

When it comes to archiving, many argue that it should be involved from the very beginning of a site, if not before. Preserving and protecting data and results from a site is highly important, regardless of their form. Naturally, in the modern, technological world we find ourselves in, digital archiving has become an important step in the archival process. Various organisations (the Archaeological Data Service [ADS] and Historic England's OASIS project to name two prominent figures) have stepped forwards to assist in this. OASIS and the ADS use the FAIR acronym to describe the aims of data and metadata construction: to be findable, accessible, interoperable, and reusable. The FAIR Guiding Principles were outlined by Wilkinson *et al.* (2016), with subsections of each one offering the aims of best practice. These principles will be used to examine my constructed dataset and metadata statement and will assist in understanding the extent to which I have achieved this with my data. Figure one lists the different FAIR principles according to Wilkinson *et al.* (2016), and this will be referred to throughout this assignment. Whilst not written specifically for archaeological data, the FAIR principles help to unite data creators and users around the world. While organisations may differ from one another, they hold similar goals; a level of cohesion needs to be reached. Metadata has proved itself a valuable tool in accessing and understanding data in this way, making it highly findable if done correctly.

Findable data and metadata rely upon consistency. Although this exercise was theoretical, if these data were to be entered into an index, consistent application of defined terminology would ensure that they were findable. In my dataset, I have used a consistent unit of measurement (metres) and ensured that any measurements given were corrected to use this unit. Furthermore, the metadata I created for the dataset I constructed reiterates the unit of measurement and notes that the

appropriate fields (length, width, diameter, etc) are written as numerical values. I took care to comb through my extensive dataset and ensure that this was consistent throughout. This should ensure that other data presented as numerical values (such as sample numbers or context numbers) are not mistaken for measurements; artificial intelligence might struggle to differentiate the two. Ensuring that this is the case with a cohesive metadata statement is more effective than relying on the likes of dataset formatting for presentation of data, which can be easily lost. This helps to minimise human error and save time. Well-described metadata ensures a level of agreement on how we define specialist terminology within an area. For example, I thought carefully on how to define descriptive phrases, such as “shape in plan”, in my metadata statement. In my experience, these terms can be difficult to understand to one unfamiliar with them or two individuals may disagree on how they might define them. I have known there to be mistakes arise from this confusion, and I wanted to reduce this where possible by clearly defining the term. Using relevant Dublin Core Metadata Element Set components helped to sufficiently describe file-level metadata (Dublin Core 2020). These metadata elements may seem insignificant (software used, file names, language etc.), but they are vital in easily finding and accessing data. Unique identifiers massively aid findability. The OASIS system advises use of an ORCID iD upon account registration, although it is not mandatory; I used this exercise as an opportunity to register for my own ORCID iD. This unique digital identifier serves to define an author beyond a name or title that may change over time, making data more easily findable. Persistent identifiers, such as Digital Object Identifiers, provide a means to locate data not solely based on its current digital location. In doing this, metadata may be accessible even when the dataset it is

formed from is not. There was no opportunity to provide a unique identifier of the dataset and metadata, but this would have been provided upon publication.

Accessibility is highly important and was a key consideration of the creation of this dataset. I chose to present the data in the form of a spreadsheet (and not a database) due to its somewhat simple nature (Niven 2011). The dataset I have presented is a transformed site context log, the first step in processing data collected on site for archiving. Each section relates to a previous section, meaning that a singular sheet is the best way to display them. On-site digital context logs used by Headland Archaeology utilise cell-specific formatting to display different types of data. This aids accessibility, findability and interoperability by ensuring that values are easy to differentiate. For example, contexts numbers, sample numbers, and measurements are all presented using integers of different lengths. A spreadsheet allows one to limit the type of numerical value that can be used in each section, reducing the possibility of human error and ensuring consistency in numbering practices. If the background of the number is lost, it should be easy to differentiate what it pertains to. Accessibility here also relates to how the data is stored and implemented. I have saved my metadata and dataset as comma separated value files, considered the best file type for long-term preservation. File types typically associated with databases (and not spreadsheets) are notoriously difficult in this sense, such as the changing file types used by Microsoft Access over the years rendering older files difficult to open. Presenting a simple dataset as a comma-separated value does not lose value in lack of formatting and benefits from smaller data files due to high compressibility. If my dataset and metadata table were to be deposited, it would be done through the ADS-approved OASIS index. This would also serve to make my data and metadata interoperable, bettering the use of

computer and software systems to exchange use information. It should be noted that accessibility of the dataset does not necessarily mean that the dataset be made open to all. While the OASIS index is free to use, it requires users to create an account that is linked to (at least) one organisation. One may wonder how this serves to make archaeological data accessible to one not attached to an organisation, such as an interested member of the public. It is a sad reality of commercial archaeology that results are often hidden or obscured from public view, often due to client wishes or project sensitivities. Data in the form of grey literature reports catalogued by the ADS is accessible to members of the public yet the time between excavation and public access can vary, often being a lengthy process. Well-constructed metadata is important as the data it is based upon can often disappear or become inaccessible. If a metadata statement were still to exist in this case, those involved (individuals or institutions) could assist in finding the data or providing more insight to those who need it. Such information can seem somewhat insignificant when compared elements describing data explicitly, but they are just as valid. It is vital that even small details such as copyright owners, authors, dates created, and software versions are used so that the original data may be understood and utilised, and I have taken care to incorporate these details accurately in my metadata statement.

Interoperable data and metadata were an important consideration. This was particularly important given that my dataset was created from site records, which are a raw form of data. The raw data had to be processed to produce a dataset that could provide valuable insight into the site without destroying any interpretative value. This is why I chose to include the “interpretation” and “discussion” columns of the dataset. The “discussion” column took particular time to process. This is an

intentionally vague section of the site context sheet, allowing the excavator to note anything which may have stood out to them during excavation. Sometimes they do not, explaining large gaps in this section in my dataset. I sought to make the data more cohesive and uniform, initially doing so by spell-checking this column thoroughly. This makes data more searchable and further eliminates time lost due to human error. I also sought to better the interoperability of my data this by checking that terminology was consistent. Given that different individuals had used the same system to input data, there were many variations. I combed through the dataset with certain key words in mind, many that I was aware might be problematic from past experience in both excavation and post-excavation. For example, I changed all instances of “post-hole” or “posthole” to “post hole”. I also ensured that correct terminology was used (“circilinear” had been incorrectly used over “curvilinear”) and that it was spelled correctly (I changed instances of “hypercaust” to “hypocaust”). Fair Guiding Principle I1 advises use of a “formal, accessible, shared, and broadly applicable language” to ensure data is interoperable. The use of FISH (Forum on Information Standards in Heritage) vocabularies is a typical solution to this in archaeology, although they are not always adhered to. Site context data as it is recorded on context sheets varies widely from person to person, and there is an understanding within archaeology that it will be transformed during post-excavation via data input. To do so, I used relevant feature type terminology found in the FISH thesauri and changed terms accordingly (such as post hole, post pad, and post pipe) (Forum on Information Standards in Heritage 2021a). Certain descriptive words were corrected to a uniform spelling (“Subrectangular” and not “sub rectangular” or “sub-rectangular”) (Forum on Information Standards in Heritage 2021b). It should be noted that the “sub” prefix is often used as a qualifier in context discussion leading to

clumsy terminology, such as “subsquare” or “sublinear”. I left these unaltered to preserve the interpretation of the excavator, but it should be noted that their usage is often confusing and somewhat bad practice. Richly described metadata can also assist in interoperability. Field descriptions allow both human and machine to distinguish the type of data being used. This would give the human user an idea of how to present, use, and recreate the data, and would enable the computer to compile and utilise data appropriately.

Reusable data and metadata are also highly important. Fair Guiding Principle R1 specifies that data and metadata should be richly described, using suitable formats and templates. I sought to achieve this by using the spreadsheet metadata template advised by the Archaeology Data Service and closely following the completed example (Archaeology Data Service 2023). The metadata has been sufficiently described to allow reusability, and I tried to provide as much detail without attempting to predict the usage of the data. Creating both the data and metadata along ADS guidelines ensures that they can be submitted to the OASIS index and assist in future research. Thorough description of the metadata allows a data user to efficiently search and discern which elements may be relevant to their research, such as relevance to a certain research framework, time period, or even individuals or institutions that were involved with creation of the dataset. By using an ADS-approved metadata template, a “clear and accessible data usage license” (Fair Guiding Principle R1.1) and “detailed provenance” (R1.2) would be provided upon publication. File-level metadata provides provenance and licensing information by noting contextual information, such as the software used in creation of the dataset and copyright holders, which I have provided. This provides information as to how data may be reused in the future. As many organisations do, Headland Archaeology

provides its employees with guidance on how to structure data, encouraging the use of FISH vocabularies (R1.3) and ADS guidance, yet also internally created digital housekeeping guidelines. Using such guidelines aims for a level of cohesion on a company- and industry-wide level, which is why I have corrected data to the FISH-advised terms as aforementioned.

Of course, all of this relies upon both data and metadata being used correctly. While I believe I have achieved the best possible representation of this data, I understand that, due to the nature of the dataset, there may be human elements that might hinder its computational value. This is due to individual differences which can muddy even the most seemingly simple aspects of a context dataset, such as taking and recording measurements. I have known there to be disputes over what constitutes a section width or length, and it can be seen in this dataset that (if infrequently) individuals have provided multiple measurements or a range of measurements. It can be baffling to attempt to comprehend how similar factors may also affect data. While bearing in mind the inherently human nature of archaeology, we cannot underestimate the importance of good data and metadata practices. Communicating with and involving members of staff (or even society) from all walks of life serves to benefit this and only improve best practice alongside technological advancements.

Word count: 2144