

Assignment 2: Construct a database or dataset of selected archaeological material/records, provide metadata statements in line with current Archaeology Data Service guidance, and a commentary explaining and justifying its structure and the categorisation / classification of the data.

Recent industry reviews such as the *Planning for Archives* and *Matrix* projects (ALGAO 2019; May, Taylor and Binding 2023) have highlighted widespread inconsistency within post-excavation processes, and the production of ‘useful’ archives - variously attributed to a lack of post-excavation manuals (ibid., section 4.5), and structural, commercial tensions with academic interests (Davies 2017, p.174). For the pottery specialist, recent decades have seen popular adoption of core standards and policies (PCRG 2015; PCRG, SGRP, MCRG 2016; MCRG 1998) enabling and renewing the production of consistent, useful datasets – so for our purposes, the production of consistent data is likely less of a concern than how its *usefulness* is affected by the manner and extent of its communication. In the age of ‘open archaeology’ (Wilson and Edwards, 2015) the challenge is to make our data as widely accessible, intelligible and re-usable as possible for the wider world of research - otherwise, why keep an archive at all?

Principles established by GoFair (2016) suggest that for a modern archive to be “Findable”, “Accessible”, “Interoperable”, and “Reusable” (2016), it must meet be prepared specifically for machines to find, search and read it. Publicly accessible digital archives such as Archaeological Data Services (ADS) provide enormous opportunity for human researchers,

but must be fed data appropriately prepared to work with the tool itself *as well as* remain intelligible by human end-users. With this in mind, the aims of this project were:

- To create a stable, archivable, ADS-compliant format for a meaningful group of data, i.e. effective preservation by record
- To support useful characterisation of a pottery group by identification of its elements and abstraction of significant attributes
- To organise and abstract data in ways that facilitate a wide range of research aims, both immediate and potential
- To create a legible, accessible and useful reference document for human and machine readers

The resultant dataset (Gifford 2024a) is adapted from data I originally recorded from projects at Quidenham, Norfolk - significantly edited/reconstituted for this assignment, to explore new modes of data construction. The dataset now does not necessarily reflect the original project archive, and only represents a thought experiment rooted in a real-world example. The following commentary explores how these aims were approached in shaping this dataset, drawing primarily upon advice from the *Guide to Good Practice* (Niven, Digital Antiquity and ADS, 2024).

Structuring the Dataset

It is a requirement of ADS deposition that datasets must have “a logical data structure” (ADS 2024a), i.e. represent a meaningful group of related data, be it geographically, chronologically, etc. This may appear axiomatic, but more logical grouping greatly aids digital search effectiveness and is key to discoverability. To that end, my dataset is overall

structured by theme and location: the data relate to a single material type (pottery), and all derive from two related investigations on the same land. The overarching 'pottery' and 'Hargham Road' taxonomies map to the spreadsheet file as a whole (*Pottery-AllPeriods.xlsx*). Hierarchically below that, data are sub-divided into broad periods (prehistoric, Roman and post-Roman), mapping to individual worksheets which each constitute a period-based catalogue (e.g. *Pottery-Prehistoric_Catalogue*). A further, less significant taxonomical element, representing the two investigations which produced data, maps simply to a mandatory attribute (*Site_code*) associated with each record.

In strict taxonomical terms, a combined table of all pottery fabrics (*Pottery-AllPeriods_Fabrics*) sits above the period-based sub-divisions, as does the table of attributes recorded (*Pottery-AllPeriods_AttributesCodes*): this hierarchy is reflected in the naming of the sheets for clarity. However, both are stored in parallel to the catalogues, at the level of worksheets within the spreadsheet: this is justified on functional, rather than taxonomical grounds, since the fabrics table will ultimately be used similarly to the catalogue tables as a primary record, and both the fabrics table and the abbreviations table should be easily accessible for close consultation when using the catalogue tables. While the dataset is constituted in its current format (*.xlsx*), this hierarchy remains structurally inherent, and hardly requires explanation. However, it is worth defining since plaintext formats (e.g. *.csv* or *.txt*), which support only single-sheet structures, would require each table to be positioned as an individual file in a sub-directory constructed to replicate this hierarchy for accessibility and dissemination (ADS 2024b).

Choosing a format

The choice of an Excel .xlsx spreadsheet format for the dataset offers significant advantages: foremost, data can be recorded and submitted for archive with no migration to plaintext formats, since the .xlsx file type is an open access format, and stable enough to be suitable long-term storage (Niven, Digital Antiquity and ADS 2024). It therefore meets our stated criteria of ADS compliance, archive stability and can immediately; furthermore, by avoiding migration and division into single-sheet formats (e.g. .csv), it does so while maintaining the dissemination-ready, accessible structure afforded by having related tables connected within the same file. Spreadsheets are also more widely used than relational databases, so this format will also be immediately familiar and usable to a wider group of users. An Access database, by contrast, would not only require migration to multiple plaintext files for storage, but the complex relationships within that database would need to be carefully mapped and stored alongside those files, in order for the dataset to be reconstructable for full functionality.

The suitability of .xlsx formats for archive does depend on how it is used: the meaningful content of the file needs to be easily translated into a plaintext format, in case such a migration becomes necessary (*ibid.*). This means certain functionalities should not be archived: these include embedded content and formulas. Illustrations, etc. are therefore stored separately to the catalogue. Any formulas used in my sheets to concatenate strings, or sum quantities (e.g. those I used to populate the *Pottery-AllPeriods_Fabrics* sheet) will also lose functionality at migration to .csv, leaving only the generated value in its place. However, since the data in the record are, by definition, intended to remain static after completion, dynamic features are no longer needed. A final consideration is formatted text: **bolded** or *italicized* text (e.g. used to highlight published fabrics, etc.) will revert to an

unformatted state at migration to plaintext. As such, this dataset expresses all meaningful information through structure and standard characters, since meaning denoted by formatting is vulnerable to loss.

To ensure compatibility with as many applications as possible, column headers have excluded potentially problematic characters such as spaces and punctuation, using only the basic alphabet and underscores. Headers such as “desc” or “date” have also been avoided, since these can be problematic with ORACLE software (Niven, Digital Antiquity and ADS 2024). These are small concessions to accessibility, but do require active consideration at the design stage of developing a new template.

Categorising the data

Good data-categorisation for recording isolates attributes for characterisation of the assemblage, and also provides equivalent points of comparison for re-use in wider research - as a broad example, splitting the assemblages into prehistoric, Roman and post-Roman catalogues is standard practice because almost any research will approach these separately. Our data categorisation is designed to closely match the attributes highlighted by the Standard for Pottery Studies in Archaeology for basic recording (PCRG, SGRP and MCRG 2016, p.12-13) – as demonstrated in Table 1.

Close comparison does reveal some slight variations between sheets (see Table 1). Two additional columns were added only to the *Pottery-Prehistoric_Catalogue* to accommodate data particularly significant to the prehistoric assemblage: *Form_features* and *Fabric_group*.

Table 1. Dataset attributes mapped to basic recording requirements (PCRG, SGRP, MCRG 2016 p.12-13)

Required data (PCRG, SGRP, MCRG 2016)	Attributes recorded		
	<i>Pottery- Prehistoric_Catalogue</i>	<i>Pottery- Roman_Catalogue</i>	<i>Pottery- PostRoman_Catalogue</i>
“Contextual unit”	<i>Context</i>	<i>Context</i>	<i>Context</i>
“Ware name, [...] type or [...] group”	<i>Fabric_group</i>	<i>Fabric_type</i>	<i>Fabric_type</i>
“Full fabric identification”	<i>Fabric_type</i>		
“Ware date range”	<i>Fabric_date</i>	<i>Fabric_date</i>	<i>Fabric_date</i>
“Sherd type”	<i>Sherd_type</i>	<i>Sherd_type</i>	<i>Sherd_type</i>
“Vessel class”	<i>Form_generic</i>	<i>Form_generic</i>	<i>Form_generic</i>
“Vessel type”	<i>Form_specific</i>	<i>Form_specific</i>	<i>Form_specific</i>
“Vessel size”	<i>Rim_diameter_cm</i>	<i>Rim_diameter_cm</i>	<i>Rim_diameter_cm</i>
“Surface treatment”	<i>Surface_treatment</i>	<i>Surface_treatment</i>	<i>Surface_treatment</i>
“Decoration”	<i>Decoration</i>	<i>Decoration</i>	<i>Decoration</i>
“Sherd count”	<i>Count</i>	<i>Count</i>	<i>Count</i>
“Sherd weight in grams”	<i>Weight_g</i>	<i>Weight_g</i>	<i>Weight_g</i>
“Number of vessels”	<i>ENV EVE</i>	<i>ENV EVE</i>	<i>ENV</i>
“Cross-fitting sherds”	<i>Crossfits</i>	<i>Crossfits</i>	<i>Crossfits</i>
“Joins within context”	<i>Comments</i>	<i>(Comments)</i>	<i>Comments</i>
“Evidence for use”	<i>Residues Condition</i>	<i>Residues Condition</i>	<i>Residues Condition</i>
“Sherd condition”	<i>Condition</i>	<i>Condition</i>	<i>Condition</i>
“pieces that require illustration”	<i>Illustration</i>	<i>Illustration</i>	<i>Illustration</i>
“[pieces that] require scientific analysis”	<i>Comments</i>	<i>Comments</i>	<i>Comments</i>

Both these columns were added to enhance the datasets useful characterisation of prehistoric pottery, given the particular heterogeneity of material from this period.

Form_features was added because by its nature, handmade pottery does not produce closely replicated forms, and it can be harder to characterise a whole prehistoric vessel's form from limited sherds than in other periods. As such, isolated formal features on sherds (e.g. rounded/angular shoulders in our dataset) may take on increased diagnostic significance (in this example, having possible chronological implications). By abstracting this data, it can be included for consideration without crowding and obscuring the

Form_specific/Comments columns with non-equivalent data types. Similarly, the attribute of *Fabric_group* was added to *Pottery-Prehistoric_Catalogue* to compensate for differences in fabric recording conventions. The localised nature of prehistoric production means fabrics require more granular detail for meaningful characterisation; PCRG guidelines recommend fabric types should be "site-specific" (PCRG 2015, p.22). Since new types are constantly created, the fabric data lacks the supervening layer of meaning inherent in Roman or Medieval types such as "Early Medieval ware" or "grey coarsewares", which relate sherds to groups, styles or periods. The inclusion of *Fabric_group* in the dataset makes space for this layer of meaning, linking the sherds to the wider PDR tradition, again without crowding the *Comments* or other columns. The addition of these columns then, not only enhanced the characterisation of the material, but by distilling the different levels of information into different attributes, makes searching and using the information easier for both machines and humans.

In line with the *Standard (2016)*, recording certain attributes, such as decoration and surface treatment, requires inputting multiple data components (i.e. technique, motif, position). For

these, the option of splitting these components into new attributes (e.g. *Decoration_technique*, *Decoration_motif*, etc.) was considered, since this would isolate specific points of comparison, and enhance data articulation for re-use. However, while this could work well for simple examples, any sherds with complex motifs or surface treatments, incorporating different techniques in different positions, would be difficult to articulate in this structure without complex splitting/merging of cells, which in itself impedes re-use and accessibility (University of Reading 2024). These attributes therefore remain as one column (i.e. *Decoration*, *Surface_treatment*), accepting that some complex textual inputs are inevitable, since more restrictive structuring would likely impede clarity in this case.

Capturing the data

Clearly, how attributes are categorised should relate to how best to express a particular type of information. In the example above, a relatively ‘free’ text input was deemed the clearest way to abstract complex visual features. However, completely unrestricted inputs are by their nature less standardised, and therefore harder to search, filter and synthesize – all key considerations for future analysis and research. An important way of mitigating this is the use of standard terminology, and strict adherence to including all required content: as such, the *Pottery-AllFabrics_AttributesCodes* sheet lists where recording follows accepted pre-defined terminology or definitions (e.g. description of prehistoric decoration follows PCRG 2015).

Similarly, typology is a key structural tool for standardising input on a regional or thematic scale. For instance, Brudenell’s classifications of PDR forms (2012) are followed in both *Form_specific* and *Form_features* on *Pottery-Prehistoric_Catalogue*, while *Fabric_type*

follows Anderson's Suffolk Fabric Series (2020) on *Pottery-PostRoman_Catalogue*, and the Suffolk Roman Pottery Series (as seen in Ashwin and Tester) on *Pottery-Roman_Catalogue*. By expressing data in terms of these typologies, coherency is instantly established not just within the dataset but on a regional scale. However, heavy use of type references can also create over-abstracted, lower-resolution data, glossing over variance, inconsistencies or uncertainties. Our dataset utilises direct description in the *Form_features* column to mitigate this effect, while still following terminology from Brudenell (2012) for consistency. The *Form_generic* attribute on *Pottery-Prehistoric_Catalogue* also deliberately does not follow the vessel classification system of Brudenell (2012) or Barrett (1978), because it would duplicate information: the components of those classifications are surface treatment, fabric and generic form, all of which are recorded as separate attributes. Sufficient data are already present to apply that structure if required for future research, but our dataset prioritises a more granular record of individual attributes, and an overall closer representational relationship to the assemblage itself.

Indeed, where a database can guarantee standard inputs with mandatory fields, number-only fields, limited text length, or only allowing values from a specified list, a spreadsheet does not have this functionality. This does allow the recorder more capacity for nuance, and avoids problems of false precision or decontextualised data that can arise navigating the restrictions just mentioned. However, it also leaves responsibility for consistency and thoroughness to the individual recorder; in reality, that consistency will likely develop as the individual's "style", perhaps regulated more by habit than design (cf. May, Taylor and Binding 2023, section 4.5; Davies 2017, p.139). For the sake of the data's future utility and intelligibility, though, habits must be made transparent by explicit identification of the

lexicon, standards and typologies employed. This information is invariably included in reports as methodology, but ideally should also be linked to the dataset itself. Indeed, the ADS metadata statement (Gifford 2024b) also contains explanation of codes and abbreviations, but for our dataset, slightly more comprehensive information regarding references and terminology is also built into the *Pottery-AllPeriods_AttributesCodes* sheet. This way, it is accessible to all users, even if they obtain the data without the ADS.

Data Clean-up

A final consideration in the production of the dataset is the 'clean-up' stage. After the completion of recording, this stage is key to improving the accessibility of the data, by improving digital search functionality and general human intelligibility. The process included ensuring abbreviations and codes were used consistently, and defined or referenced clearly (see *Pottery-AllFabrics_AttributesCodes*), checking spelling and accurate formula references, and removing any 'notes-to-self' from the working draft or moving notes worth keeping to *Comments*. In addition, at this stage, I checked for entirely blank columns, and populated them as necessary with a negative datum (e.g. 'undiagnostic', 'NA'). This is a precaution to make the data more compatible with screen-reader software, which can be confused by blank columns or rows (University of Reading, 2024). While this final scrutiny should be applied to all datasets, it should be said the process is undoubtedly lengthier when using a format with no input restrictions (i.e. a spreadsheet, rather than a database).

References

*References marked with and *asterisk are available in supporting documentation.*

ADS (2024a) *Instructions for Depositors* Available at:

<https://archaeologydataservice.ac.uk/help-guidance/instructions-for-depositors/data-structure/> (accessed 17 January 2024)

ADS (2024b) "File naming" in *Instructions to Depositors* Available at:

<https://archaeologydataservice.ac.uk/help-guidance/instructions-for-depositors/file-naming/>

[ALGAO \(2019\) *Planning for Archives: Opportunities and Omissions*](#)

ALGAO (2019) *Planning for Archives: Opportunities and Omissions*, Historic England [Online].

Available at: <https://historicengland.org.uk/images-books/publications/planning-for-archives/planning-for-archives/> (Accessed 23 January 2024)

Anderson, S. (2020) *Suffolk Medieval Pottery Fabric Series* Available at:

<http://www.suffolkmedpot.org/>

Ashwin, T., and Tester, A. (2014) *A Romano-British settlement in the Waveney Valley: excavations at Scole, 1993-4*, East Anglian Archaeology 152.

Brudenell, M. (2012) *Pots, Practice and Society: an investigation of pattern and variability in the Post Deverel Rimbury Ceramic Tradition in East Anglia*. Unpublished PhD thesis.

Davies, D. (2017) *The Development Of Archaeological Post-Excavation Within British Professional Archaeology*, BAJR Guide Series 46. Available at:

http://www.bajr.org/BAJRGuides/46_PostEx_Theory_Application/46%20Archaeological%20PostExcavation.pdf (Accessed 23 January 2024)

Go FAIR (2016) *FAIR Principles* [PDF] Available at: https://www.go-fair.org/wp-content/uploads/2022/01/FAIRPrinciples_overview.pdf (Accessed 23 January 2024)

*Gifford, C. (2024a), *Gifford_2318131_HPAR7004_2_Pottery-AllPeriods_Record* [Microsoft Excel Spreadsheet]

*Gifford, C. (2024b), *Gifford_2318131_HPAR7004_2_ADS_spreadsheet_metadata.xlsx* [Microsoft Excel Spreadsheet]

May, K., Taylor, J.S. and Binding, C. (2023) Stratigraphic Analysis and The Matrix: connecting and reusing digital records and archives of archaeological investigations, *Internet Archaeology* 61. <https://doi.org/10.11141/ia.61.2>

MCRG 1998 *A Guide to the Classification of Medieval Ceramic Forms* Occasional Paper 1. Available at: https://medievalpottery.org.uk/wp-content/uploads/2019/12/a_guide_to_the_classification_of_medieval_ceramic_forms.pdf (Accessed 5 June 2023)

Niven, K., Digital Antiquity and ADS (2024a) "File Formats" in *Guides to Good Practice* Available at: <https://archaeologydataservice.ac.uk/help-guidance/guides-to-good-practice/basic-components/databases-and-spreadsheets/creating-databases-and-spreadsheets/file-formats/> (accessed 17 January 2024)

Niven, K., Digital Antiquity and ADS (2024b) "General Considerations" in *Guides to Good Practice* Available at: <https://archaeologydataservice.ac.uk/help-guidance/guides-to-good-practice/general-considerations/>

[practice/basic-components/databases-and-spreadsheets/creating-databases-and-spreadsheets/general-considerations/](#)

Niven, K., Digital Antiquity and ADS (2024c) "Structuring your data" in *Guides to Good Practice* Available at: <https://archaeologydataservice.ac.uk/help-guidance/guides-to-good-practice/basic-components/databases-and-spreadsheets/preserving-databases-and-spreadsheets/structuring-your-data/>

PCRG 2015 *The Study of Prehistoric Pottery: General Policies and Guidelines for Analysis and Publication* Occasional Papers 1 and 2, 3rd edition. Available at:
<https://www.prehistoricpottery.org/publications> (Accessed 2 January 2023)

PCRG, SGRP, and MCRG (2016) *A Standard for Pottery Studies in Archaeology* Available at:
<https://www.prehistoricpottery.org/publications> (Accessed 2 January 2023)

University of Reading, "Accessible Spreadsheets" in *Digital Accessibility*. Available at:
<https://www.reading.ac.uk/digital-accessibility/resources/spreadsheet> (Accessed 22 January 2024)

Wilson, A. T. & Edwards, B. (2015) *Open source archaeology : ethics and practice*. [Online].
Berlin: De Gruyter Open