Data Archiving and Networked Services

DANS

ARCHAEOLOGICAL METADATA

Guide to entering metadata of archaeological datasets

Version 1 – July 2011 (translated to English in December 2013)

Data about data (metadata)

To ensure that other archaeologists can use the deposited data for their research project, the actual meaning of the data needs to be thoroughly documented. Aside from explicitly stating to what project the data relates, it needs to be explained *what* was documented and *how* this was done. For example; which data was produced in an analogue manner and subsequently digitised, and which data is only available as a digital file? How did the files receive their names? What kind of codes and variables were used? These examples of *metadata* supply other archaeologists with the necessary information to understand and re-use the files. Sufficient metadata will also ensure that data will not be misunderstood or used incorrectly.

Metadata exists on three levels:

1. Metadata on project level. These describe the whole research project and the context of the related files. With data deposits in EASY, the project level metadata is recorded in fields based on the international Dublin Core metadata standard during the deposit procedure. This metadata offers a structured overview of the entire project and includes specific archaeological features such as the type of site.
2. Metadata on file level. File metadata provide a description of the content and of technical aspects of each individual file. It is documented in a file list (usually a spreadsheet such as Excel), in which multiple features of a file can be documented. File-specific metadata can be technical (the file name, the file size, the software used, …) or descriptive (the file content, methodological aspects, …).
3. Metadata on the level of variables and codes. These are recorded in a codebook. Any necessary description of variables or codes is strongly dependent on the file type in which those variables or codes are used. It may further be dependent on the software and the manner by which the data is utilised. Because of the large variety, there are no rigid standards where codebooks apply; each project and/or file(type) can be supplied with its own codebook. A codebook or data dictionary is regularly already available: it may be included in the project planning documents or with the reference tables of a database. In some cases, adding a document to the dataset detailing minor additions or corrections to existing documents would be sufficient for providing a general dataset codebook. A codebook should be provided as one or more separate documents, so that other researchers can effortlessly comprehend the structure and content of a collection of data.

Metadata on file level: the extended file list

In addition to the description of the research project as a whole, it is necessary to provide (meta)description of each deposited file. This consists of both technical and content-relevant information on file level. In the list below, several features are enumerated that could be used for the description of a digital file. These metadata clarify the meaning of the original files. The more extensive the content of a file will be explained during data creation, collection or depositing, the easier it will be to use the digital file in the future. Just like with other metadata, very few fields are mandatory. The more fields, the better; but not if it requires a disproportionate amount of effort enter these fields.

The aspects below can be considered as a list of recommendations and points of concern more so than actual mandatory features to include in the file list. Not all items are relevant for each file-type, where certain features may already be derived from the file name or extension. For example, features detailing spatial descriptions may only be applicable to – and very important for – CAD-files or GIS-tables. Conversely, the type of software is not applicable to digital photos. This is because digital photos are typically saved as jpg-files which are not associated with a single software application.
It is important to include a codebook if the contents of a file cannot otherwise be fully understood. Within this codebook, the exact filename and contents should be recorded.

In the list below the fields displayed in **bold** are mandatory. An example is given within the brackets.

file_name	File name of the data file (samplelist.xls)
file_content	Description of the contents (list of soil, seed and wood samples)
data_format	General technical description (relational database)
software	Software and version used (MS-Word, version 2000)
hardware	Original system (Intel-PC)
original_OS	Original operating system (DOS)
data_collector	Specific individual responsible for the contents of the file (hired surveyor)
purpose	Specific goal for the collection of data in this file (random sample)
collection_mode	method of collection (height measurements in a grid of 5x5 metres)
analytic_units	analysis/storage of features, observations or records (description of individual features)
data_appreciation	evaluation of data quality, re-use value or limitations (5 to 10cm error margin in height measurements)
geog_cover	spatial coverage area (500 x 250m around central point 134790/352200)
geog_unit	features where spatial coordinates have been used (in kilometres)
mapprojection	name of the card projection or local measurement system (RD)
local_georef	minimum of 2 reference points for the conversion of local to national coordinates
source_document	(analogue) source of digital data (digitalised field measurements 1:50)
othmat_citation	name of the files that contain related, supplementary information (see the attached plan of approach: plan_schipluiden.doc)
othmat_codebook	exact file name of the related codebook (samplelist_code.txt)
notes	supplementary and specific instructions for (re-)use

A minimal description of files could be limited to the four features displayed in bold in the list above. However, an archaeologist would not gain a sufficient overview of the data if the description is limited to only these four features. It is therefore strongly recommended to use as many of the 19 features as possible, if applicable to the file. For an even more accurate description, the number of features can be expanded upon with elements from the full list of description elements of files (see attachment in Dutch). The use of features will depend on what the data creators regard to be important information, or whether certain features can be added with little extra time investment (i.e. automatically). The primary objective should be kept in mind: the metadata should aid other archaeologists in their re-use of files without having to contact the original creators of the dataset.

It is not uncommon for datasets of excavations to include hundreds of digital photos of ditches, sections, profiles and finds. The documentation of all these digital photos in a large file list would involve a disproportionate amount of effort. In many cases, this is not necessary, as photo descriptions are usually digitally recorded during the project. Lists produced during the excavation are typically directly or indirectly corresponding with the mandatory fields in the list above. The file name (usually including the photo number) and the file content (ditch, grid, feature, …) are in most cases adequately recorded. At the same time, the fields ‘software’ and ‘codebook’ are not applicable in this situation.

The same is valid for (scanned) field drawings (grid, ditch, profile or section drawings). The list of drawings created during the excavation is highly congruent with the relevant fields. If the file name of a scan can be added to this list, it will suffice as metadata in the file list.

Both examples show that the creation of metadata should never call for an extensive (re-)description of a large amount of files. Existing information and the smart application of name conventions of files can save a lot of work.
Some groups of files may share all of their metadata by their nature. A typical example in Dutch archaeology would be the multitude of MapInfo GIS-files for digital map drawings. Each grid, ditch and feature layer is made up out of four or five separate files (.TAB, .ID, .MAP, .DAT or .DBF, .IND). This would result in hundreds of files in a typical excavation. It is sufficient to only include the .tab-file in the metadata file list: the .tab is seen as the main file, which has its corresponding included. When naming the files, the following convention could be applied:

<projectcode>_<trench number>_<grid number>_<feature>.tab

alp02_wp012_vl01_trenchcontour.tab

alp02_wp012_vl01_features.tab

alp02_wp012_vl01_sectionlines.tab

alp02_wp012_vl01_heightmeasurements.tab

In this scheme, the content is immediately clear to the viewer. The metadata can be extracted automatically from the file name in the extended file list. Very few additional features remain to be added, and most of these can be added automatically. This can be done automatically for a whole group of files with a specific query or copy/paste command.

The same is true for technical metadata such as the file type, the software or the file size. These can, using the appropriate tools, be added (semi-)automatically to a file list.

The extensive file list will usually be created in a database or spreadsheet. Every metadata element will be added only once. The file name is the key and has only one possible unique value. However, it should be possible to add multiple persons under 'data_collector'. Most metadata systems are based on XML, where a repeat of the same element is possible. In a database or spreadsheet, this is often less easily done. The solution is separating entries with a semicolon (;), as per international standards (for example: Butler, J.; Brongers, J.A.).

Due to specific limitations of XML, special characters such as the & (ampersand), < (less than) and > (larger than) cannot be used in the file description. The same is true for quotation marks (", ', `), percentages (%), and diacritic marks such as trémas and umlauts.

The extensive file list has to be delivered digitally. An obvious choice would therefore be a spreadsheet or database table. The names of the files have to correspond exactly to the actual file names, including the case of the letters. Metadata elements have to correspond exactly to those in the list on page 2.

Metadata on the level of variables and codes: The codebook

Next to the description of each individual file, it is also necessary to provide a detailed description of the contents of each file in regards to any codes and conventions used. Such a description is provided with a document known as a codebook. The structure of the codebook can be chosen freely and can be adapted to the research project, the research plan or the computer applications used.
A codebook should explain the form and content of a specific digital file. An example of a codebook is visible on page 1 in Dutch.

Codebooks can include all names of tables in a database, the sheets in a spreadsheet, or the layers in a CAD-drawing. Subsequently, the codebook should clarify the meanings of the variable names in the database table, spreadsheet or GIS-table. The codes used within the files need to be explained in a code list or reference table. Additionally, the units used (kilometres, centimetres, litres) and/or accuracy (for example, to two decimals) need to be specified in fields with numeric features.

The file list and codebook form the foundation from which future researchers can evaluate and understand the digital files. Making an encompassing codebook will cost time and attention. This meta-information is usually already available in one form or another, or is roughly the same for different projects undertaken by the same organisation. For larger research projects, much is determined and described in advance as part of the research plan. The research plan contains a detailed description of the research methods and the (digital) end products wherein results must be recorded. A codebook may be easily created from this description, with the addition of extended documentation wherein irregularities and exceptions are recorded. However, documenting irregularities that occurred solely during fieldwork may create uncertainty, in which case it may be better to create a separate codebook for the dataset instead.

A database usually contains reference tables, wherein the codes used within the data tables are recorded and explained. The structure of tables, variables and their relationship is typically logged and displayed graphically with the database management software. This information can easily be exported to a single text document. Microsoft Access, for example, possesses a special ‘Documenter tool’. At the end of a project, the ‘end result’ of a database can be documented in this fashion.

From the perspective of a future user, it would be user-friendly to provide a codebook for each individual file. In some cases, these codebooks can be created easily. In many cases however, a group of similar files will have one common codebook, such as Mapinfo files or tables in a database. This is not an issue for archiving or re-using the data, as long as the codebooks are referenced correctly for each file in the filelist (othmat_codebook). If a codebook is used as a supplement to the research plan, it is required to add the digital version of the research plan to the archive.

Under no circumstances should a future user of the archive be confronted with a complete account of all e-mail correspondence between the client and the contractor in which small, large, temporary or definitive changes are recorded. The final fieldwork documentation should be detailed in one or two textual codebooks.

The definitive file list contains information about which codebook is applicable to which file. The codebooks themselves will be archived in a long-term preservation format, like ascii text, or a pdf document (should the mark-up of the text be important).