented, right? Please correct me if I am wrong and elaborate this a little bit in Sec. 3, because even it is a prototype, as I reader I want to get some idea about the project's future.
8. Page 5, Section 4: Here I have the same problem as above. Where can I see the source code, do I have an option to access those data directly? $>>$ Below I found the option, but I kept the comment here to show that this was a question that crossed my mind while reading the paragraph.
9. Page 6, Sec. 4.1: I do understand why the author splits the system, I am not really in favour of this decision, but I don't have to. What I was missing on the webpage was a link back to the landing page where I can see all the sub-projects.
10. Page 5, the social engineering aspects: How users can contribute datasets? Do they get reviewed or accepted without a review? I certainly like to see tall these

data available, but why a user should contribute data to the project and what happens with them they decide, maybe later, that they don't want to have them anymore in the system?

11. Page 8, lines 12–16. I asked this two times above (sorry, I saw it just now) and here it comes, the code is available on request only, and the tool development is non-transparent. This aligns with the announcement that the system will be *"continually updated"*. The author should reconsider both decisions.

    (a) These days agile software development goes hand in hand with rolling releases in conjunction with poor release versioning and an even less transparent recording of changes (if not developed via an open-source repository). Do we want to have this for scientific software tools? A proper and transparent (means visible to all users without having to place an inquire) bug and change tracking should be standard for scientific software solutions.

    (b) Proper and transparent versioning of tools has the advantage that users can refer to it. This helps to track (potential) differences in the calculation. The author argues that the strength of the proposed solution is that ages are dynamically calculated, but maybe they don't have to be (re-calculated). Perhaps they had been all calculated with the same version or even with a different version, but the differences were related to cosmetic issues only (but we would not know). I understand that the calculator is somehow disconnected from the database and shows version numbers in the output, but also here. If the development was hosted in some kind of *git* repository (or similar), changes would be easily understandable.

Additional thoughts

12. **Sustainability**: Recently, I read an interesting (non-scientific) blog entry about problems open-source software development projects face if the principal maintainer, here it appears to be the manuscript author, simply cannot continue the work. Who would take over? If everything would be available for download (or in a git repository), others could also help with the development and maintenance, and the system and all the work put in the past would sustain. After I saw that the webpage asks for donations to keep the service running (which is partly financed using private money from the author; my respect!), this issue the should think about.

13. **Usuability**: Is there a particular reason why I have to copy & paste the numbers to recalculate everything in the online calculator (either the exposure age calculator or *CREPp*, maybe this can be realised via sent button that fills the form?

14. Finally, I have two other questions that crossed my mind while testing the webpages (1) I could not find a proper Impressum, it would be good to add one since it will help to prevent legal problems and make clear who is running the page and for what kind of purpose. (2) Under what type of licence the datasets are published? No licence? Probably not, it seems that they can be used and re-used freely (there are in the database), but a licence should be clarified and added. Extrem case: Another group recycles the data and comes up with a completely new data interpretation based the conduced data mining. It should be clear from the start what kind of licence does apply if the data are available via *ICE-D*.

Conflict of interest

I have no conflict of interest to declare. I am not a co-author or otherwise a beneficiary of the suggested references to be cited.

Sebastian Kreutzer – Bordeaux – March 31, 2020

References

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., t Hoen, P.A.C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B., 2016. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3, 160018. doi:10.1038/sdata.2016.18

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