Review by Robert Bischoff

This paper describes a database of lapidary artifacts from the Ceramic period of the Caribbean islands. A brief background of the setting and cultural history is provided, as well as a discussion on prior lapidary research. The bulk of the paper is a description of the data collection methods, database structure, and repository locations. The description contains the necessary metadata to interpret and access the database (URLs are provided), including caveats regarding data quality and limitations. Many archaeologists and researchers in other disciplines have called for more open-access databases with appropriate metadata. This article and database answer that call and provide a large database with location data that should be useful for researchers in the Caribbean as well as researchers and educators looking for high-quality datasets. I would prefer to see a short case study demonstrating the potential this database has for providing new archaeological interpretations; however, the article as written merits recommendation with the addition of some minor revisions to the article and by addressing a few problems with the database. The strength of this paper is its clear description of the different elements of the database and how they fit together with enough detail to confidently use the database to address research questions. The tables and figures are helpful and of good quality. The inclusion of the Rmarkdown document is an excellent example of open source science, and I had no trouble reproducing it. This will be useful for anyone looking to adopt a similar format for their own work. The database itself is of good quality compared to many archaeological databases and contains a number of useful fields. The location data does not have any obvious errors. There are a few important issues that need to be addressed prior to a recommendation.

The Islands table discussed in the text is not included in the OSF repository and has not been reviewed.

Thank you for pointing that, we indeed did not integrate this table. We agree with the reviewer that this is a mistake, and we thus included it in this new version. It is now described in the manuscript and is available from the OSF repository.

I used the tables in the OSF repository to examine the database. I was not able to join all of the sites to the beads tables in English or French. I also tried using the Index*Site field to join to the first part of the Index*B field (I removed the three-digit numbers at the end of Index*B), but I found a similar problem to just using the Site field. This should be fixed prior to a recommendation. I suggest including the Index*Site field in the BEADS table as it will be easier to avoid problems with typos and will be a more stable key for joining tables. There are three rows in the BEADS and PERLES tables that are exact duplicates and should be removed (GR-01-044, GR-01-099, and GR-01-128). There are also several rows that are included in the BEADS table but missing in the French version called PERLES.

These problems, they have been solved. The rows from a few sites in French Guyana were remaining in the English version of the table BEADS while we do not want to let them in this version of the database since the inventory on the continent is very preliminary. This explains the discrepancy between the number of rows of the BEADS and PERLES tables, as well as the problems of joining the SITES_EN and BEADS tables. Also, some artifacts found in the storage *of some museums unfortunately do not have any known site of origin. These artifacts thus have a Index_B like GD-00-001 but this GD-00 does not have its row in the SITES_EN table.*

Some of the terms in the English version have not been translated from French. For example, the same field has entries for “Bead-Pendant” and “Bead-pendentif.” This will cause problems aggregating the data.

Thank you for pointing this, scattered absences of capital letters created problems in the automatic translation script. We worked on this and hope there is no more issues.

There is one small problem I identified in the Rmarkdown document. The size of the database is calculated by the length of the Index*P field, but the English version is loaded which is called Index*B and results in a value of zero. Perhaps using nrow(Data) would be simpler.

This typo in the Rmarkdown document has been corrected and, as adviced, we now use nrow(Data) to minimize the risk of error.

Some minor changes to language and phrasing in the article are suggested.

We thank the reviewer for these suggestions to improve the text. It is very much appreciated as we are making efforts but are not native English speakers. They were all included in the new version.

Review by Clarissa Belardelli

I am going to articulate my review into three parts: 1. Scientific content 2. Bibliography, figure captions 3. Translation

1. Scientific content The paper is very good and the argument amazing. method is well conscructed; the overview maybe is not too vaste but there is a good bibliography.

Thank you for your comment.

However something is not clear: P. 4: “Presently, the dataset of lapidary artifacts contains 0 entries, originating from 87 sites”: so, how many entries does the dataset contain?

This was an error in the Rmarkdown document also recognized by other reviewers. It has been fixed now.

P. 5: “We will thus describe only the English tables Islands, Sites and Beads. Each table also exists in French 3.0.1 Islands table (ISLANDS and ILES)” but it is shown also in french. So, what do you mean actually? It is not clear.

We tried to make this sentence clearer now by changing the sentence that has now become : “Each table exists in French and English. For the sake of simplicity we will describe here only the English version.”

P. 5: 3.02, Sites Table. ”Nb_beads is the calculated number of artifacts related to this site in the BEADS table” That give place to confusion! A. Artifacts or only beads? B. In each site or in the structure linked to the site (see below in your paper)? It would be pretty useful to add something like presence of other lapidary artifacts from the SITE: Y/N and then specify which type of, and the number of each type in each structure.

Thank you for pointing this. This Nb_beads indeed encompass beads and pendants and other artifacts related to lapidary production (raw material for example). We then changed the name of this field to Nb_artifacts. As it is written in the description, this field relates to the site globally, and not the structure. Most site do not give enough detail on the position of the artifacts in structures so this would be irrelevant for too many artifacts.

P. 6: “Perforation is the number of perforation”. But if you found that object in bibliography, and you cannot explore the object, how can you decide whether the hole is only one or more than one? It would be much helpful PERFORATION: Y/N for doubtful cases, so you would have at least a generic information.

We worked on this part of the dataset thanks to your comment. We added the number of perforation(s) (Nb_Perforation), and simplified the content of the position (Pos_Perforation) and shape of the perforation (Shape_Perforation).

2. Bibliography, Figures captions There are some problems in the rendering of bibliography, but perhaps it depends on my pc. Cody's quote is different in the three titles of the bibliography: why?

Thank you for pointing this, we did not notice this discrepancy. The name Cody is indeed different on the three references listed here. Her PhD thesis states "Ann K. Cody", one of the article contains "Annie Cody" and the other one "Annie K. Cody". We will finally keep the name written on her PhD thesis since we think this is probably the most reliable.

In the text, figures are referred to with "Fig." but each caption has "Figure": why?

We used a LaTeX template that makes this choice for the layout. As I can see from several journals, it happens that in the text it is (fig. x) and the caption of the figure is "Figure x." (for example Scientific Report or Bulletins et Mémoires de la Société préhistorique française). We thus chose to keep it that way for the preprint. If we publish it in a journal we will follow the journal's template.

Fig. 4: Figure 4: A. Screenshot of the ArkeoGIS application, a simplified GIS online system. B. Zoom on Guadeloupe, sowhing the potential of the ArkeoGIS visualization tool. (showing) To be corrected!

Done

3. Translation In general, your english could be better. You should ask for a native speaker who helps you.

We tried our best to write a text that is fully understandable to English speakers, and it seems that this is the case. We are not-native English speakers indeed, and appreciated the comments of other reviewers to improve the wording. The text is apparently understandable and, despite being aware of the importance of spreading our results in English, we are also aware of the barrier English is for non-native English speaker (e.g. Hanauer et al. 2019 Linguistic Injustice in the Writing of Research Articles in English as a Second Language: Data From Taiwanese and Mexican Researchers) and do not want to have to pay for or use time of a colleague to improve our English language when it is already understandable.

There are some minor mistakes. I noticed two of them but there could be more. Here, one: - P. 6: State specifies is (IF) the object is complete or broken To be corrected!

Thank you for pointing this typo. With your help and the one from the other reviewers, we hope there are no more mistakes like this.

Finally, the paper is good but needs some more attention.

Review by Stefano Costa

This data paper describes an interesting and very wide-ranging dataset about a specific artifact category found in archaeological contexts across the Caribbean region. I can recommend its publication with some revisions described in the following review comments.

Section 1

Spatial coverage

Since the spatial distribution of sites is concentrated in the Eastern part of the Caribbean, with the exception of Jamaica and Bahamas, it seems useful to provide a more detailed description of the coverage beyond the geographic bounding box, perhaps including the names of smaller archipelago groups.

We added this sentence to the Spatial coverage paragraph: "We registered the lapidary artifacts for all the regions of the Caribbean arc which can be divided in the Lesser Antilles (Leeward Antilles, Leeward Islands and Windward Islands), the Greater Antilles (Puerto rico, Hispaniola, Jamaica, Cuba) and the Lucayan archipelago (Bahamas, Turks and Caicos Islands)."

The authors mention that some islands did not yield any lapidary artifact, or at least none that we could find in the literature and I think it is important to give a list of what islands were investigated but yielded no results.

The different islands of the Caribbean arc were not investigated one by one, but the literature was thoroughly analyzed to find all the lapidary artifacts in this region from this period. Given the way we created this database, the islands which are not listed in the SITES_EN table can be considered has having no lapidary artifact in their archaeological record.

The map in figure 1 is a little confusing because some of the site IDs do not match those in the table or the sites listed in the table are not labeled on the map. It could be useful to draw the bounding box in the map, perhaps differentiating islands that are included in the dataset versus those that are not included.

The map in figure 1 is extracted from a QGIS project and therefore the names of the sites visible or not on the map are based on this software handling of visibility. We updated the map so that the names on the map match those in the database. As for the differentiation of the islands, since the literature review was not based on geography but on keywords, they have been all included in the research and the absence of sites on Cuba and Hispaniola do reflect the published archaeological record.

Section 2

Steps

When the authors write that the dataset of lapidary artifacts contains 0 entries. This **must** be corrected with the actual number, by changing the code snippet at line 86 of the PAAF-datapaper.Rmd source file from r length(Data$index_P) to r length(Data$index_B), which correctly gives 5011 entries.

This has been corrected in the new version.

Section 3

Dataset description

The ISLANDS and ILES tables are not included in the archived dataset at SocArXiv / OSF. These **must** be added for the dataset to be complete.

This problem has been also underlined by other reviewers and the ISLANDS and ILES tables are now available in the repository.

The dataset description is very detailed and well organised. There is a lot of potential for reuse and further analysis with the radiocarbon data, as published JOAD data papers have shown: I may suggest to split the *Date_BP* field in two separate fields, to have better error checking (it becomes two integer fields instead of a text field) and make it easier to reuse the information. If the field contains more than one date, perhaps a separate table could be added.

For this new version of the database, we moved the information about radiocarbon data from the BEADS table to the SITES_EN table. We followed your advice to split the fields of calendar and calibrated ages and we integrated three dates whenever it was possible. New fields are therefore integrated in the database and explained in the text. The screenshot of the relations between tables has also been updated. There is a recent review paper about radiocarbon dates in the Caribbean (Napolitano et al. 2019) so our work will not be that useful for this specific purpose probably, even if we add some dates that are not in this review paper.

It is unclear why the dataset is described as CSV but the reproducible paper is based on the XLSX version, this could create unnecessary complexity and/or mismatch between different versions of the dataset.

We follow your advice about simplicity here and we now work only with CSV files. It was indeed an unnecessary potential source of error.

The Filemaker server is useful for quick interaction and most importantly contains photographs and drawings. A reason for not including those in the archived dataset should be given, since it represents a major source of information for the study.

The photographs and drawings visible in the Filemaker application are from very diverse provenance, mostly from screenshots taken from publications or from our own work on archaeological material. Understanding and using correctly the intellectual property of each and every picture to put them all in a freely available folder would be way too complicated. Since the Filemaker server only shows the pictures one by one, it would be the same work to *download them all than taking them from the publications, so it is very unlikely that someone will do that.*

Review by Li-Ying Wang

This paper describes the dataset of lapidary artifacts during the Ceramic period in the Caribbean islands. My comment below is based on the criteria listed on JOAD website that focuses on the description of the data and best practices for data deposition.

For the method section, I think the paper provides good information for understanding how the dataset was created, including the source of data, the methodology for collecting data, sampling strategy, quality control, and constraints. But I would like to suggest to the authors to provide more details about how they searched the ornament-related words in literature. For example, they mentioned "the words…have been systematically searched for." What is the exact method for systematically searching? I think it would be better to specify.

We added the following sentence to the new version of the manuscript: "Numeric literature was the main source of documentation and the search for specific words was thus done thanks to the pdf reading software. For less recent literature or scanned documents, Optical Character Recognition (OCR) was first applied to the documents. In physical books, we read the text, looked at figures and tables, and we used the index to find the information we were looking for."

Similar issues for the quality control section, how did they do for data cleaning?

Data cleaning for the numerical values is already explained in the manuscript and the use of dropdown menus, as described in the manuscript, is meant to avoid a big need to data cleaning afterwards. To be more precise, we modified the first sentence and added a second one: "Data cleaning and consistency have been realized thanks to the use of standardized thesaurus with dropdown menus to avoid typos for most of the fields. For other fields, we created lists of values for each variable to spot the discrepancies."

For constraints, how did they indicate the missing values in the dataset? I can see most missing values are indicated by blank in csv files, but some are indicated by a dash. Are these the same? This needs to be corrected or clarified. This can be addressed in the main text since incompleteness is a constraint in their dataset as they recognized.

Thank you for pointing that. It was indeed a problem to have sometimes blanks and sometimes a dash. We have changed that and now it is always a blank for missing values.

In addition to the above questions, there are some mistakes in the text:

1, Page 4, 2.1 steps: "the dataset of lapidary artifacts contains 0 entries, originating from 87 sites". Surely zero is not correct. There should be a number instead of 0.

This was an error in the Rmarkdown document also recognized by other reviewers. It has been fixed now.

2, Page 4, 2.4 constraints, “The quality of information has been problematic for several topic of the database” should this be ‘topics’ as a plural?

Done.

3, Also, this sentence “the quality of the reproduction of ancient photographs in the numeric documents now accessible for this literature” is unclear and would be better to rephrase.

We clarified our meaning by modifying this sentence to: “the quality of the reproduction of ancient photographs in scanned or photocopied documents”.

For the openness of data, they deposited their dataset on OSF where can be easily accessed and examined. The data is actionable and mostly labeled nicely. However, the information in the manuscript and repository about the number of files is not consistent with each other. In the 3.7.1 Download section, there are six csv files listed in the main text, but only four csv files are uploaded to OSF. The “ILES” and “ISLANDS” files are missing.

The ISLANDS and ILES files have been added in the directory while the review process was ongoing since we continued to work on it. We should have waited until the end of first round of review indeed, sorry for that. These files are described in the new version of the manuscript.

Also, in OSF, what are the xlsx files under the folder of Data paper JCA2020? The file structure and naming on OSF is confusing and needs to match exactly what is described in the paper. They should make it clearer in the paper and use more informative names instead of just “Table 01.xlsx” to guide readers or users.

The authors added to their OSF repository another folder for a former article in the same research project, while this first round of review was ongoing. We have improved our knowledge of the OSF framework platform and now both datasets are in different components of the same OSF project. We modified the link to the dataset for this specific dataper accordingly.

This paper meets most of the requirements for publishing. It is well organized with detailed archaeological contexts at the beginning, followed by data information, but the current version has some incorrect descriptions that are required to be solved. I recommend acceptance after the corrections of the issues mentioned above.