

A commentary of the context description dataset from Pontefract Gatehouse Community Excavations

Introduction:

The dataset discussed in this essay are the context records of the Pontefract gatehouse excavation for which the fieldwork took place in September-November 2019 and July-August 2020 (Casswell et al. 2020). The dataset is a working document that is used as an ongoing record, which can be used for analysis and interpretation. The specific uses of context record dataset include:

- Converted into tables to be used as an appendix in reports.
- Shared with specialists to contextualise finds or samples.
- Analysis tool when producing site narratives.
- Primary form of the data to be included in the archive.

The dataset was designed to be easily usable, editable and understandable by all those contributing to and reading the data. The dataset is by necessity set up to be archived (Tsang 2020) though it should not be considered in itself the final archive.

Why excel?

The dataset is structured in Microsoft Excel, this is because of the usability, exportability, accessibility, longevity and sustainability of the software. Excel is user-friendly with extensive online support. It is a program many people are familiar with, this is important because most people involved in a project will need to interact with the context record dataset in some form (Tsang and Corley 2014). Using Excel means that both internal and external staff can add to or read the dataset in a program they are familiar with and feel comfortable using. Despite being proprietary software, the Microsoft Office suite is considered a business standard, and due to its ubiquity most open source dataset software (such as LibreOffice Calc) is compatible with it. Due to this it also presents the lowest barrier of entry as the level of technical skills required is minimal and widely possessed by the professional population.

Files created in excel [.xlsx] are an acceptable format for Archaeological Data Services (ADS) as they are a stable file type, however the ADS prefer delimited text files (Niven 2011; Archaeological Data Service 2021). It is easy to export spreadsheets into delimited text files, especially if the document is well set up for this export. The potential downside to this is that if the data is stored in non-optimal ways such as formatting etc., this may not export correctly or be lost entirely (Niven 2011; Graham et al. 2020). The easiest way to avoid this is by providing clear templates. The sample dataset provided will export as a comma separated delimited text file [.csv]. Fields with commas in them (e.g. descriptions) can be placed in quotation marks to ensure the export works correctly, and columns which contain lists are separated by semi-colons. All fields are formatted as text, which helps minimise issues with exporting the data, and no formatting adds information to the data (Niven 2011).

The csv filetype also functions well as an interim storage format. It is a common result of exporting data from more complex relational databases (i.e. exporting from Microsoft access databases via PowerBI), including DigVentures' own relational database (DigVentures 2015). At the same time many databases are set up to accept spreadsheet/csv inputs and transform them into the format required for the database. This means that spreadsheets are usually the most effective way to move data from one database to another. When datasets like this are archived, the future uses of that data are unknown. Therefore, storing it in such a flexible filetype maximises the chances that it can be easily

integrated and used in future databases and datasets, the purpose, software and capabilities of which are currently unclear.

Data categories:

While it is widely considered important, the machine searchability of archived data in the heritage sector is often relatively limited. This is because archaeological data, even where it is deposited digitally, is usually stored in non-standardised templates and often using non-standardised or flexible terminology. In response the Pontefract context descriptions spreadsheet contains columns that will enhance the machine searchability. Each context has a unique identifier made up of the site code, "CXT" and the context number (which in itself is made up of the trench number and a 3 digit code). All levels of the data hierarchy (e.g. features or finds) have these unique identifiers associated with them. This allows the data to easily be integrated and queried in databases relying on strong "many to one" linkage (Binding and Tudhope 2016; Croix et al. 2019). These are particularly valuable for joining disparate data and enabling projects such as meta-analyses.

However, machine searchable fields tend to be very prescriptive and inflexible, which can be a problem when archaeology is so varied. Limiting ourselves to fit everything into neat boxes often removes the small nuances and differences that were experienced on site. This leads to a problem where many examples of datasets that strive for machine searchability are neither easily machine searchable, nor hold much archaeological meaning, a sort of half-way house that satisfies nobody (Binding and Tudhope 2016; Croix et al. 2019). Therefore, the dataset contains two types of columns, each with an express purpose, either to aid with searchability or uses descriptive language to precisely articulate the nuanced, specific and lived experience of the excavator. A good example of this in the attached dataset is the "phase" field versus the "period" field. The "period" field utilises the MIDAS Forum on Information Standards in Heritage (FISH) thesaurus classification of periods to archaeologically date the contexts. This is a widely recognised chronology that is used throughout projects and organisations in England (MIDAS Heritage 2016). As such it is interoperable with other datasets using the same classification. The "phase" field on the other hand is specific to the project and provides site specific information and its values are not directly comparable to other sites. Both fields fulfil a similar function, but one is standardised and therefore machine searchable while the other is helpful to human interpreters.

The data in this dataset is limited to only contexts and their associated data. There are several practical reasons for this. Different data requires different information and ideally each field should be fillable for each entry. This is a working document that is intended to aid analysis; too broad a dataset runs the risk of creating a largely empty dataset, making it less usable and efficient. Additionally archaeological data exists in a hierarchy, with finds being within contexts, which in turn are within features, which are in trenches, which are all within the site. Mixing different levels of the hierarchy within a dataset would either result in inconsistent treatment of different entries or information on the lower levels of the hierarchy would be lost (e.g. several finds type within a single context would be difficult to express within one row of a spreadsheet). Additionally, the clear hierarchy of the data makes it easy to link datasets once they are uploaded to a relational database.

The dataset includes a version control which is required due to it being a working document. When many different people are editing a document it is important to track what changes have been made, when these were made and who made them, particularly as the dataset sits on a shared server.

While it is important to include metadata in the document, the dataset should as much as possible be self-describing and not require reading external documents to understand what the data is. Column headers should be clear and self-explanatory, for example in the dataset the dimension headers

include the unit of measurement. Similarly within the data fields the use of abbreviations and acronyms should be minimised, for this reason “north” is used rather than “N” etc. (Tsang 2020).

Metadata:

ADS recommends the provision of three levels of Metadata (Niven 2011; Graham et al. 2020):

1. Project Level
2. Resource Level
3. Document Level

As the dataset consists of a single document, level 3 metadata is appropriate, though the full project archive would be supported by all three levels. While the first two levels are more about discoverability, the document level metadata is primarily supposed to aid the technical interpretation of the document. As such, it needs to be as specific as possible about each field and practically helpful for a human reader, professional or otherwise (Kintigh 2006; Tsang 2020). For this reason a description is provided of every field where the meaning of archaeology specific words (such as “context”) needs explaining, listed any abbreviations and expanded on how each field would be filled in (such as formatting, e.g. dd/mm/yyyy) depending on the type of data recorded.

Additionally, to help with the issues of standardised vocabulary alluded to above, links to the thesauri used for technical terms (such as the MIDAS FISH thesaurus for chronology under the “period” field (MIDAS Heritage 2016)) are included. At the moment archaeology isn’t perfectly standardised in the vocabulary used. This makes it difficult to achieve interoperability as even major heritage bodies can use the same word to mean different things (Binding and Tudhope 2016). Having metadata helps future proof the dataset. If a standardised industry thesauri is devised at some point in the future, comprehensive metadata will allow people to understand what was meant when specific terms were used. It will allow the dataset to be updated to reflect any changes in terminology and be used as part of larger datasets in the future.

It’s important that this level of metadata is present in the document from its inception (Kulasekaran et al. 2014). This will clearly provide the resources (such as thesauri or recording manuals) for those contributing to the dataset to ensure consistent meanings and vocabulary are used across the document, with those meanings being clear to everyone accessing the document. This is particularly important in a document such as context records which will be distributed amongst finds specialists (among others) to aid their interpretation of finds. This will allow anyone to look up the meaning of the vocabulary used and have the best possible understanding of the data (Binding and Tudhope 2016).

Paradata:

Another consideration for this kind of dataset is paradata. If metadata describes what the data is, paradata describes how that data came to be. This is important for external readers to help them interpret and contextualise the data (Tsang 2020). This allows them to perform their own quality control on the data, e.g. by understanding whether a measurement was taken using a hand tape, a differential GPS or a total station, as these methods provide varying levels of accuracy. However, this dataset does not require its own specific paradata as the methodology for recording contexts is included in the site recording manual. As such it is sufficient to link to this resource within the metadata, and the recording manual should be sent alongside the dataset.

Conclusion:

To summarise the principles that the dataset is built on:

- The dataset is accompanied by comprehensive metadata that outlines how every field has been compiled.
- Each context should have its own row within the dataset with a unique identifier.
- There is no formatting that is necessary to understand the data that would be lost if the data is converted to a CSV and no symbols are used that would make this conversion difficult.
- Anyone editing the file should update the version and the version control.
- Where technical vocabulary is used the relevant thesaurus is referenced or the meaning is described in the metadata.
- Columns should either be designed for machine searchability or human understanding, and both should be present in the dataset.
- The dataset should be as self-explanatory as possible.
- Paradata should be provided alongside the dataset, or relevant methodological documents (e.g. recording manual) should be referenced within the metadata.

Word Count: 1,929

Bibliography:

Archaeological Data Service 2021. Guidelines for depositors: Downloads. Available at: <https://archaeologydataservice.ac.uk/advice/Downloads.xhtml#File-level> [Accessed: 15 November 2021].

Binding, C. and Tudhope, D. 2016. Improving interoperability using vocabulary linked data. *International Journal on Digital Libraries* 17(1), pp. 5–21.

Casswell, C., Jackson, N., Jago, I., Pina-Dacier, M., Tatton, H., Ungemach, J. and Wallace, D. 2020. *Gatehouse Project, Pontefract Castle Community Archaeology Project Assessment Report*.

Croix, S., Deckers, P., Feveile, C., Knudsen, M., Qvistgaard, S.S., Sindbæk, S.M. and Wouters, B. 2019. Single Context, Metacontext, and High Definition Archaeology: Integrating New Standards of Stratigraphic Excavation and Recording. *Journal of Archaeological Method and Theory* 26(4), pp. 1591–1631.

DigVentures 2015. Digital Dig Team. Available at: <https://digventures.com/digital-dig-team/> [Accessed: 1 November 2021].

Graham, S. et al. 2020. The Open Digital Archaeology Textbook. Available at: <https://o-date.github.io/draft/book/> [Accessed: 15 November 2021].

Kintigh, K.W. 2006. The Promise and Challenge of Archaeological Data Integration. *American Antiquity* 71(3), pp. 567–578.

Kulasekaran, S., Trelogan, J., Esteva, M. and Johnson, M. 2014. *Metadata Integration for an Archaeology Collection Architecture*.

MIDAS Heritage 2016. Chronological guide for recording sites and monuments (Periods List). Available at: <http://www.heritage-standards.org.uk/chronology/> [Accessed: 15 November 2021].

2011. Databases and Spreadsheets: A Guide to Good Practice. Edited by Niven, K. Available at: <https://guides.archaeologydataservice.ac.uk/g2gp/Main> [Accessed: 15 November 2021].

Tsang, C. 2020. *Barriers to the Production of Metadata for Archaeological Datasets*. Historic England.

Tsang, C. and Corley, H. 2014. *Archaeological Digital Archiving Protocol (ADAPt) Project 6896 Digital Preservation Policy*. Historic England.