Repository Report

Data and Metadata Profile

The data I will be discussing is a collection of compiled observational data regarding European prehistoric archaeological digs containing jewelry and other adornments. The data consists of four distinct datasets. These datasets are divided into two distinct categories of data: raw and processed. The raw datasets consist of two datasets, the first is sites: a dataset listing the sites where prehistoric adornment items were found. Then there is structures: a dataset listing the structure type and name where prehistoric adornment items were found. The processed datasets consist of records: a dataset of adornment items found at these sites and structures records: a dataset displaying more detailed information about the locations where these items were found. Despite their similar-sounding names, all datasets contain unique data regarding the archaeological sites. All four datasets are available via download as CSV files. I have successfully downloaded/viewed them on Mac and PC. They are accessible via Numbers or Excel.

The data is from the Prehistoric Europe's Personal Adornment database. An EU-funded database for collecting and sharing European prehistoric site data. The data was "extracted from bibliographic resources". The specific criteria for which archaeological resources were chosen and why is absent from both the Zenodo page and the PEPAdb. However, this omission may be rectified in the future as the database is far from complete.

This data has a slew of potential stakeholders. The first stakeholders that come to mind are archaeologists. Archaeologists can use this data to identify important sites for further excavation, relevant sites for their research, and previously investigated sites. Archaeologists are also the people creating and curating this data. Thus, Archaeologists have many incentives to ensure the data remains current and useful. The PEPAdb's goals are to encourage this data to be shared and used by all. Therefore, scholars of all kinds are another stakeholder. Scholars in other fields could analyze and incorporate this data into their work and specialties. Given the PEPAdb's goals, ensuring these stakeholders can use the data is a major goal. The EU is another stakeholder. PEPAdb is funded by the EU. The EU has an ongoing interest in ensuring this data is accessible and useful. Anyone with an internet connection is also a potential stakeholder. Anyone can find and access this data. This is an important consideration especially when considering accessibility. This universality ensures that this data can/may be approached from a variety of perspectives. It also means that any culturally sensitive or identifiable information needs to be very carefully considered.

This data does not appear to have any use restrictions. The website and the Zenodo page don't list any. There may be restrictions as there are funding documents on the Zenodo page in Spanish that I couldn't read. However, it seems unlikely that these restrictions wouldn't be mentioned elsewhere.

The very basic metadata for the datasets can be downloaded in the form of an xml file meeting Dublincore standards from the Zenodo page. However, the more detailed metadata needs to be downloaded from the PEPAdb website. It is available in TTL (Terse RDF Triple

Language). A format designed for storing RDF (Resource Description Framework) metadata more efficiently. TTL requires specific software to open. For example, GraphDB or Blazegraph will open them. These require seeking out and are harder to locate than most software. The title metadata feature for the datasets is confusing due to their use of similar terms. However, the titles appear to be standardized making change unlikely. The dataset keywords only contain personal adornment omitting other potential access points like archaeology or prehistory. The metadata elements for the data vary by dataset. For example, sites only lists the site, whereas structures has elements like site, country, and structure type. While there are repeating metadata elements each dataset has different features. The metadata features are not explained or defined. Users have to learn what the metadata features represent via context clues, additional research, and/or expert knowledge. For example, adm1 and adm2 are not explained. It is through research one learns that these refer to two levels of administrative government. The metadata features are not very comprehensive. Even the most comprehensive datasets do not possess many or particularly in-depth features. Furthermore, all datasets feel like they could be expanded with additional metadata. Sites that only feature one metadata feature could be expanded with additional features like state, climate, etc. The database is still a work in progress; perhaps more robust metadata features will be forthcoming.

The Zenodo page does not mention any metadata standards. However, the Prehistoric Europe's Personal Adornment database states that the geographic metadata meets ISO metadata geographic standards. The raw and processed datasets meet RDF standards.

As addressed above the metadata feels both vague and lacking in explanation. Additional keywords regarding archaeology or prehistory should be added to the dataset metadata to increase findability. The metadata for the data would also be improved with additional metadata features. For example, features like culture associated with the sites, additional information regarding the digs, or descriptions of the adornments would be useful for researchers. Currently, this data paints a useful but broad picture. It tells us that 50 beads were found in a cave near Madrid. However, additional details like what shape those fifty beads were, is this site associated with a bead-making culture, and nearby sites with similar objects could be useful. This additional metadata would help users better understand these sites. I especially feel this additional metadata may be useful for users lacking familiarity with archaeology and prehistory. Currently, the lack of specificity and lack of explanation of terms may exclude some users from using the data to its full potential. The presence of additional explanatory metadata features may remove some barriers. It would also be helpful to enhance findability with additional keywords. For example, presently researchers looking for data on digs involving the beaker culture may not find this database.

Zenodo contains no citations of these datasets. However, the Zenodo page is recent, having been published on January 21, 2024. This data may simply be too new to have been cited yet. While I found many papers referencing the topic I found no papers that cited this data in the UW libraries. I advanced search using multiple terms and individually via subject, and keyword searches for adornment, prehistoric, archaeology, Prehistoric, Jewelry, and Prehistoric Europe's

Personal Adornment database. I also tried searching by author for the contributors listed on the Prehistoric Europe's Personal Adornment database page. They had papers, but none citing this exact data.

Repository Overview

The repository I have chosen to discuss is Paradisec (Pacific And Regional Archive for Digital Sources in Endangered Cultures). Paradisec is a repository of musical works, cultural works, grammar tables, language snippets, fieldwork, and other data that helps preserve endangered cultures/languages. Paradisec's main goal is to preserve works that are in danger of being irreparably damaged. The majority of the data in the repository is anthropological recordings or anthropological field notes.

I chose this repository for two reasons, firstly because its subject matter interests me. I am interested in the ways researchers preserve endangered languages/cultures. I especially wanted to explore the role repositories serve in this work. Secondly, I wanted to investigate how this repository handles issues of Indigenous data sovereignty. Indigenous data sovereignty is an extremely important, but often under-discussed topic in data management. Indigenous data sovereignty is a very complex and nuanced idea. A basic explanation is that it refers to Indigenous people controlling their data and the data created about them. A lack of Indigenous data sovereignty can cause major harm. For example, the Havasupai Tribe vs Arizona State legal case, where the Havasupai tribe provided blood samples to Arizona State University for a study. However, these blood samples were used for additional tests without the Havasupai tribe's consent or knowledge (Kukutai and Taylor 48-49). This included tests that invalidated Havasupai religious and cultural beliefs like DNA mapping (Kukutai and Taylor 48-49). Sacred/culturally sensitive knowledge is also an important concept to consider with a repository like this. This refers to sacred and/or extremely important Indigenous knowledge that the Indigenous community has decided should not be public knowledge. (Burkhart 17-18). I was curious to explore how the repository would handle these concerns while accomplishing its own goals.

Paradisec is not open to general submissions. Paradisec defines their criteria for submissions as "deposits from linguists, ethnographers, ethnomusicologists and other researchers". Paradisec presents its role as safely storing data with an eye toward future use. Paradisec also aims to digitize reposited data and offers to digitize received analog data if possible. To deposit data one needs to sign up for an account and then fill out a metadata form. One can then provide Paradisec with the data. Digital data can be directly sent to Paradisec. Physical data requires an explanation/description of the data to be emailed to Paradisec which they then review before deciding. An account simply requires an email and password. While the repository is closed it seems likely that exceptions are made. The repository aims to preserve as many works regarding endangered cultures/languages as possible. Therefore, it seems doubtful that rare collections in vulnerable formats would be outright rejected because they were not submitted by ethnomusicologists.

The geographic scope of the repository should be briefly discussed. Paradisec's geographic scope is hard to determine. Their website primarily mentions Indigenous groups in Oceania. For

example, a photo on the website shows people in Vanuatu reading a dictionary of their language. Paradisec's process statement indicates that they utilize special criteria to gauge whether they will accept a deposit. These criteria are: the language's status, the state of the deposited work, and the work's regionality. However, Paradisec's website never explicitly states what geographical areas they accept. Furthermore, there is data from/regarding distant geographical locations such as Italy, Denmark, and the Czech Republic. This data is part of a massive collection of anthropological interviews from Arthur Capell, an Australian anthropologist. It is difficult to determine if similar work without the Australian connections would be accepted. However, there is a large collection of Nepali recordings that have no overt ties to the Pacific. Therefore, currently it seems reasonable to assume that work about Endangered languages/cultures from around the world are donatable. The geographical scope may be more explicitly addressed when actually donating.

The repository is very cognizant of and prepared for data containing culturally sensitive knowledge. The submission page mentions that donors can set whatever viewing guidelines they require. This seems an important safeguard that allows data with sacred knowledge elements to be stored without being widely accessible. Paradisec also provides contact methods for Indigenous communities to use if they discover sacred information.

Viewing and interacting with the material stored in the repository requires users to create an account. As mentioned above, accounts simply require an email address and a password. However, while this will allow you access to most data, not all data is accessible or interactive. Accessible data has the word open in green. The site also mentions that some data stored in the repository is not viewable by users as well. These methods seem designed to safeguard Indigenous data sovereignty and sacred knowledge.

The data is available in a large number of formats and is accessible via several access points as well. For example, this interview about weaving is provided via jpeg, video, audio, and XML. The interview can be downloaded, read on the browser, listened to via the browser, or can be watched via video file in the browser. The files available are tied to the content itself. For example, this collection of questionnaires only has two available formats. However, it does still offer a downloadable and a browser-usable access point.

Paradisec's catalog is run using Nabu, a media management system designed by Paradisec. Nabu is open-source and provides a GitHub repository with the source code. Paradisec's catalog uses metadata that confirms to Dublincore and Open Languages Archives Communities Standards. Paradisec also offers a PDF explaining their catalog's metadata features. Paradisec's metadata files are provided in XML format.

Regarding the DIP of Paradisec, each file is in a different format. Currently, all formats when downloaded seem to simply provide the chosen file without additional metadata or identifying data. There may be additional files provided depending on the downloaded file or the format. The repository contains a wide range of formats and additional information may be present for different formats. There may also be additional documents provided to explain

modifications of data for reasons of cultural sensitivity. However, both are mere speculation on my part.

Recommended Data Citation-

Prehistoric Europe's Personal Adornment database. *Prehistoric Europe's Personal Adornment database raw and processed tabular data.* QUANTA[2]S, 2024. https://zenodo.org/records/10545573

Human Subject Considerations- The data in this repository has not been anonymized in any way. There is no data present in this repository that contains any personally identifiable information.

Long-Term Preservation Considerations-

Data: The data should be preservable in the long term. The format was chosen by a group considering long-term preservation. Furthermore, the data is part of an EU initiative. Therefore, it seems reasonable to assume that this data's long-term preservation is a high priority. The processed and raw data is also able to be accessed via any software. It can be opened using Excel or numbers.

Metadata: The metadata on the other hand seems more likely to be lost in the future. Two files are XML. XML is a very standard format that is easily accessible. Therefore, the likelihood that metadata will be lost is low. The other metadata however is in TTL format. TTL format requires specific software to open. There are lots of software options for opening TTL files and they are easily findable via searching the internet. However, the software capable of opening TTL files appears to be mostly fan-maintained. I couldn't find anything that could open it and was easily accessible on the Windows or Apple store. This lack of a more official licensed software may create preservation issues. While currently there are a lot of options these options may become poorly maintained. This may result in the metadata becoming inoperable.

Copyright-The data does not appear to be copyrighted. The PREDAP never indicates that the data is copyrighted.

Works Cited

Burkhart, B.Y.. What Coyote and Thales can teach us: An outline of American Indian epistemology, In *American Indian thought: Philosophical Essays*. (p.15–26). Blackwell Publishing., 2004.

Kukutai, Tahu and John Taylor. *Indigenous Data Sovereignty : Toward an Agenda*. 1st ed., vol. 38, ANU Press, 2016.