Review of – How FAIR is bioarchaeological data: with a particular emphasis on making archaeological science data reusable

Emma Karoune

General review statement:

I fully support the overall message of this article, which is advocating for the use of the FAIR data principles in Bioarchaeology, however, I think it needs considerable reworking and revisions to make it of publishable standard.

- The paper would benefit from being restructured – you have background sections after the materials and methods section. The background sections on FAIR and the intros to the different types of bioarchaeological data should be part of the introduction. Analysis section should be part of the discussion section.
- Methods section needs much more development - your methods of data collection, data processing and analysis need to be described in much more detail and your article would benefit from additional documentation around the methods such as a questionnaire document that showed all the questions.
- You are lacking many references relating to general archaeological, zooarchaeological and archaeobotanical data management.
- I also don't think that your data does support your analysis – I think you have overstated how well bioarchaeologists are making their data reusable.

Detailed feedback:

Title – clear title

Abstract – I would like to see you add some results into the abstract – can you give some number/percentages.

Keywords – add findable, accessible, interoperable, archaeology

1. **Introduction** – I think that you need to move the section called background and then sections 3, 4 and 4a into the introduction. Then have materials and methods section, then results.

Line 44 – ethical considerations – here you need to add some references and sentences about the CARE principles as this is relevant to your argument here.

- CARE References that could be added –
    - Carroll, S.R. et al. (2020) 'The CARE Principles for Indigenous Data Governance', Data Science Journal 19, p. 43. doi:10.5334/dsj-2020-043.
    - Carroll, S.R. et al. (2021) 'Operationalizing the CARE and FAIR Principles for Indigenous data futures', Scientific Data 8(1), p. 108. doi:10.1038/s41597-021-00892-0

- Marwick, B., Pham, T.S. & Ko, M.S. (2020b) ‘Over-research and ethics dumping in international archaeology | Nghiên cứu mang tính lối mòn và sự tha hóa về mặt đạo đức trong khảo cổ học quốc tế | ိ ိုင်ငံတကာေိရှုုးေိဟာင်ုးသုေိတသနပညာရပ်မှ မဆီေိလ ာင်ေိသာ သေု ိတသနမ ာုုးိငု ် မသငေ ် ိလ ာ ် ေိသာ က ငဝ ် တမ ် ာုုး’, SPAFA Journal, 4. doi:10.26721/spafajournal.v4i0.625.

The section from 37 to 48 is very similar to my own recent PCI article introduction section on why is reproducible archaeology important, so I would suggest also referencing that

- Karoune, E., and Plomp, E. (2022) Removing Barriers to Reproducible Research in Archaeology. Zenodo, ver. 5 peer-reviewed and recommended by Peer Community in Archaeology. https://doi.org/10.5281/zenodo.7320029.

Line 50 – reference main FAIR publications after the word FAIR as this is first time you mention it so da Silva Santos et al. 2016 and Wilkinson et al 2016. This should also be where you first explain FAIR so add findable, accessible, etc in brackets here.

Line 52 – change word thesis to paper or article.

2. Methods and Materials

This section needs much more detail so that readers can fully understand what you have done to collect the data, to process and analyse the data. Essentially, you want anyone to be able to reproduce your results and be able to replicate the method.

Line 58 – remove duplicate of methods and materials title.

Line 59-60 – Did you select people to send the survey to? If so, how were they selected?

Or did you use archaeology mailing lists? This process of getting participant’s for the survey needs to be explained more thoroughly.

Line 60-61 – ‘primarily in the UK, as well as in the rest of Europe, …’ – Your dataset also includes the rest of the world and I think you need to make sure that you mention this here as otherwise it seems like a very focused, or biased, towards first world countries survey.

Line 63 - Please make a document that has all of your survey questions as an appendix or put on zenodo for a doi and add ref on this line or 64.

Line 73-74 – ethical guidelines – put a reference for them here or provide an ethics form that you completed as reference.

Line 75-77 – I don’t think this is needed as it is not part of the methods.

You also need to add in to this section how you cleaned/processed the data.

Dataset – I don’t think this is the raw data as the question numbers are not sequential. If this is the case, where is the rest of the data?

You do say that some data is redacted for privacy but I can see names in the dataset so it is not clear to me how the redacting you have done keeps the data private??? Also bioarchaeology is a small

field and even people that you have redacted the names for, I can tell who they might be from their institution and specialism answers, so you need to fully redact all that information if you want an anonymous dataset.

If this is not the raw dataset, then please document what steps you have taken to get to this cleaned data set. And it is a good idea to name the dataset as cleaned so it is distinct from a raw dataset.

- Add analysis steps here too. Did you use statistical analysis?

Background – this section should be part of the introduction.

Line 88 – not just the researcher but also the domain has an influence on how the FAIR principles are/can be implemented.

Line 91 – give some examples of persistent identifiers such as DOI.

Line 113 – You have this one line on data management plans and then you refer to them later on in the discussion/conclusion – you need to expand this into a paragraph to describe what they are and how they link to FAIR and why you think there are needed for FAIR data.

Paragraph from line 115 to 127 – I agree with your broad definition of bioarchaeology but you have left out archaeobotany, so I think it needs mentioning in here somewhere. Especially as you have archaeobotanists in your dataset!

Then you have sections 3, 4 and 4a – I do think these need to be expanded to be more comprehensive of all relevant articles/info from each discipline.

Also, you do not intro zooarch or archaeobotany – it seems a bit odd not to have sections on them too and their relevant data management articles.

Zooarch references – zooarchaeology is fairly well advance in terms of data management – a few examples below.

– LeFebvre MJ, Brenskelle L, Wieczorek J, Kansa SW, Kansa EC, Wallis NJ, et al. (2019) ZooArchNet: Connecting zooarchaeological specimens to the biodiversity and archaeology data networks. PLoS ONE 14(4): e0215369. https://doi.org/10.1371/journal.pone.0215369

- Neusius, S., Styles, B., Peres, T., Walker, R., Crothers, G., Smith, B., & Colburn, M. (2019). Zooarchaeological Database Preservation, Multiscalar Data Integration, and the Collaboration of the Eastern Archaic Faunal Working Group. Advances in Archaeological Practice, 7(4), 409-422. Doi:10.1017/aap.2019.33
- Too Many Bones: Data Management and the NABONE Experience – http://dx.doi.org/10.5913/2017956.ch02

Archaeobotany –

- Lodwick, L., 2019. Sowing the Seeds of Future Research: Data Sharing, Citation and Reuse in Archaeobotany. *Open Quaternary*, 5(1), p.7.DOI: https://doi.org/10.5334/oq.62

- Karoune, E., 2022. Assessing Open Science Practices in Phytolith Research. *Open Quaternary*, 8(1), p.3.DOI: https://doi.org/10.5334/oq.88
- Kerfant, C., Ruiz-Pérez, J., García-Granero, J.J. *et al.* A dataset for assessing phytolith data for implementation of the FAIR data principles. *Sci Data* **10**, 479 (2023). https://doi.org/10.1038/s41597-023-02296-8
- Neotoma - https://www.neotomadb.org/about

Figure 2 – needs to have a reference unless you made that figure yourself and check the license to see if you can re-use it.

4. Stable isotopes – I think you need to add in IsoArcH – IsoArcH | Isotopic database for bioarcheology

- Neotoma - https://www.neotomadb.org/about

5. Results

You might want to consider a saturation analysis to show that you have collected enough data to be representative.

5a. – in figure 3 you have emails – so I assume this is direct emailing to participants?

- What is further dissemination?
- This is what you need to describe in your methods section as this diagram is not really understandable otherwise.
- Figure 4 – also needs more description – what are the institutions from other countries other than UK. Why are there only a few other world regions represented?
- The commercial, freelance, online responses would not really be included in the analysis as they are incorrect responses to the question.
- I would also put the questionnaire question number on each of the figure titles as this will link it to the dataset better.
- What are the other specialisms in figure 5?

Generally all of the figures need more description in the text, such as figure 10, which is very complex and only has 1 line to describe it.

There are probably too many figures though so you could just describe some of the results in more detail but not include a figure.

6. Analysis

This needs to be part of your discussion section.

I do not think that the analysis actually reflects your dataset.

Figure 31 – gives a very over estimation of how FAIR each discipline is and you need to reanalyse this. For example, you have aDNA 100% making their raw data accessible but I think you have based this on very few participants that said aDNA as their main specialism, some of these do not share raw data and there are multiple other participants who say they do aDNA that do not provide raw data.

Line 319 – you say that almost all specialisms make their data open access and ensure availability from raw data – this is not true from your dataset. Figure 14 shows 40% in the category of 'in published reports'. This data is really not in line with the FAIR data principles and therefore not accessible. These reports could be in close access articles or even in site reports that are only

accessible to few researchers. Data in reports is usually very poor quality for reuse, so not raw data and not in a file format that can be reused.

There is often a misunderstanding of what raw data is by researchers and I think you have this problem in your dataset. Figure 16 say 68% are making raw data available but then figure 24 has 53% saying that they publish raw data. Probably the people that are not using excel, csv, tsv files to publish their data are not publishing fully raw datasets. And it is very unlikely that people are publishing raw datasets in published reports as there just is not enough room to do so.

I do think you need to rework the analysis to reflect the data more accurately.

7. Discussion

7a. – You have not really asked about standardization in your survey as far as I can see but you have this section on it. So, can you present some results about the use of standard vocabularies/ontologies from the survey?

7c. I don't think your data supports the claim that data accessibility is relatively high – most in published reports does not mean they are accessible for reasons I have mentioned above.

8. Conclusion

Line 425 – you mention sample IDs here for the first time – it would be good to describe what these are earlier in the paper.

Line 429 – word thesis needs changing for paper or article.

References:

Put your data and code as references in your reference list and then add in text in methods section and data/code availability statement.

General archaeology references on data sharing/archiving that could be added:

- Marwick, B., & Birch, S. (2018). A Standard for the Scholarly Citation of Archaeological Data as an Incentive to Data Sharing. Advances in Archaeological Practice, 6(2), 125-143. doi:10.1017/aap.2018.3
- Huggett, J. (2018). Reuse Remix Recycle: Repurposing Archaeological Digital Data. Advances in Archaeological Practice, 6(2), 93-104. doi:10.1017/aap.2018.1
- Kansa, E. 2012. Openness and archaeology's information ecosystem. World Archaeology, 44(4): 498–520. DOI: https://doi.org/10.1080/00438243.2012.737575
- Kansa, EC and Kansa, SW. 2013. Open Archaeology: we all know that a 14 is a sheep: data publications and professionalism in archaeological communication. Journal of Eastern

Mediterranean Archaeology & Heritage Studies, 1(1): 88–97.
DOI: https://doi.org/10.5325/jeasmedarcherstu.1.1.0088

- Kansa, SW, Atici, L, Kansa, EC and Meadows, RH. 2020. Archaeological analysis in the information age: guidelines for maximising the reach, comprehensiveness and longevity of data. Advances in Archaeological Practice, 8(1): 40–52.
DOI: https://doi.org/10.1017/aap.2019.36
- Kansa, SW and Kansa, E. 2014. Data publishing and Archaeology's information ecosystem. Near Eastern Archaeology (NEA), 77(3): 223–227.
DOI: https://doi.org/10.7183/0002-7316.79.1.5