FAIR Phytoliths - Data Collection Methodology Development

Conducted by:

- Emma Karoune
- Carla Lancelotti
- Juan José García-Granero
- Javier Ruiz-Pérez
- Marco Madella

Questions:

How FAIR are existing datasets?

- How easy is data to find?
    - Is it in the paper/supplementary files/repository?
    - Is there a data availability statement?
        - What are the policies for these journals on data availability statements?
- Is data accessible?
    - article access - where is the data, what file format?
    - repositories used for articles and data.
- How interoperable is the data?
    - Nomenclature - what used and how is it used?
    - Confirming use of nomenclature by providing pictures
- Is the data reusable?
    - Is there a license on the data?
    - Type of data
    - Format of data
    - Methodological information provided - processing, counting method, instruments, stats.

Trial of data categories

Everyone is to select one article to try using the data categories:

- Are there any data categories that are too subjective?
- Are there any data categories missing?
- Are we capturing the information we need for FAIR? - We want this to be useable at the end. FAIR is not about open data but making data available through clear metadata, doi's, licensing and the process for obtaining data being clear.

Articles used in the trial: - the aim was to try to get a mix of articles that gave geographical and type of study spread.

Delhon, C., Binder, D., Verdin, P., Mazuy, A., 2020. Phytoliths as a seasonality indicator? The example of the Neolithic site of Pendimoun, south-eastern France. Veget Hist Archaeobot 29, 229–240. https://doi.org/10.1007/s00334-019-00739-0

Ge, Y., Lu, H., Wang, C., Gao, X., 2020. Phytoliths in selected broad-leaved trees in China. Sci Rep 10, 15577. https://doi.org/10.1038/s41598-020-72547-w

Kaczorek, D., Puppe, D., Busse, J., Sommer, M., 2019. Effects of phytolith distribution and characteristics on extractable silicon fractions in soils under different vegetation – An exploratory study on loess. Geoderma 356, 113917. https://doi.org/10.1016/j.geoderma.2019.113917

Premathilake, R., Akhilesh, K., Anupama, K., Pappu, S., Prasad, S., Gunnell, Y., Orukaimani, G., 2017. Phytoliths as indicators of Quaternary vegetation at the Paleolithic site of Attirampakkam, India. Journal of Archaeological Science: Reports 14, 479–499. https://doi.org/10.1016/j.jasrep.2017.06.013

Watling, J., Shock, M.P., Mongeló, G.Z., Almeida, F.O., Kater, T., De Oliveira, P.E., Neves, E.G., 2018. Direct archaeological evidence for Southwestern Amazonia as an early plant domestication and food production centre. PLoS ONE 13, e0199868. https://doi.org/10.1371/journal.pone.0199868

Learnings from trial:

- Making click down lists of standardised answers for standardising our own responses.
- Include full reference for article and get rid of title and DOI categories.
- Period/date - need to simplify collection of this - just add what is stated at the start of the article - this is usually in the abstract. Use modern for non-archaeological and palaeoecologial papers.
- Type of study - choose the most prominent as there could be several focuses of the study.
- Structure of questions is hard to collect answers - need to reorder the questions to match where they come in articles to make data collection easier.
- We are going to use R package to get information about open access status.
- Make a list of criteria for question about use of ICPN 1.0 and 2.0 - common misuse for each to help us detect this.
- We need to start slowly and calibrate ourselves so we are collecting data consistently. We will work together to start with and moderate after we have done a few papers each.

Suggested data collection method for FAIR assessment of phytolith articles from Europe and South America:

1. Google scholar search - using publish or perish

- Use term phytolith - Europe or South America
- Refine date range - 2020 to 2016 (5 years).
- Download csv.file into folder.
- Download the first 50 articles that meet our criteria - primary data, in date range 2016-2020 first published online, in region of study as defined by list of countries on wikipedia lists for Europe and South america.
- Record data using google form to generate google sheet.
- Check for data saturation - if not met, collect more article, 1 year at a time using the same search criteria as above.
- Possible additional steps:
  - *Make data into csv file from each article - keep private until agreement from authors (if not open access)*
  - *Send email to author about FAIRifying their data.*

2. Method we are not using - select specific journals:

- Use journal website to do search
- Phytolith - Europe and Phytolith - South America
- Refine date range - 2020 to 2016 (5 years).
- Download file into folder.
- Record data using google form to generate google sheet.
- Check for data saturation - if not met, collect more article, 1 year at a time using the same search criteria as above.
- Possible additional steps:
  - *Make data into csv file from each article - keep private until agreement from authors (if not open access)*
  - *Send email to the author about FAIRifying their data.*

Could also find all articles, book chapters and grey literature in a time period - but probably too difficult to find some of this.

Data analysis plan

Aim: create a pipeline from raw data to data visualisation for reproducible workflow.

For each dataset, once data collection has finished:

1. Save raw data files as csv in data folder - then do not alter these files (file names to be date_fairphytos-samerica.rawdata.csv and date_fairphytos-europe.rawdata.csv) - date should be like this example 2021-12-01.
2. Use Openrefine to clean the data - save cleaned data csv in data folder (file names to be date_fairphytos-samerica.cleandata.csv and date_fairphytos-europe.cleandata.csv).
3. Import cleaned data csv to Rstudio - link project to Github repo - FAIRphytos-data-management

4. Transform data tables
5. Do statistics/calcs
6. Data visualisation
7. BINDER