



OPEN

DATA DESCRIPTOR

A dataset for assessing phytolith data for implementation of the FAIR data principles

Céline Kerfant¹✉, Javier Ruiz-Pérez², Juan José García-Granero³, Carla Lancelotti^{1,4}, Marco Madella^{1,4} & Emma Karoune⁵

Phytolith research contributes to our understanding of plant-related studies such as plant use in archaeological contexts and past landscapes in palaeoecology. This multi-disciplinarity combined with the specificities of phytoliths themselves (multiplicity, redundancy, naming issues) produces a wide variety of methodologies. Combined with a lack of data sharing and transparency in published studies, it means data are hard to find and understand, and therefore difficult to reuse. This situation is challenging for phytolith researchers to collaborate from the same and different disciplines for improving methodologies and conducting meta-analyses. Implementing The FAIR Data principles (Findable, Accessible, Interoperable and Reusable) would improve transparency and accessibility for greater research data sustainability and reuse. This paper sets out the method used to conduct a FAIR assessment of existing phytolith data. We sampled and assessed 100 articles of phytolith research (2016–2020) in terms of the FAIR principles. The end goal of this project is to use the findings from this dataset to propose FAIR guidance for more sustainable publishing of data and research in phytolith studies.

Background and Summary

FAIR stands for Findable, Accessible, Interoperable and Reusable. The FAIR guiding principles¹ were developed to maximise the sustainability and reuse of data as a framework for data management and stewardship. FAIR aims to make data available whenever possible, and when it is not, metadata is required to allow a full understanding of the data. For example, if data cannot be openly accessible, metadata describes the data and therefore enables the findability, accessibility, interoperability and reusability of these datasets^{1,2}. Metadata plays an important role in each step of the FAIRification process, including both human- and machine-readable metadata^{3,4}. The FAIR principles can be applied to create FAIR standardised tools for self and community research assessment^{1,2,5}, which can be used to improve (meta)data quality and searchability⁵, as well as reproducibility⁶ of research (Fig. 1).

This dataset was developed as part of the ‘Increasing the FAIRness of phytolith data’ Project (commonly known as the FAIR Phytoliths Project)⁷, which is a community initiative to address the lack of data sharing and the difficulty of reusing data in phytolith research. Phytoliths are microscopic bodies of silica that develop within living plants from the uptake of groundwater^{8–11}. In the last two decades, phytolith research has evolved quickly and is used in a wide range of scientific disciplines including archaeobotany, palaeobotany, geology and plant physiology, among others^{12–19}.

However, this increased application of phytolith analysis in a variety of disciplines has resulted in a dispersion of terminologies and methods^{10,20–23}. Two nomenclatures (ICPN 1.0 and 2.0^{8,10}) exist to standardise phytolith naming but these are limited to certain morphotypes and taxonomic groups. The wide range of phytolith methodologies and applications, as well as the diverse background of phytolith researchers, makes collaborative work difficult and hampers the sharing and reusability of data. In fact, a recent study looking at open science

¹Universitat Pompeu Fabra, Department of Humanities, Barcelona, 08005, Spain. ²Texas A&M University, Department of Ecology and Conservation Biology, College Station, TX 77843, Texas, USA. ³HUMANE—Human Ecology and Archaeology, IMF, CSIC, 08001, Barcelona, Spain. ⁴ICREA, Catalan Institution for Research and Advanced Studies, Barcelona, 08010, Spain. ⁵Historic England, Investigative Science Team, Portsmouth, PO4 9LD, England. ✉e-mail: celineemmanuelle.kerfant@upf.edu

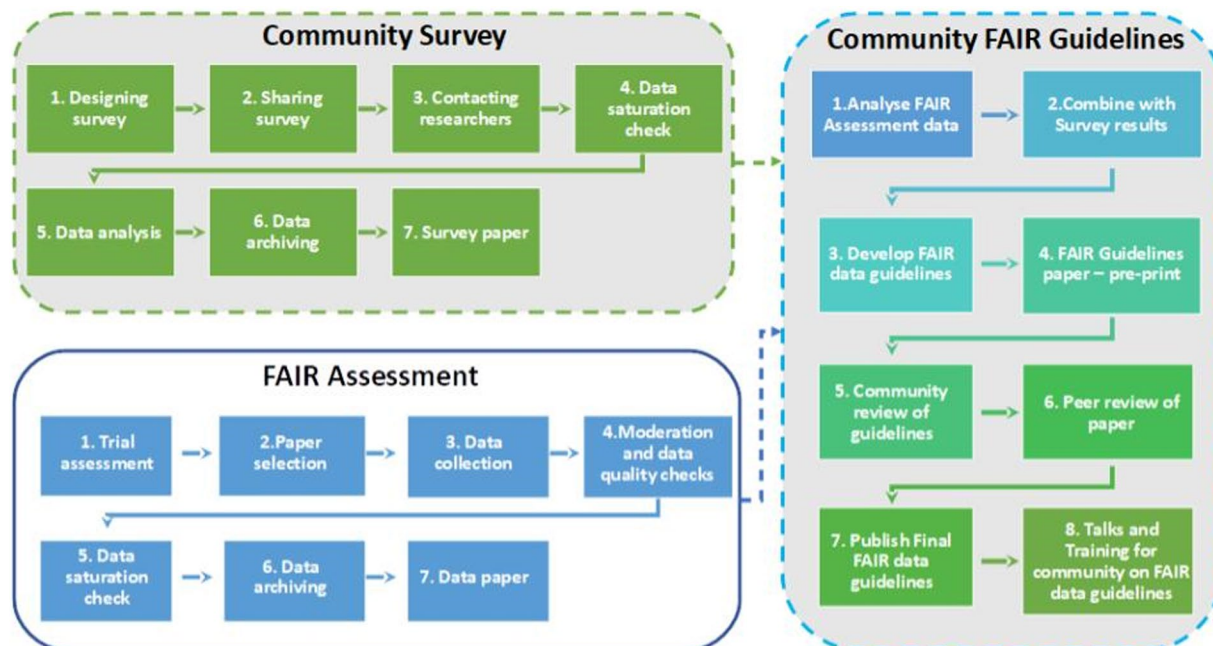


Fig. 1 FAIR Phytoliths project workflow - the FAIR assessment is the workflow for the data paper and will be combined with the survey paper results to produce the first draft of the Community guidelines. Dotted lines indicate work packages still to be completed on the FAIR Phytoliths Project.

practices in phytolith research revealed a clear lack of data sharing (53% shared data in any format) and data reusability (only 4%) from a sample of 341 published papers with primary phytolith data^{23,24}.

The FAIR Phytoliths Project focuses on examining published phytolith research in terms of how data sharing and management can be improved using the FAIR principles as a framework. Inspired by the wheel of FAIRification²⁵ and the need to define objects and variables to be fairified in the pre-FAIRifying stage, we have gathered a wide sample of phytolith research from different disciplines to assess the full extent of published phytolith data. We chose to take published research articles from studies conducted in two geographic areas, Europe and South America, and published between 2016 and 2020 (both included). These two regions were chosen because they have different traditions of phytolith methodologies (e.g., counting strategies, nomenclatures) and therefore a selection of papers from both regions was thought to give suitable breadth of published phytolith data.

Once we had selected the published articles, we scrutinised all aspects of the published material to highlight what elements were already in line with the FAIR principles and how other elements could be improved. The results of this dataset will be used to develop domain-specific FAIR guidelines for the phytolith community (with community review) and make a FAIR assessment tool to help phytolith researchers assess their own datasets and publications. Following this study, we will develop training materials for phytolith researchers so that a wide range of tools will be available to improve data sharing and collaboration among phytolith researchers.

Methods

The method for data collection used in the FAIR Phytoliths Project is a development of the method created by Karoune^{23,24} to assess open science practices within phytolith research. It is designed to provide a transparent record of the FAIR assessment dataset and aims to be reproducible. The full data set is available here <https://doi.org/10.5281/zenodo.7851930>²⁶.

Six phytolith researchers (Emma Karoune, Carla Lancelotti, Juan José García-Granero, Javier Ruiz-Pérez, Marco Madella, Céline Kerfant) conducted the whole study in two steps: A first team of five assessors carried out the trial of data collection methodology and a second team of four assessors conducted the main data collection. One of them did not take part in the trial assessment.

We have developed a transparent methodology in this study to ensure trust and a deeper understanding of this project's work within our community. This will hopefully generate greater uptake of the resulting guidelines and therefore initiate a movement to the desired improvement in data management and stewardship within the phytolith community.

The full description of the experimental design is grouped into two sections: trial of data collection methodology and data collection.

Trial of data collection methodology. An initial method for data collection was developed based on previous work by Karoune^{23,24}. A data categories table and a matching Google Form²⁶ were created. The terms *data categories* and *variables* are being used interchangeably in this article - the data dictionary has been standardised using the term variable but prior to this both terms will be used in this article.

The trial of data collection assessed the clarity of meaning for each variable before starting the FAIR assessment. The document containing the full details of the trial can be accessed through our repository²⁶. The methodology for the data collection was evaluated, firstly, to check if any data categories (variables) were missing; secondly, to make sure that we were capturing all of the information needed for assessing all elements of the FAIR data principles in the selected articles; and, lastly, to determine if any of the categories produced a wide range of results among the assessors.

During the trial process we linked the FAIR principles to the variables that are specific for phytolith research. In the trial, the variables were ordered in terms of the FAIR acronym²⁶.

The trial was conducted by five researchers. We each selected one article (five articles in total) from different world regions and different types of studies. We each assessed all five papers and used the Google Form to collect data. We then compared our results and discussed differences recorded for each article and any problems that we had with collecting data for each data category.

Data collection methodology. The first step was to conduct two literature searches to create a sample of publications from Europe and South America. We standardised how we determined what countries are in Europe and South America according to Wikipedia:

- List of European countries: (https://en.wikipedia.org/wiki/List_of_European_countries_by_area)
- List of South American countries: (https://en.wikipedia.org/wiki/List_of_South_American_countries_by_area)

Due to multiple members of our team collecting data, we needed to get a static list of papers rather than using online searches. We used Publish or Perish software (version 0.7.2) (<https://harzing.com/resources/publish-or-perish>) to run two keyword searches ('phytoliths' and 'Europe' or 'South America' with a date range of 2016 to 2020, both included) using the Google Scholar as search engine and then we exported the results as a .csv file. This produced a list of 989 articles for Europe and 996 for South America.

We selected the first 50 articles from both searches by manually looking at the articles to check they met our search criteria of being a phytolith article, in one of the chosen areas (Europe or South America) and a study with primary data on material. In the case of articles with two different publication dates (online publication date and final publication date), the online publication date was chosen.

Each article was then assessed using the prepared Google Form. We collected data in two spreadsheets - one for European publications and one for South America publications. Notes were put in a README file²⁶ (there is one README file for each spreadsheet with the date, data collectors name and what was done in each working session). This made it easier for us to work asynchronously.

Once all the articles were assessed the R package *roadoi* version 0.7.2 (uses Unpaywall with R) was used on our list of articles to find information about the open access status of the articles. The documentation for the data categories in this package can be found here: (<https://unpaywall.org/data-format>). This information was integrated into the data sets after the rest of the data was collected.

Finally, the datasets were merged to create the final dataset for assessing FAIR practices in existing phytolith publications.

Standardisation and transparency of data collection was vital in this study, as the data collection was being conducted by four assessors working in a distributed team. Specifically, we needed to avoid duplication and control subjectivity in the data collection. This was achieved by using README files for each dataset during the data collection, so that each data collector recorded their work during every session. The progress, issues and changes of the FAIR assessment datasets were recorded in the README files for each dataset and were discussed every two weeks during online meetings as well as in a dedicated Slack channel.

We also collected data using Google Forms whereby most of the variables displayed a drop down list of responses that had to be selected²⁶ and very few of the variables had open text boxes. This meant that there was little room for typing errors in the dataset and therefore less need for data cleaning before analysis. This also ensured that subjectivity in the answer was kept at a minimum and avoided the creation of many single-entry variables.

Data Records

The entire dataset is deposited in the following Zenodo repository²⁶, the structure of the data sheets is as follows.

Raw data files can be found in the project's Zenodo repository²⁶. This folder contains the two raw datasets for Europe and South America as .csv files and also the final combined dataset as a .csv file.

The Zenodo repository also contains a fully transparent record of data collection including methods development: data collection, data validation and data saturation checks²⁶.

During data collection we used a working data dictionary to help us with standardisation of our decision in line with guidance from the FAIR Cookbook (<https://fairplus.github.io/the-fair-cookbook>). This has been changed into a standardised data dictionary²⁶.

The first section of variables (1–7) gathers general information about each article such as journal name and full reference of the article. This section also includes the information about our sampling criteria, which comprises year of first publication online (between 2016 and 2020) and the location of the study in the article (either from an European or South American country, or countries, or regions using Geonames standardised list of countries or regions in each region). We also gathered information on time period (e.g., Holocene archaeological site, modern plant cultivation experiment) and type of study (e.g., archaeology, plant physiology). This was used at the analysis stage to group articles in different ways to look for any differences in alignment with the FAIR principles.

The second section (variables 8–11) captured information about open access. The use of open access publishing and type of journal in terms of openness may influence the way data is published as there are differences in publisher data policies such as in the policy of the journal *Vegetation History* and *Archaeobotany* authors are encouraged to deposit data that supports their findings, whereas in *Plos One* authors are required to make all data necessary to replicate study's findings publicly available without restriction²³.

The third section focused on the methods reported in the articles (variables 12–16). For data to be reusable, full and transparent reporting of methods is required. For phytolith research, the method of extraction needs to be reported fully and this information is captured in variable 12. The “counting method” (variable 13) assesses the level of replicability of this analytical step as this is a paramount detail for comparing datasets or considering merging datasets. We looked for information included in the articles about the number of phytoliths counted per slide, if those phytoliths were single-celled or multi-celled, if unidentified phytoliths were represented in the count and if any morphometric study was performed. If any of this information was not included or unclear, then we did not consider it as a replicable counting method.

We also collected information about the instruments used (variable 14). We looked specifically for two pieces of information: the general type of microscope (not the brand or model of microscope) and the magnification used to carry out the counting.

We collected data on the use of specific nomenclatures (variable 15), such as ICPN 1.0⁸ and ICPN 2.0¹⁰, but also other naming conventions developed and used locally by specialists within their own laboratories^{27–30} or by researchers creating ad-hoc reference collections^{31,32}. Specifically, variable 16 was used to collect data on whether the standard nomenclatures were being fully used - this meant if the articles showed no adaptations to the standard nomenclatures or these adaptations had been fully described. This was only done for those articles that had stated the use of either ICPN 1.0⁸ or ICPN 2.0¹⁰.

The last section (variables 17–25) aimed at collecting information on the data presented in or linked to the articles. We recorded where the data was located (variable 17) and, if in a repository, which repository (variable 18) was used, including institutional repository or long-term repository, among other possibilities.

The “type of data” (variable 19) and “data format” (variable 20) variables assess the data made available in or linked to the article for reuse. The only reusable data are raw data: in the case of phytolith research an editable file providing the sample weights (processing weights recorded during extraction of phytoliths from sediments or other materials) and the absolute counts of each morphotype in each of the samples. It is difficult to reuse processed and reorganised data as this limits the reuse potential of the data²³. In this dataset, we therefore recorded the most reusable data presented in the article. This means there could have been different types of data in tables or graphs but we only recorded the most reusable data presented. When only graphs were presented in the article, we collected this as “no data provided-only graph” due to the difficulty in extracting data from graphs.

The data availability statement (variable 21) allows other researchers to easily find data presented in articles. We checked if a data availability statement was provided and, if so, made sure that the statement did relate to the data. Some data can be available but under some specific conditions and each of them have to be separately assessed. For example, some datasets can be available on request (from authors, organisation or third party), a practice that hampers data reusability as data and metadata are not easily accessible³³. In other cases, data can be fully available but in different ways. They could be presented within the article, in a supplementary file or could be linked to the article with a DOI. All these possibilities were recorded.

The data licence (variable 22) needed to be assessed as it is the legal framework in which researchers will find indications about how to reuse the dataset and how to cite the dataset. Therefore, we recorded what kind of licence was specifically added to the data.

The presence or absence of phytolith microphotographs (variable 23) was also assessed. Illustrations are part of both ICPNs^{8,10} principles to describe phytolith morphotypes. This documentation aspect is important for confirming identifications of phytolith morphotypes and discussions of nomenclature standardisation. We recorded under three responses: none, only significant ones, and all. ‘Only significant ones’ were recorded when there were some photographs but not all the morphotypes found in the data.

The last data variable concerned the software used to carry out the statistical analysis (variable 24). We recorded any mentioned software used for statistical analysis in the paper. This is important to understand the types of analysis tools that are currently used on phytolith data and particularly the use of open source software that can aid reproducibility of research and reuse of data.

Technical Validation

At the start of data collection. We started data collection slowly and spent time discussing the first few papers assessed so that we were being consistent with our decisions as a group. The data category table²⁶ and the standardised version of this table - the data dictionary (see the supplementary table 1)²⁶ was a tool that helped us to maintain the same level of rigour throughout the assessment. The explanation column aimed at avoiding misinterpretation or overinterpretation of data categories by providing clear examples of the intended answers or particular aspects to consider when making decisions. Supplementary table 1 can also be found at: <https://github.com/open-phytoliths/FAIR-assessment-data-paper-documentation/blob/main/FAIR-assessment-final-documents/Data-dictionary-FAIR-assessment-final.csv>.

Moderation during data collection. After each assessor had collected several papers worth of data, a moderation exercise was conducted. One of our team members looked independently at the three papers assessed by others and compared answers with the data collected by the other team members. We discussed the findings as a group and went back to recheck where any issues arose or changes needed to be implemented.

We also implemented a system of moderation for further data quality checks. After doing the data collection on each paper, if we were unsure about any of the data collected we made a comment in the readme file and highlighted the comments in yellow. This meant that the paper needed moderation by another team member.

A moderation meeting was conducted near the end of the data collection period to discuss the problematic papers and any data categories that had been hard to assess. Any changes agreed upon were then implemented on the whole dataset.

All discussions about the problems encountered during the evaluation can be found in this document in our Zenodo repository²⁶. After manual data collection, we used the *roadoi* R package to capture the information for data categories 8 (open access) and 10 (What repository for green access article). We decided not to collect data for category 11 (Is it a signed up repo/ ResearchGate and Academia categories) using R as the *roadoi* package does not take into account academic social networking sites (ASNS) such as ResearchGate and Academia. Some information on academic social networks was entered in this data category manually but not for every article in our sample. These types of academic social networking sites are regarded as places to upload data and research articles by some researchers, but they do not comply with long-term repository requirements, among other responsible science requirements^{34,35}. The results captured by using the *roadoi* R Package are presented in two files, one for each region, in their respective folders and the code is also available²⁶.

Data saturation. Once 100 papers had been assessed, data saturation was tested. A saturation method³⁶ that assesses the thematic saturation of qualitative data was used to evaluate the two datasets separately (Europe and South America) in terms of representativity and saturation qualities.

The representativity was ensured by an adequate redundancy of the journal's name.

The journal name, year (first published online), type of study, period/date and geographic location were the categories used to evaluate the data saturation.

The data saturation assessment showed we had collected enough data from the 100 papers to cease our data collection²⁶.

Usage Notes

The transparent recording of this dataset means that the methods used can be reproduced or replicated by others that want to conduct a FAIR assessment of the data in their own research community. Our data collection methodology could be used directly to look at phytolith data from different geographic areas of the world or easily adapted for use in related disciplines. There are other communities that have limited data sharing, especially in the use of open repositories^{33,37}, and our methodology of assessing FAIR data would transfer well to these communities.

We also hope that our methodology can be reused by other related disciplines to undertake similar FAIR assessments for the improvement of data management and stewardship. Similar results for data sharing to those found for phytolith research by Karoun²³ have also been found for another archaeobotanical discipline (macrobotanical remains)³⁷, demonstrating that phytolith research is not alone in the need for change in this particular area.

We included articles from a wide range of phytolith study types and disciplines in this dataset, therefore data is not only related to archaeology and palaeoecological disciplines. This dataset can be reused by non-archaeologist phytolith researchers and help us to fill the gap between our approaches and methods.

This dataset can also be reused for teaching activities related to reproducibility and transparent recording as each step of our study is recorded and can be reused independently for different training courses about responsible science.

Full description of the Zenodo repository files. *FAIR assessment final documents.* This folder contains the methodological files for our data collection, readme files, data collection forms and code used to collect data.

- Data-dictionary-FAIR-assessment-final.csv
- FAIR assessment data categories table
- Observations made during data assessment table
- Saturation and representativity of data collected.pdf
- Specifications table (Detailed overview of the study design).pdf
- README files
 - README file for European dataset
 - README file for South American dataset
- Data collection forms
 - FAIR Phytoliths Data Assessment Form – Europe
 - FAIR Phytoliths Data Assessment – South America
- *Roadoi* R package documentation
 - *roadoi* R package_Europe.csv
 - *roadoi* R package_South_America.csv
 - *roadoi* R package_Europe_code.md
 - *roadoi* R package_South_America_code.md

Subject	Phytolith research
Specific subject areas	Archaeology/Ecology/Geochemistry/Methodology (related to phytoliths)/Palaeoecology/Palaeontology/Plant physiology
Dataset description	This dataset contains information collected from 100 phytolith articles to inform guidelines for implementing the FAIR data principles in phytolith research - 50 selected from European studies and 50 selected from South American studies, over a 5 year publication time period (2016–2020).
Data format	Primary and secondary data
Type of data	Qualitative data
How data were acquired	Data were collected by six assessors in total (Emma Karoune, Carla Lancelotti, Juan José García-Granero, Javier Ruiz-Pérez, Marco Madella, Céline Kerfant). A trial of FAIR variables was conducted on five articles to check the criteria relevance by five assessors while the main data collection was carried out by four assessors. Access to the literature was carried out using <i>Publish or Perish</i> software (Harzing, A.W. (2007) <i>Publish or Perish, available</i>) to conduct two searches using the decided criteria ('phytoliths' and 'Europe' or 'South America') to produce a static list of papers. The first 100 articles selected from these lists using the inclusion criteria ('primary data', 'phytoliths' and 'Europe' or 'South America') were assessed on the basis of the availability and format of data and metadata presented. The R package <i>roadoi</i> (version 0.7.2) was used to transparently assess the accessibility of the articles. The languages of most of the studies are English and Spanish.
Open science tools used	<i>Publish or Perish</i> software (Harzing, A.W. (2007) <i>Publish or Perish, available</i>) - https://harzing.com/resources/publish-or-perish <i>roadoi</i> R package (version 0.7.2) - https://github.com/ropensci/roadoi
Data availability	Zenodo - https://doi.org/10.5281/zenodo.7215571 GitHub repository - https://github.com/open-phytoliths/FAIR-assessment-data-paper-documentation Data is available under the Creative Commons Attribution 4.0 International licence

Table 1. Specifications table (overview of the study design).

FAIR assessment trial. This folder contains information about the trial we conducted prior to starting our main data collection phase.

- Data collection methodology development document
- Data collection form for trial
- Table summarising how the FAIR principles were linked to a practical set of questions shaped specifically for phytolith research.pdf

data-raw
roadoi R package_South_America_code.md

- Contains <https://doi.org/10.5281/zenodo.7851930/data> raw
- Also contains the two separate raw datasets for Europe and South America as .csv files.

data-search
This folder contains the files of the results of the searches we conducted to produce lists of articles.

- contains .csv files of the searches conducted using *publish or perish* to get the list of relevant articles for Europe and South America articles.

paper-tables

This folder contains the tables in the article and as supplementary files.

- Table-1-Specifications-table.pdf
- Supplementary-Table-1-Data-dictionary-FAIR-assessment.pdf

Code availability

All code and methods documentation for this study can be found at the Zenodo repository²⁶.

It can also be accessed through this link to our GitHub repository: <https://github.com/open-phytoliths/FAIR-assessment-data-paper-documentation>.

Received: 20 October 2022; Accepted: 7 June 2023;

Published online: 21 July 2023

References

1. Wilkinson, M. D. *et al.* The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* **3**, 160018, <https://doi.org/10.1038/sdata.2016.18> (2016).
2. Devaraju, A. *et al.* From Conceptualization to implementation: FAIR assessment of research data objects. *Data Sci. J.* **20**, <https://doi.org/10.5334/dsj-2021-004> (2021).
3. Gonçalves, R. S. & Musen, M. A. The variable quality of metadata about biological samples used in biomedical experiments. *Sci. Data* **6**, 190021, <https://doi.org/10.1038/sdata.2019.21> (2019).

4. Wilkinson, M. D. *et al.* Evaluating FAIR maturity through a scalable, automated, community-governed framework. *Sci. Data* **6**, 174, <https://doi.org/10.1038/s41597-019-0184-5> (2019).
5. Wilkinson, M. D. *et al.* A design framework and exemplar metrics for FAIRness. *Sci. Data* **5**, 180118, <https://doi.org/10.1038/sdata.2018.118> (2018).
6. Karoune, E. & Plomp, E. 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> (2022).
7. Karoune, E. *et al.* FAIR phytoliths project archive april 2022 (v1.0). *Zenodo* <https://doi.org/10.5281/zenodo.6435441> (2022).
8. Madella, M., Alexandre, A. & Ball, T. International code for phytolith nomenclature 1.0. *Ann. Bot.* **96**, 253–260, <https://doi.org/10.1093/aob/mci172> (2005).
9. Hodson, M. J. The development of phytoliths in plants and its influence on their chemistry and isotopic composition. Implications for palaeoecology and archaeology. *J. Archaeol. Sci.* **68**, 62–69, <https://doi.org/10.1016/j.jas.2015.09.002> (2016).
10. International Committee for Phytolith Taxonomy (ICPT), Neumann, K. *et al.* International code for phytolith nomenclature (ICPN) 2.0. *Ann. Bot.* **124**, 189–199, <https://doi.org/10.1093/aob/mcz064> (2019).
11. Rashid, I., Mir, S. H., Zurro, D., Dar, R. A. & Reshi, Z. A. Phytoliths as proxies of the past. *Earth-Sci. Rev.* **194**, 234–250, <https://doi.org/10.1016/j.earscirev.2019.05.005> (2019).
12. Carnelli, A. L., Theurillat, J.-P. & Madella, M. Phytolith types and type-frequencies in subalpine–alpine plant species of the European Alps. *Rev. Palaeobot. Palyno.* **129**, 39–65, <https://doi.org/10.1016/j.revpalbo.2003.11.002> (2004).
13. Fernandez Honaine, M., Zucol, A. F. & Osterrieth, M. L. Phytolith assemblages and systematic associations in grassland species of the south-eastern Pampean plains, Argentina. *Ann. Bot.* **98**, 1155–1165, <https://doi.org/10.1093/aob/mcl207> (2006).
14. Sandweiss, D. H. Small is big: The microfossil perspective on human–plant interaction. *Proc. Natl. Acad. Sci. USA* **104**, 3021–3022, <https://doi.org/10.1073/pnas.0700225104> (2007).
15. Shipp, J., Rosen, A. & Lubell, D. Phytolith evidence of mid-Holocene Caspian subsistence economies in north Africa. *Holocene* **23**, 833–840 (2013).
16. Calegari, M. R. *et al.* Potential of soil phytoliths, organic matter and carbon isotopes for small-scale differentiation of tropical rainforest vegetation: A pilot study from the campos nativos of the Atlantic forest in Espírito Santo State (Brazil). *Quatern. Int.* **437**, 156–164, <https://doi.org/10.1016/j.quaint.2016.01.023> (2017).
17. Strömberg, C. A. E., Di Stilio, V. S. & Song, Z. Functions of phytoliths in vascular plants: an evolutionary perspective. *Funct. Ecol.* **30**, 1286–1297, <https://doi.org/10.1111/1365-2435.12692> (2016).
18. García-Granero, J. J., Lancelotti, C. & Madella, M. A methodological approach to the study of microbotanical remains from grinding stones: a case study in northern Gujarat (India). *Veget. Hist. Archaeobot.* **26**, 43–57, <https://doi.org/10.1007/s00334-016-0557-z> (2017).
19. Lentfer, C. J., Crowther, A. & Green, R. C. The question of early Lapita settlements in remote Oceania and reliance on horticulture revisited: new evidence from plant microfossil studies at Reef/Santa Cruz, south-east Solomon islands. *Tech. Rep. Aust. Mus.* **34**, 87–106, <https://doi.org/10.3853/j.1835-4211.34.2021.1745> (2021).
20. Shillito, L.-M. Grains of truth or transparent blindfolds? A review of current debates in archaeological phytolith analysis. *Veget. Hist. Archaeobot.* **22**, 71–82, <https://doi.org/10.1007/s00334-011-0341-z> (2013).
21. Zurro, D., García-Granero, J. J., Lancelotti, C. & Madella, M. Directions in current and future phytolith research. *J. Archaeol. Sci.* **68**, 112–117, <https://doi.org/10.1016/j.jas.2015.11.014> (2016).
22. Strömberg, C.A.E., Dunn, R.E., Crifó, C., Harris, E.B. in *Methods in Paleocology. Vertebrate Paleobiology and Paleoanthropology* (eds Croft, D., Su, D., Simpson, S.) Ch. 12 (Springer Cham, 2018).
23. Karoune, E. Assessing open science practices in phytolith research. *Open Quat.* **8**, <https://doi.org/10.5334/oq.88> (2022).
24. Karoune, E. Data from “assessing open science practices in phytolith research”. *JOAD* **8**, <https://doi.org/10.5334/joad.67> (2020).
25. David, R. *et al.* FAIRness literacy: The Achilles’ heel of applying FAIR principles. *CODATA* **19**, 1–11, <https://doi.org/10.5334/dsj-2020-032> (2020).
26. Kerfant, C. *et al.* Data from ‘A dataset for assessing phytolith data for implementation of the FAIR data principles’ (2.1.0). *Zenodo* <https://doi.org/10.5281/zenodo.7851930> (2022).
27. Iriarte, J. Assessing the feasibility of identifying maize through the analysis of cross-shaped size and three-dimensional morphology of phytoliths in the grasslands of southeastern South America. *J. Archaeol. Sci.* **30**(9), 1085–1094, [https://doi.org/10.1016/S0305-4403\(02\)00164-4](https://doi.org/10.1016/S0305-4403(02)00164-4) (2003).
28. Golyeva, A. Biomorphic analysis as a part of soil morphological investigations. *Catena* **43**(3), 217–230, [https://doi.org/10.1016/S0341-8162\(00\)00165-X](https://doi.org/10.1016/S0341-8162(00)00165-X) (2001).
29. Golyeva, A. in: *Plants, People and Places. Recent Studies in Phytolith Analysis* (ed. Madella, M.) Ch. 17 (Oxbow Books, 2007).
30. Henry, A. G. & Piperno, D. R. Using plant microfossils from dental calculus to recover human diet: a case study from Tell al-Raqāʿi, Syria. *J. Archaeol. Sci.* **35**(7), 1943–1950, <https://doi.org/10.1016/j.jas.2007.12.005> (2008).
31. Novello, A. *et al.* Phytolith signal of aquatic plants and soils in Chad, Central Africa. *Rev. Palaeobot. Palyno.* **178**, 43–58, <https://doi.org/10.1016/j.revpalbo.2012.03.010> (2012).
32. Watling, J. *et al.* Differentiation of neotropical ecosystems by modern soil phytolith assemblages and its implications for palaeoenvironmental and archaeological reconstructions II: Southwestern Amazonian forests. *Rev. Palaeobot. Palyno.* **226**, 30–43, <https://doi.org/10.1016/j.revpalbo.2015.12.002> (2016).
33. Tedersoo, L. *et al.* Data sharing practices and data availability upon request differ across scientific disciplines. *Sci. Data* **8**, 192, <https://doi.org/10.1038/s41597-021-00981-0> (2021).
34. Bhardwaj, R. K. Academic social networking sites: Comparative analysis of ResearchGate, Academia.edu, Mendeley and Zotero. *ILS* **118**, 298–316, <https://doi.org/10.1108/ILS-03-2017-0012> (2017).
35. Manca, S. ResearchGate and Academia.edu as networked socio-technical systems for scholarly communication: a literature review. *RLT* **26**, <https://doi.org/10.25304/rlt.v26.2008> (2018).
36. Guest, G., Namey, E. & Chen, M. A simple method to assess and report thematic saturation in qualitative research. *PLoS ONE* **15**, e0232076, <https://doi.org/10.1371/journal.pone.0232076> (2020).
37. Lodwick, L. Sowing the Seeds of Future Research: Data Sharing, Citation and Reuse in Archaeobotany. *Open Quat.* **5**, <https://doi.org/10.5334/oq.62> (2019).

Acknowledgements

We would like to thank EOSC-life for funding this project and for their support and training in the FAIR data principles. This project has received funding through EOSC-life from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 824087. The FAIR Phytoliths Project is based at two institutions: Universitat Pompeu Fabra and Historic England. We would like to thank both these institutions for their support and also financial commitment to this project. Thank you also for in-kind support from Texas A&M University and the Spanish National Research Council for time on this project by Javier Ruiz-Pérez and Juan José García-Granero. Ruiz-Pérez was supported by the National Science Foundation under award number DEB-1931232 to J. W. Veldman. García-Granero acknowledges funding from the Spanish Ministry of

Science, Innovation and Universities (Grant No. IJC2018-035161-I). Carla Lancelotti acknowledges funding from the European Research Council (ERC-Stg 759800). Celine Kerfant is a recipient of Juan de la Cierva-Formación contract FJC2021-046496-I funded by MCIN/AEI/10.13039/501100011033 and by the European Union “Next Generation EU”/PRTR. We would also like to thank the reviewers of this article that helped to improve and refine the content.

Author contributions

Conceptualisation and methodology: E.K., C.L., J.R.-P., J.J.G.-G. and M.M. were responsible for the conceptualisation of the project. Investigation: E.K., C.L., J.R.-P., J.J.G.-G., M.M. carried out the trial before the data acquisition. E.K., C.K., J.R.-P., J.J.G.-G. collected the main data. Formal analysis and validation: J.R.-P. performed the R *roadoi* package analysis and CL carried out the data saturation analysis. Funding acquisition: E.K., C.L., J.R.-P., J.J.G.-G. and M.M. Writing: E.K. and C.K. wrote the initial draft of the manuscript with contributions from C.L., J.R.-P., J.J.G.-G., M.M. All authors read and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to C.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023