

Section 4. Archive Files#

4.1 Working files#

These are all the geophysics files created during the duration of the project, ranging from the raw data files downloaded from field instruments, to processed geophysical data files, their image representations, GIS files that integrate these files and interpretation diagrams. They will mostly be stored and managed in proprietary data formats that are best suited for the processing flow (e.g. ArchaeoSurveyor XML files, Geoplot composites, dxf or shp files). All these files should be included in the Archive as they can be useful for subsequent analysis and re-evaluation. They form the primary information and are equivalent to a collection of archaeological finds from an excavation. Their arrangement in hierarchical folders is discussed in [Section 4.8](#).

4.2 Preservation files#

In order to allow future users access to the data, it is important to provide them also in a format that can be used without the proprietary software with which they were created. These files are included in the Archive in a preservation format, which is well documented so that data can be accessed with other software. The Archiving Body will be able to migrate the data, which means that old file formats will be read and saved in a new format (e.g. Microsoft Excel xls files to be saved as xlsx files; or AutoCAD R13 dxf files being saved as dwfx or dwg files). Such a continuous migration of preservation files makes access much simpler and faster for potential users. The folders of the Archive that contain these preservation files should be clearly labelled, for example by the prefix 'PRESERVE'.

It is inevitable that during the export from the proprietary format to the preservation format some information is lost and it is hence essential that this is captured separately in the geophysics metadata (e.g. line spacing, orientation of subsequent lines, instrument used) as detailed in [Section 5](#).

4.2.1 Horizontal Mapping

Data from magnetometer surveys or earth resistance area surveys are usually assembled into a 'composite' by placing all collected data in their spatial raster position, usually defined in the geophysics coordinate system. The best preservation format for composite data are as XYZ text files (maybe better referred to as XYV files; for details see [Appendix 1](#)) both in their unprocessed ('raw') form and after final processing. It may even be useful to export the data grids that represent the smallest unit of field data (see [Section 2.2](#)). The units for XY and Z/V must be specified (e.g. meters for X and Y; nanoteslas (nT) for V). Exporting to XYZ text files is preferred over so called 'Z text files' (i.e. a list of only Z values) as the latter require additional information about the order of the data stream and the length of individual rows and columns.

The final processed data should also be exported as XYZ text files in georeferenced form, where XY are in map coordinates (the datum used must be specified, e.g. UTM 36N, or Irish National Grid). This allows importing the data directly into other software packages, for example into a GIS.

4.2.2 GPR

For GPR data the preferred preservation file format is SEG-Y rev 1 (see [Appendix 1](#)). The files should conform to the well-defined SEG standard (Society of Exploration Geophysicists 2002) and location information should be recorded in the sx/y and gx/y fields that specify the location of transmitter and receiver antennas separately (this information also allows establishing whether

a bistatic antenna pair was used in longitudinal or transverse arrangement). Where data are acquired in lines, SEG-Y files for individual lines are preferred over a single large data cube as file management is then considerably easier. However, duplicate or repeat lines must be clearly identified or eliminated.

4.2.3 Electrical Resistivity Imaging

Data for Electrical Resistivity Imaging (ERI) are collected as earth resistance measurements using a set of surface electrodes that are connected in turn to a current source and a measuring unit for the electrical potential, with different electrode spacing and configuration. The electrodes may either be along a straight line (usually referred to as a 2D survey) or in a pattern on the horizontal surface (3D survey). At the moment, there is no standard format for ERI data and a simple text file should be used, for example in an AMNBV format (see [Appendix 1](#)). If an inversion process is used to estimate the ground resistivity distribution that may have caused the measured ERI this can be exported as XYZV text file (see [Appendix 1](#)).

4.2.4 Interpretation Diagrams

Interpretation diagrams are usually created as overlays of the measured geophysical data and can hence be thought of as secondary data. They are often vector diagrams (lines and polygons) and standard preservation files can be used (e.g. dxf/dwg, shp, see the Guide's sections on [Vector Images](#), [CAD](#) and [GIS](#)). If the diagrams were produced using a defined coordinate system (to be specified as either being in geophysics or map coordinates) these files can be directly used with CAD or GIS software and this is hence the recommended approach. If they were only sketched over an unscaled image, relevant information has to be provided to enable the georeferencing of these sketches (e.g. scale, coordinates of several control points).

4.3 Image Files#

Although image files are only a limited representation of the original measurements (see [Section 2](#)) they can be useful to quickly locate data in the Archive, for browsing and for comparison with re-imported data ("does it look right?"). For example, complementing Geoplot composite files with image files of the same name allows to identify quickly those composites that are most useful for a particular purpose. Image files may also be integrated into a GIS instead of a direct import of the data.

In all cases it is important to supply information about the spatial extent of the image (e.g. "120 m x 60 m") or the resolution of the pixels, and the range of values it represents (e.g. "-2 nT white to +3 nT black"). Where possible georeferencing information should also be provided to facilitate integration into a GIS. This additional information can be supplied as a textual description (e.g. in a text document), as world- or other header files, or incorporated into a GeoTIFF file. Sometimes a scalebar is provided on the image but estimating the size of a survey area from an image is usually less accurate and more cumbersome than a direct provision of this information. The same applies to the inclusion of a value-scale as a greyscale or colour bar.

For the best results when converting data to images, loss-less image formats should be used (tiff, bmp, png) and not the lossy compression of jpg files. If ever these image formats were to become obsolete, they would also be migrated as part of the archiving process.

4.4 Project notes#

Field notes form an important part of the Archive and many practitioners will have experienced that referring back to these often makes understanding one's own data and possible problems in them, easier. This is no different to the case of an archaeological excavation where field notes are carefully kept. Even if the information contained within these notes is usually already captured in some metadata records it should be considered to scan paper notes and make them available as part of the Archive. An appropriate file format is pdf/a.

4.5 Project report#

Reports are often produced in common word processing or desktop publishing packages for which the files will usually remain useable for a considerable time. In addition to these, it is advisable to also include a pdf/a copy in the Archive as it can be easily read and printed, including all pictures (see also the [Guide to Documents and Texts](#) for recommended file types) and can be migrated to newer formats.

4.6 Geophysics and project metadata#

As noted above, most proprietary geophysics software packages maintain some record of the geophysics metadata but when exporting to preservation formats, this may be lost and geophysics metadata have to be created to capture this information. Similarly, many contractors and academic practitioners use databases to store information about their projects. This can be exported to form the core of the project metadata to be augmented with further information. [Section 5](#) provides a comprehensive overview of possible geophysics and project metadata. [Section 6](#) then lists subsets of this list of metadata that are used in different archiving schemes.

There is currently no database standard to store these metadata in and it is therefore recommended to create a spreadsheet, text document or XML file. It should be an aspiration of the archaeological geophysics community to develop an XML schema that could capture all this.

4.7 Geophysics georeferencing#

Although georeferencing information could be considered as metadata it is more difficult to store in a database and is hence treated separately here. At the moment there is no uniform way to provide this information and the user will have to document it as well as is possible and in accordance with their best practice. [Appendix 2](#) provides some background information and a conceptual framework.

Collecting georeferencing information can be done using tapes or a Total Station in conjunction with a reliable basemap, or directly in map coordinates using a GNSS/GPS with sufficiently high accuracy (e.g. 0.1 m). It usually involves linking features at given map coordinates to locations within the geophysics grid given in geophysics coordinates. For example, tape or Total Station measurements from/to characteristic points in the landscape that are also marked on a map could be linked to points in the geophysics coordinate system. Alternatively, GNSS/GPS coordinates can be listed for several geophysics coordinates. Georeferencing information should always be provided in a form that can be accessed with common software. For example as a list of control points (simultaneously in geophysics coordinates and map coordinates) or in the form of distance and bearing values. The units and datum used need to be specified. Implicit georeferencing information in a proprietary software format should be avoided; examples would be cal files of AirPhoto, or aux files in ArcGIS. Usually, a list of control points can easily be exported from these packages to text files.

Information is required on three aspects:

- the geophysics coordinate system (procedure used to lay out the grid; where is the coordinate origin; estimates of accuracies for re-establishing the grid)
- coregistration to site grid (coordinates of control points, both in the geophysics coordinate system and the site grid, together with their respective accuracies; geophysics or site coordinates of map-features, if coregistration with a map is required; geophysics or site coordinates of ground features used for georeferencing).
- georeferencing (description and approximate location of ground features for reference; distance to ground features for key points along the baselines; compass bearings of baselines)

4.8 File description#

All the files that make up the Archive should be listed in the *File Description Document* that outlines what the files are, explains how they are arranged in the Archive (e.g. naming conventions, or that each Geoplot composite consists of a cmp, cmd and cms file) and what they mean. This can take the form of a text document or spreadsheet that is clearly labelled as *File Description Document*.

It is essential to organise the project files well in a hierarchical folder structure so that relationships are clear and relevant files can be found easily. Although every project has its own storage strategy it is recommended to use a hierarchical folder structure, for example in the form

<Project> \ <Site> \ <Survey_Block> \ <Technique> \ <Format>.

The <Survey_Block> would be the smallest survey area for which all measurement parameters are the same (e.g. same traverse interval, same dates and hence same moisture) and for each <Survey_Block> the different survey techniques have their own <Technique> folder, which then saves data of different formats in separate <Format> folders (e.g. ArchaeoSurveyor, Surfer, XYZ). Those formats that are considered preservation formats should be clearly indicated, for example by a prefix 'PRESERVE' The survey blocks may be allocated to separate <Sites> that could be grouped under individual folders (e.g. A1 Widening, 2010 Surveys). As an example for data from the monastic grange at High Cayton, the folder structure of Table 2 could be useful.

| |
|--|
| Monastic_Granges\High_Cayton\North_Field\Mag\ArchaeoSurveyor |
| Monastic_Granges\High_Cayton\North_Field\Mag\Images |
| Monastic_Granges\High_Cayton\North_Field\Mag\PRESERVE_XYZ |
| ... |
| Monastic_Granges\High_Cayton\North_Field\GPR\Raw |
| Monastic_Granges\High_Cayton\North_Field\GPR\Processed |
| Monastic_Granges\High_Cayton\North_Field\GPR\Images |
| Monastic_Granges\High_Cayton\North_Field\GPR\PRESERVE_SEGY |
| ... |
| Monastic_Granges\High_Cayton\North_Field\GIS |
| ... |
| Monastic_Granges\High_Cayton\Kilns\Mag\Geoplot |
| ... |

Table 2: Folder structure for data from High Cayton

The relationship between different files can be captured well in such a hierarchical structure and requires only few explanatory notes in the *File Description Document* (e.g. what is the 'North Field' and where are the 'Kilns').

Some extra information is however needed to describe the individual data sets in these folders. For proprietary data it is essential to provide information about the software with which they were created (e.g. "all files in folders \GPR\Raw were collected and stored with PulseEKKO software"). It should be considered that there are many programs that produce dat files ? but these are all different and to make use of them requires additional information. It is also desirable to specify the groups of files that belong together. For example a single 'shape file' consists of at least three actual computer files (shp, shx and dbf) and Geoplot composites require three computer files as well (cmp, cmd and cms; note that with certain settings in Microsoft Windows cmd files are not shown in a folder listing and may hence be forgotten when compiling the Archive). Although such information may be difficult to obtain for proprietary data formats, it only needs to be compiled once and can then be used generically for all other projects.

Working Files included in the Archive should be explained as comprehensively as possible in the *File Description Document*. For example what do the names of the composites mean? (e.g. "HCT04P01 contains High Cayton Survey Block 04, processing step 1"). What is the naming convention of the GPR transects? (e.g. "line01 to line22 for Site 1, line23 to 26 for Site 2, with line24 being repeated and superseded by line25").

Similarly, additional information should be provided for preservation and image files. It is, for example, important to know how the composite data were exported to XYZ text files with respect to the exact coordinates (e.g. measured from the centre of a raster cell or from a corner). This information can be included in the *File Description Document* or in form of a screen capture (Figure 8) during the conversion process. For image files it is essential to record how they were created, the range of values they represent and their size (see [Section 4.3](#)).

Figure 8: Screen capture of XYZ data export.

[Previous](#) | [Next](#) | [Contents](#)