DataTrain Archaeology Module 3

Working with Digital Data

Presentation Notes

Written by Lindsay Lloyd-Smith

(July 2011)

Slide 1 Working with Digital Data

This session covers the everyday practical aspects of working with digital data.

Slide 2 Working with Digital Data

For example:

- How to best organise our files;
- Good practice in naming files,
- And what background information we need to provide for different types of digital data to be understandable and re-usable in the future.

We will also cover some basic technical issues concerning the most common digital file and format types that most of you will use in your research.

However, this technical information is not best learnt in a class room and you are advised to, in your own time, refer to these slides and the online Guides to Good Practice on the Archaeology Data Service website.

This session is intended as a 'whistle-stop' tour of the common file format types to provide an overview of the range of details that need to be thought about during data management planning.

At the end of the session you'll begin to turn some of what you written in the first two exercises into part of a more formal Data Management Plan concerning the File Structure and Naming conventions of research projects.

Slide 3 File Structure

A systematically organised file structure is really important.

While you are the one who will be working with it every day and way you organise your data might be obvious to you, the sign of a logical file structure is if it is easily understandable to others who know nothing about the research project.

Many people work closely with other researchers or as part of larger projects and a logical system helps to share and exchange data.

Think carefully about a sensible file structure. For those working with GIS for example, the file structure needs to be consistent for maintaining the retrieval of files from the Geodatabase into ArcGIS.

A way to create a logical and simple file structure is to define the 'end product' of a project.

Slide 4 Data Hierarchies

When deciding how to organise your research data it is useful to first decide what the primary data of the project is. In archaeology this often comes down to projects either being organised by:

- Material – in the widest sense, so everything from types of material culture to archaeological samples (bones, soils, genetic samples, etc).

Or,

- Location – where data are grouped by region or archaeological site.

A third possible way would be to organise the data chronologically but often there is so much temporal overlap with sites or material spanning several periods it is difficult to create distinct sets of data.

Another useful tip is to **define the 'end-product' of the research project**, that is:

- The data that will comprise the project archive and what will be made public and shared with others in the future.

And,

- Try to keep this clean of temporary folders and files.

Bear in mind that you need to be able to distinguish between different projects you are working on, and in particular, distinguish between sub-folders, which might end up with same name as other folders in other projects.

Two last things:

- It is important to acknowledge that research designs change and so must file structure.
- Some people advise against the over use of folders as it takes forever to find files, though this might be easier said than done.

Slide 5 File Naming 1

File naming should be considered from the very outset of a project:

- Names tell us what a file is – that is they contain **contextual information** about the file so we know what it is without having to open it.
- Names order files – thus making stuff easy to find.

The most important thing is to:

- Define your system - and **stick to it.**

Slide 6 File Naming 2

Some useful tips when deciding upon your own system for naming files are:

- Different data may require different naming conventions:
  - Should different data/file formats be identified as part of same project?

- File names can contain contextual information:
  - Date, Author or Initials, Site or Project, Material.

- Capitals in file names affect ordering – be consistent.

- Numbers order files only if zeros are used before units and tens:
  - 001, 002, 003, etc will order files up to 999.
  - Dates are useful for version control and ordering files.
  - YY-MM-DD (11-03-02) first in a filename orders files by date.
  - Name_YY-MM-DD orders files of the same name by date.

- Spaces between file names cause havoc in GIS. **Use_underscores**

- / Slashes / in file names can cause problems too, for example for files that will be uploaded to a website.

- CAPITALS ARE HARD TO READ!

Slide 7 Consistency?

This is a made-up dummy project as an example of common misdemeanours in file structure and naming.
What inconsistencies are there in the file structure?

- Different file structure in different years of the same project.
- Folders out of order due to inconsistent naming.
- Stray files – some including sensitive personal data which is an issue we will discuss in a later session.

Slide 8 Version Control

Being consistent with what you call files makes keeping track of which version is the most up to date and the current version of a document you are working on much easier. Particularly,

- When we get feedback from others in the form of a document with reviewer's comments or sections highlighted etc.

Or,

- When we are working on a multi-authored document and people keep sending the file back and forth.

It may sound pretty simple but you'd be surprised how many people say that they didn't have a system for naming the version of the file – until they couldn't find the right one and lost loads of work.

- Add a draft or version number to the file name and/or the date.
- Initials in file names tell people who worked on the file last.
- Another important thing is to clean out older drafts of the same data.

It is wise to keep older drafts until the final version is made but whether you want to keep old versions of files and data is debatable.

You have to ask yourself: are you ever going to look at them again?

Slide 9 File Formats – why they are important

It is probably fair to say that most of us aren't really interested in the internal workings of different computer programs. We just want to get on with our work.

However, knowing just a little about file formats is important because:

- File formats facilitate exchange of data.
- It will make life easier when you need to work on different computers / software packages
- And knowing about file formats makes you appreciate the issue of preserving digital data for future re-use.

Slide 10 File Formats – key issues

There are four key issues that affect the ease with which different files can be read on different computers with different software, and the quality of information/data contained in the file.

Proprietary versus open formats
Proprietary file formats are tied to particular software programs – e.g. MS Word. Open formats are just that, designed to be read by as many programs as possible.

Non-Standard versus ISO Standard (International Standards Organisation)
File formats that use ISO Standards in their code are obviously safer in being readable in the future.

The move from Binary to Plain Text
Over the last few years there has been a move in many file formats from Binary to Plain Text code. The benefit of so-called XML code is that it is human readable and if the file became corrupted the data stand a better chance of being recovered.

Compressed and non-compressed
To make digital images smaller in size many formats compress the digital code that makes up the image. This will nearly always result in a loss of data. Although on first compression this may not be noticeable to the eye, once compressed these data cannot by retrieved. In a long-term perspective we want to ensure that high quality data is preserved and accessible.

Slide 11 File formats – further information
There is a lot of good, in-depth, technical advice available online. This information covers all file format types, as well as file structure and naming, sensitive data and ethical issues relating to digital data. Further information can be found at:

National Digital Repositories:

- Archaeology Data Service
- United Kingdom Data Archive

Institutional Digital Repositories:

- University Research Data Management Guidelines.

Perhaps the most useful documents to seek out are the Deposit Guidelines, as these tell you what is required to archive data. If you work within these guidelines from the outset then you know you can't fail.

There are also some very useful national advisory bodies on digital data formats:

- Jiscdigitalmedia.ac.uk for all types of digital images, audio and visual formats.
- Collectionslink.org for advice of digital material in museum collections.

We will now quickly run through the most common file types: text documents, raster and vector images, CAD and GIS, spreadsheets and database files, and lastly audio and audio-visual recordings.

Slide 12 Text Files

Text files are the most common form of digital data that we use and create during research. In terms of digital formats, there are three types of document:

- Those produced on computer – Microsoft Word documents for example.
- Scans of printed material – often made into PDF documents – which can be converted into text files using Optical Character Recognition (OCR) software.
- Marked-up formats – for viewing as web-pages – such as THML (HyperText Markup Language)

The fundamental division is between those text files which are essentially a digital version of paper document, for example word documents produced on a computer for publication, and text files which are designed to be viewed as a web-page, called Markup Language text formats such as HTML or XML.

Slide 13 Text File Formats - Table

Text file formats are generally not a problem. For essential information on project data a simple Text File should be used. It's common to have so called 'Read Me' files in Text File format containing essential contextual information to accompany other digital data, say images.

.DOCX files have smaller 'footprint' size that traditional MS DOC files, which might not be such a priority for us, but in terms of digital depositories this is a serious consideration.

One thing to note is that while Open Office can open Microsoft word documents, Microsoft word cannot open OpenOffice (ODT) documents. Remember this, if you use Open Office save the files as DOC files, if you want to use them on another computer or send to other people to use.

PDFs are the standard way to archive text files – and they are aimed at cross platform document exchange, designed to retain the layout of a document. PDF documents can also store a range of embedded media including images, raster and vector images – for example, you can save Illustrator files as PDFs, and the vector information is maintained.

PDF files can also be secured to limit editing or printing.

Something called a PDF/A – Archive format, is now preferred for longer term archiving. Basically the key point about these is that **the actual file is self-contained in terms of the information regarding fonts, colour profiles, images,** etc.

PDF/A also confirms to International Standards, thus guaranteeing future software will know how to open and read the files.

PDF/A documents can be created directly in Microsoft Word 2007 files. Also, ordinary PDF documents can be converted into PDF/A documents.

PDF/A documents have smaller size than PDFs and are good for e-mailing.

Slide 14 Archiving Text Files

A general consideration for final versions of text files intended for deposition in an archive project archive is to ensure that they are complete, self-explanatory, and self-contained in terms of the data and information they contain.

Data embedded in text files (images, audio, video, tabular data, etc) should also kept as separate files in appropriate formats. This will ensure that the original qualities, e.g. image resolution, will be maintained.

External links referring to data outside the text file should not be included – even to material within the same folder, and especially to material on the internet.

Digital repositories recommend that when making PDF or PDF/A documents, you do not delete the original file format, but keep it alongside the PDF.

It is important to retain what are known as a text file's **significant properties:**

- The words and word order
- Correct script for non-English words
- Hierarchical structure of headings and sub-headings
- Formatting within the text (italics and bold, but not necessarily the font type).
- Page numbering.

Slide 15 Digital Images

Archaeology is a visual and image rich discipline. There are a range of formats with different properties, functions and capabilities which affect what they are good for in terms of the task at hand, be that dissemination and sharing of data, publication either on printed paper, as a digital pdf document, or on the internet.

Different image formats also have different qualities for long term preservation.

There are two types of digital image: raster and vector.

An example of a raster image is a digital photograph. They are composed of a matrix or grid of dots/pixels with a fixed resolution or size. Each dot/pixel contains information - the amount and range of what that information is varies between file formats.

Vector files are digital illustrations where the different parts of the image comprised of points, lines, polygons, which the software constructed using geometric mathematical formulae – for example Corel Draw is a common vector image drawing program.

Slide 16 Routes to Raster Files

First we'll cover raster files. There are different ways to create raster files. In deciding the appropriate file format it is important to think of the end purpose of the image:

- Is it for publication or reference?
- For print or screen viewing?
- Black/white, greyscale, or colour?

These factors will affect the nature and resolution of the file. For example a minimum of 300 dpi (dots per inch) is required for print publications, although some may require higher resolution, while scanning documents for reference only, for example field note books, 150 dpi is often sufficient will take shorter time to produce and result in smaller files.

Slide 17 Raster Files – Technical Stuff

If you are going to produce a lot of raster files – for example digital photographs – then it is very useful to learn what these files are made up of and how they work. This way you will appreciate the issues concerning their long term preservation.

Each pixel of a raster image contains data about the colour range – the wider the colour range the larger the file size. We won't go through all these details. Save to say that they all affect the quality of the image and the file size. Again, think of the purpose of the image and decide what factors are important. For preservation of raster images uncompressed files are preferred.

Slide 18 Raster format table

Again, we're not going to bore you with lots of details now. These slides and other information can be downloaded from the ADS website.

The important thing to remember is the difference between formats that lose data – called "*lossy*", and those which don't lose data, called "*lossless*".

Compression of raster jpeg files take place every time the file is open in a photoshop like program. The pixels in the jpeg are averaged out to reduce the file size. Over time this results in loss of resolution and a blocky appearance to the picture.

There is a version of the jpeg format, called jpeg2000 – which might become more widespread and is good because it uses lossless compression, but this format has not been taken up by camera manufactures.

If you have a fancy digital camera – you have the option of taking pictures in raw DNG format, which the recommended format for keeping an archive copy of the image.

If not, **save an archive version of jpeg**s and do any work for outputs on copies.

The red bracketed formats are the important ones to remember as these are the recommended ones by digital repositories.

Slide 19 Archiving Raster Files

To summarise the 'rules of thumb' for archiving different raster images, the three most common files to use are: jpegs, DNG files, and TIFFS.

Slide 20 Raster Files – Key Points

To reiterate some key points about rasters:

- Think of the purpose of the image.
- Document the rationale of image creation, including:
  - Why were certain image properties used?
  - Be consistent with file naming.
  - Describe content of image.
- Save high resolution raster file (TIFF) of final vector images.
- Archive original of jpeg files and work on copies to produce output images.

Slide 21 Vector Files

Because vector images represent objects as geometric entities they are scalable without loss of quality. The images can by scaled up in size without any major increase in file size – unlike raster images where the size and resolution of the image do have a major influence on the size of the file itself.

Vector images are best used for creating clean line-drawings for publication (print, digital, and for web-sites). There is a range of software for creating vector illustrations for example Corel Draw and Adode Illustration. AutoCAD is a vector program and is used to represent scalable representations of real world objects and it is often used to work with survey data of archaeological sites.

Each program works with its own proprietary file format but at the same time the files can be saved or exported as different formats for moving between software packages or saving file formats that are preferred for long-term storage.

For temporary storage or sending by e-mail, illustrations produced in vector programs can be saved as a PDF document which preserve the vector information held in the illustration and can thus then be opened by most drawing programs. PDFs can also be opened in raster illustration programs such as Photoshop.

A recommended preservation formats are:

- Encapsulated PostScript (.eps), or
- Scalable Vector Graphic (.svg).

The important thing about SVG files is that they are written in human readable XML code, and use an open standard developed by the World Wide Web Consortium. For example, the OxCal Radiocarbon Calibration program produces graphical outputs in SVG format.

Again, documentation of the layers, names, conventions used in vector illustrations should be recorded and kept in a separate file alongside the illustration file. This should be updated when these aspects of the file are amended.

Up-to-date back-up copies of the file should be saved in EPS format and replaced with latest version of the illustration file every time the file is substantially changed.

For output files save as TIFFs or PNG raster images with high print quality resolution.

Slide 22 Vector file example 2 – West Mouth, Niah Cave

Here is an example of an Illustrator file of an archaeological site. The different layers should be documented in a separate text document.

Slide 23 Vector file example 1 – West Mouth, Niah Cave

This slide shows the export function where the vector file can be saved in variety of formats, including SVG.

Slide 24 CAD & GIS – Overview

CAD and GIS are digital worlds unto themselves. Anyone working with these is advised to refer to the Archaeological Data Service Guidelines specifically on these file types. Here, the generic issues for data management in these formats will be summarised.

Both CAD and GIS are primarily used to make plans and figures (maps, site plans, building plans, etc) for archaeological publications. In creating these figures they use layers which can be switched on or off depending on what will be included in the final figure. The figures can be made to the desired scale.

The distinction between CAD and GIS formats is this:

- Computer Aided Design (CAD) was developed as a technical drafting tool for making models of precise geometric objects in the real world. The layers can be connected to data tables which include information on what the layer is – but this information **cannot** be analysed.
- Geographical Information Systems (GIS) link graphic objects (points, areas on a map, etc) to associated data tables upon which a number of analyses can be performed, for example looking at the distribution of sites or artefacts in relation to soil types, landscape aspect, or view sheds. Or, to investigate relationships between artefact distributions.

Slide 25 CAD and GIS - 2

Common Data Management Issues
The most important issues of data management in CAD and GIS are the same as digital formats, namely:

- Documentation of the methods of data caption or collection.
- Record processes carried out on the data (recorded as a log).
- Document terms and conventions used for layers.

Common Archiving Issues

- Both AutoCAD and ArcGIS use proprietary software that is new with each program version. It is important to update files to the latest version.
- The raw survey data should be kept.
- Final archive files should be saved in an exchange format, for example DXF or SVG.

Slide 26 Example of CAD file – West Mouth Niah Cave

This slide shows the same archaeological site as shown before in the Illustrator vector format example. Again, these layers need to be documented in a separate text file.

Slide 27 Example of GIS file – Ruma' Ma'on Dakah
This slide shows an ArcGIS file, with imported raster map image and vector points, their respective layers, a related data table, and image.

Slide 28 Spreadsheets and Databases: Overview

Spreadsheets and databases have different functions but essentially most people use them in similar way: as standardised formats to store, analyse, and retrieve data.

Spreadsheets were designed on paper accounting worksheets and are primarily for ordering numerical data and performing calculations on these data, then producing charts and figures from these data.

Database software are designed to store a wide variety of types of data (numerical, text, images) and provide complex search/reporting on these data.

Significant properties of both are:

- Data values themselves, and
- Structure of the tables/sheets in which the data is stored.

Slide 29 Spreadsheets and Databases: Common Data Management Issues

The key elements, characteristics of these files and processes carried out on the data need to be documented. There are three important aspects:

1. Data Consistency

- Standardised data entry is of primary importance and a method for data entry control should be used.
- Standardised terminology is necessary for re-using and carrying out searches of the data.
- If codes are used for field headers or data entered into rows, then these must be explained in a separate document saved alongside the spreadsheet/database.

2. Embedded Objects

- Embedded media (visual charts, images, etc) should also be stored in a separate folder alongside spreadsheet/database.
- For spreadsheets, embedded media may be lost when file is transferred to different format or archived.
- For databases files, embedded media are more often stored as links to external files which should be stored in separate folder alongside database.

3. Non-data content: formatting of data presentation

- Formatting that highlights certain aspects of tabular data (font size/style, cell colour, border styles, etc) may be considered important in terms of presenting the results of analyses or an interpretation of the data.
- Such formatting can be lost when migrating data, especially to plain text archive format.
- An explanation of any important formatting should be documented and stored in separate file alongside data file.
- Database input forms, search query results, and generated ‘report’ outputs, comprise non-data aspects of databases. If these are considered significant they should be documented and saved as text or image files and stored in a folder alongside database file.

Slide 30 Save screenshot of database table relationships.

For more complex relational databases which comprise a number of related tables, the structure of the database must be recorded so that the database can be reconstructed from each individual archived table.

A simple screenshot saved as jpeg image is good enough to record this.

Slide 31 Audio and Video Files

Audio and visual recordings can be important parts of an archaeological archive.

Again there is a wide range of file formats each of which have different uses.

Thinking in terms of long term preservation of such data, open and uncompressed formats are recommended by the Archaeology Data Service: the ones indicated here.

Again, think of the purpose of producing the files. It may be that a compressed copy will be needed for its viewing on a website for example.

If you will produce lots of audio/video files, then talk to a digital repository early to start off on the right foot.

Slide 32 Documenting Audio and Video Files

And, you guessed it, it's important to document your files including both technical data and contextual information.

Slide 33 Documentation and Metadata

By now you'll have the gotten to the message that it's important to document your digital data files. This information can go in a number of places:

- The methods chapter of your thesis.
- If you have appendices to the thesis, then there should be an introductory document that explains the structure and nature of the research data.
- For information about individual files, if the technical details of lots of the files are the same then you don't need to repeat it for every single file.
- For complex CAD and GIS files try to keep a log of what you do.

We will return to discuss standardised forms of metadata creation in the module on Archiving Data.

Slide 34 Selection – *Chucking stuff away!*

We tend to think that we can keep every digital piece of data: every text document, every version of every draft chapter or journal article that we write, and every digital photograph or image. However, is this a good idea?

Certainly it is a good idea to keep hold of draft versions of papers until the final version has been finished, but the question is will you ever go back and look at and have use of those drafts after that?

Unless you have a very good and consistent file naming system, which despite what we have been taking about in the session, is easier said than done, it's amazing how quickly harddrives fill up with stuff that you don't know what is, making it hard to find what you really need.

Not only is it your own data that you create, but data that people send you – particularly if you are carrying out your PhD as part of a larger project. Is it really your responsibility to keep hold of all this extra stuff?

Slide 35 Example of lots of files and folders on a Hard-drive

To show just how quickly it builds up, this is screen shot of a typical researcher's hard-drive six years after starting a Master's degree, completing a PhD, and working on an 18 month post-doc.

Some of the 42,000 files will be systems files, but nevertheless it is quite scary.

And scarier still is that this built up without any systematic data management in place.

Leading to the question: when will there be time to go back a sort out the important data from the stuff that can be chucked away?

Slide 36 Selection – *chucking stuff away!*

To avoid ending up in the same situation, some simple tips might include:

- Define the core data which will form the project archive.
- Keep the core data folders clean of other files that do not belong there.

Ask yourself the question:

- Can we keep hold of data that other people send us?

And,

- Try to chuck stuff away during the project.
- Try not to hoard multiple versions of the same file.

The important thing is to keep the data 'clean'.

If you make a copy of an image or drawing file to work on or manipulate have a naming protocol that identifies it as such, as does not confuse with the primary data file and possibly save it in a separate sub-folder

The question of **what to do with e-mails** is a tricky one. It is likely that you will work at several institutions through the course of you careers, each with a different e-mail address. While it is possible to forward e-mails from account to account, you'll very soon run out of space to store them.

Thinking longer term, what will happen to the e-mails when you retire?

E-mails can contain a lot of contextual and background data about a project that is only found there, and in this sense it is important to document as part of the research process.

One answer is for any e-mail that you think is important to archive, make a print-version PDF of the file and save in a Correspondence Folder.

Make sure you have an appropriate and logical naming system to identify when the e-mail was sent, from whom to whom, and what it is about.

Slide 37 Exercise 3: Project File Structure and Naming

The next short exercise asks you to describe the structure of your project folders, and define and describe the file naming system for each different data types.

Obviously this will evolve as you begin to work with new types of data and files during the course of your research. But if you have consciously thought through a system at the start of your project it is easier to add to it in a logical way later.

It is important to be consistent so maybe to begin by pinning this form above your desk to remind yourself.

Slide 38 Acknowledgments

Module 3 Working with Digital Data

Written by Lindsay Lloyd-Smith (2011)

Acknowledgements
This material was created by the JISC-funded DataTrain Project based at the Cambridge University Library.

Project Manager: Elin Stangeland (Cambridge University Library).
Project advisors: Stuart Jeffrey (Archaeology Data Service), Sian Lazar (Department of Anthropology, Cambridge University), Irene Peano (DataTrain Project Officer: Social Anthropology), Cameron Petrie (Department of Archaeology, Cambridge University), Grant Young (Cambridge University Library), and Anna Collins (DSpace@Cambridge Research Data and Digital Curation Officer).

Image credits:
Slides 22, 23, 26: Screenshots of Abode Illustrator and AutoCAD files of the West Mouth of Niah Cave, Sarawak, by L. Lloyd-Smith.
Slide 27: Screenshot of ArcGIS file of Cultured Rainforest Project, created by Lucy Farr and by courtesy of Graeme Barker (Cultured Rainforest Project, Cambridge University).
Slide 27: Screenshot of the structure of the Cultured Rainforest Project's database, created by Lucy Farr and by courtesy of Graeme Barker (Cultured Rainforest Project).