DataTrain Archaeology Module 1

Creating and Managing Digital Research Data in Archaeology: An overview

Presentation Notes

Written by Lindsay Lloyd-Smith

(July 2011)

Slide 1   Creating and Managing Digital Research Data in Archaeology

This course is about good practice in looking after our digital research data.

While we will be focusing upon on digital data – that is material created or manipulated on a computer as opposed to physical data (pots, stones, bones, soils, etc, and all the paper records that archaeological research generates) – later in this first session we will consider the relationship between the physical and digital data sets, particularly in terms of who 'owns' the data, where it will archived, and who will have access to it in the future.

Before we go any further though, a number of important points about this course:

1. Nobody is perfect at looking after their research data. Even if a system is in place, it is hard to keep to it all the time and files, folders, and hard drives get messy.

2. By acknowledging this fact we can learn from our own and other peoples' bad habits and try to avoid the same in the future.

3. However, this course is NOT so much about computers, software, the Internet, or technical gadgets.

4. Rather, good practice in data management is more about the principles of good working habits rather than technical stuff.

5. That said, the basic technicalities of the common file formats will be presented in a later module, BUT we will not be covering software packages.

6. Needless to say, it is highly recommended that for those of you who will work with specialist software, for example lots of digital imagery, databases, or GIS, you should learn how to use it in the best way possible, particularly in terms of any data management applications, for example the creation of metadata.

Presentation suggestion
Depending how the separate modules are put together, possibly as 2 x 2 hour classes (Modules 1-4 and 5-8) or all eight modules as a half-day workshop, the course leader may want to show a slide of how the overall course is organised.

Slide 2   Session 1 Outline

This first session covers introductory aspects of data management and gets us thinking in very general terms of the types of digital data we work with during a PhD, how best to look after our research data, and where it will go once the project is finished.

There is a presentation lasting about 15-20 minutes covering what digital data are and some general issues regarding its use and preservation.

We'll consider the types of digital data used in archaeology and whether there are any special issues to do with archaeological digital data.

That's then a good point to cover the background to these teaching sessions: why they are taking place now and the general background to wider changes in Research Data Management taking place across all academic disciplines.

At the end of the session there is a group exercise followed by a general discussion and summing up of some key points.

The emphasis of these modules is on group discussion and learning from each other, so to get the ball rolling, we'll have a quick round of the room where everybody can introduce themselves and say what they are working on.

Slide 3   Digital Research Data – *What is it?*

To start off from first principles let us consider what the word the phrase 'Digital Research Data' means?

When commencing doctoral research we are constantly reminded by our supervisors to define the data which we will study. A major box to tick on funding applications is that you've defined your data, have arranged access to it, and know what you're going to do with it.

In general terms data can be defined as "*a reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing.*"[1]

We can think of research data as "*the principal output from scholarly research that when assembled in context and interpreted expertly will produce new knowledge.*"[2]

This, by the way, is what defines a piece of doctoral research: the creation of new knowledge.

It's probably true to say that amongst the humanity subjects, archaeology, in the broadest sense, studies the widest range of data, from genes to pots, and soils to satellite photographs.

Knowledge is created, processed and communicated digitally, so even if lots of us work with hard physical data (pots, bones, stones, etc) we create masses of digital data as part of the research process, whether captioned from a physical source for example scans of field drawings or measurements taken on an assemblage of pottery and inputted into a spread sheet, or digitally created in the first instance in the case of digital photographs or read-outs from scientific analyses, for example isotope data.

So, digital research data is anything created, manipulated, or studied, on a computer in the course of research. So, e-mails are form of digital data, and even Youtube videos can be thought as data.

1. Digital Curation Centre (http://www.dcc.ac.uk/)
2. Graeme Pryor "Librarian's doing data – a paradox?" Paper presented to the Libraries@Cambridge conference, 6th January 2011. (http://www.dcc.ac.uk/webfm_send/319)

Slide 4   Some General Issues with Digital Data

There are four key issues with digital data:

Volume
The size of computer memories doubles every two years, so one might think storage shouldn't be a problem. However, as memories have increased, so have file sizes, particularly of digital images. In the long run we have to make a decision what we keep and what we throw away. How long we will keep what we keep? Should we try to keep everything, for ever?

Management and preservation
By is very nature digital data is fragile. It is often kept on a fragile physical media – magnetic tape, floppy disks, compact discs, and USB pen-drives. The second point here is whether the format the data are in will be readable in the future.

Legal
We cannot and should not ignore the legal issues of digital data within our research. While it is conceivable that we may have to deal with the issue of whether data are real or not, for example in the case of doctored photographs, but this may be taking things a little too far.

However, the authenticity e-mails is a real issue, particularly for those working in government or official bodies around the world in the course of research. The same is the case for sensitive data: both personal data on living individuals, or the locations of archaeological sites which may need to be protected from looters.

Copyright in the digital age is an issue that affects us all, particularly when it comes to putting doctoral theses online.

Re-use
To be able to best re-use digital data collected or created in the past not only do we need to be able to read the format it is stored on, but we need to know the technical details of how the data were created in the first place. This is particularly important for digital images, audio and video files.

The last issue brings us back to the beginning – what we decide to keep and archive must be understandable to others in the future. Not only must we record how the data were created, but we must also document why the data were produced and what information they contain. For example, when were digital pictures taken, why they were taken, and what they show?

Slide 5   Digital Data in Archaeology – *A special case?*

Archaeology deals with a vast range of physical and digital data. The digital data is both derived from physical sources (measurements of artefacts, bones, or scans and digital versions of field drawings and illustrations) which if worst came to the worst could with luck be re-measured or re-scanned. There is also an increasing range and quantity of data that are digitally created and exist purely in that format. Apart from all text documents prepared for publications and scholarly correspondence in the form of e-mails, there are digital images including photographs of archaeological sites and individual contexts, artefacts, and rock art drawings, as well as digital reconstructions, and images of survey data, maps, and digital files produced by a range of scientific analyses, e.g. isotopic or genetic data, etc.

Can archaeology be considered in any way a special case in terms of how we create, manage, and archive digital data?

The simple answer is no. The issues of how best to manage digital data and safeguard it preservation in the long term are broadly the same across all disciplines.

The same goes for individual archaeological projects. Even though some might think that their own project is a special case in terms of complicated digital data, or for the fact that they will produce very little in the way of digital data, at the heart of it, the same issues apply, just on a larger or smaller scale.

A key issue which does vary from discipline to discipline is that of what are private data and what are public data. This does arise in archaeology particularly in regard to sensitive data of site or artefact locations, or sensitive personal data collected during the course of a research project.

What perhaps sets archaeology apart from other disciplines is the appreciation of the historical significance of what we do. And the fact that very often, the practice of archaeology is a destructive process and the physical and digital data obtained represent a unique archive – an experiment that cannot be repeated.

Slide 6   Why Digitise Physical Data?

All academic archaeologists still work, at least in part, with traditional paper data: hand written field notes and context sheets, hand drawn field drawings, and artefact illustrations. Some still take black and white negative film and colour slide photographs.

The paper archive is still in many peoples' eyes the primary archive, and provision has to be made for where these physical data, along with the artefacts, animal bones, soil samples, etc, will be archived.

Digitisation of field documentation should be a routine procedure. Like everything that is tedious and routine it really helps to get this out of the way as soon as possible after the production of the paper material.

Ideally input field data onto a computer during field work. If left until after the fieldwork, make a full digital copy of your physical documentation as soon as you can.

The reasons for this are straightforward: it helps you tidy up your field notes; enables analysis providing searchable spreadsheets and databases; and it makes data portable, easy to share and re-use in the future.

As part of data analysis, interpretation, and presentation, we digitise most, if not all, of the paper documentation. All publication preparation is digital.

The digital archive provides a back-up of the paper archive.

However, the original material should not be discarded. In terms of volume of shelf space, field note books, drawings, and photographs do not pose a serious threat to storage space.

This sounds all pretty straightforward and easy, so why are we here telling you stuff you already know?

Slide 7   Post-Graduate Teaching in Digital Data Management

Although within university research frameworks it is publications which count and are the means by which academics gain recognition, prestige and hopefully progress in their careers, good quality publications must be built upon good data.

Good data underpins high quality research, supporting credible and – importantly – verifiable interpretations. To be able to go back and check other researchers' interpretations in the future, take work further and re-use old data, the long term preservation of data – and increasingly this means digital data – is of central importance.

In this sense, the production of well managed and accessible data sets does gain oneself academic and professional recognition and will improve your reputation.

If you become known as somebody who produces reliable and re-useable data, people are going to want to work with you.

The importance of looking after research data is increasingly being recognised by major funding bodies and one of the conditions of receiving grants is the provision for good practice in data management, preservation, and archiving.

Of course, your immediate concern is to finish your post-graduate research on time – and the primary aim of these teaching resources is to help you do that.

Slide 8   Managing Research Data: The wider context

The concern for good practice in research data management is part of a general shift taking place in academia. It is useful quickly sketch out this background, not least because it has shaped the way this course has been designed.

- Across all scientific disciplines there has been an exponential growth in the quantity of digital data, so much so that it has been proposed that some disciplines are moving into *A Fourth Scientific Paradigm* of Data Exploration.[1] The importance of data management in the sciences in reflected in a special issue of the journal *Science*[2] in February 2011.

- The rise of digital data led to a debate over who is responsible for looking after research data in the long term.

- One of the outcomes of this debate is that more and more universities are implementing Data Management Policies.

- Many universities have set up digital repositories either within their libraries or run by computing services.

- On a national level for archaeology there is the Archaeology Data Service, based at York University, which is a government-funded organisation the purpose of which is to preserve digital archaeological data for the future. It hosts a huge, open access archive of archaeology data from the UK, including both research project data sets and also thousands of rescue archaeology site reports. We will return to look at the ADS website in later.

- Data Management is also on the national educational agenda and there has been a series of national and institutional projects to provide infra-structure, resources, and training across many disciplines, including online guidelines and training resources. The DataTrain open access educational resource was part of that initiative.

1. http://research.microsoft.com/en-us/collaboration/fourthparadigm/
2. http://www.sciencemag.org/site/special/data/

Slide 9   Funding Body Requirements in Data Management

Research funding bodies require that grant applications include plans for how the data will be looked after during the life of a project and what data will be deposited at the end the project.

This is particularly the case for projects that will produce what is termed a "*significant digital resource*" – for example a large body of digital imagery or an internet/web-based scholarly resource.

The tenet behind this requirement is that research needs to be verifiable, and, for research that has been publically funded, the data produced should, at the end of project, be accessible and open access.

For archaeology, the main digital repository is the Archaeology Data Service. Many universities have institutional digital repositories that may equally be a suitable place to archive project data for example of an inter-disciplinary nature.

We will return to the implications of these requirements in a later session – but it is something to bear in mind from the outset.

Slide 10   Approach to Data Management

Everyone here already manages their digital data so in one sense this course will not be about telling you a lot of new stuff. Also, data management is not best learnt by listening to someone talk on and on.

While there will be presentations summarising the basic ideas and facts behind digital data management, more important are the group exercises in which we discuss the types digital data we work with and how to look after our data in a better way; arriving at the answers through learning from each other.

In many ways learning about data management by trial and error is part and parcel of the research process and part of the trials of doing a PhD whereby we learn from our mistakes, ask our friends and peers what works and what doesn't, and tailor things to best fit our research needs.

Key aspects to bear in mind from the outset:

First is to think of data management as part of the publication process. You need to be able to demonstrate the validity of your interpretations and the best way to do this is to make accessible the data as you publish your findings.

Second is that responsible data management is about researchers' relationships with, and responsibilities to, wider communities (research communities, the general public, and other interested groups and 'stake holders')…and the future.

Many PhDs are based on the re-use of old excavation data. Similarly, we have a responsibility to make sure that the data we produce is accessible and re-usable in the future.

Slide 11   Aims and Objectives

As well as all being much better at looking after your digital research data than no doubt I was during my PhD (and still am), the aim of these sessions is for students to:

- Understand the bigger issues relating to the management and archiving of digital data;
- Be aware of the requirements and resources of not only this university but also nationally;
- And to be prepared for data management in the real world - be that within university or professional archaeology or indeed in any discipline or job.

This last point is important because good practice in data management is an extremely useful transferable skill.

The objective of these sessions is:

- The creation of a document - Data Management Plan - in which you describe how you intend to look after your research data;
- And to have started to make plans for what will be done with your doctoral research data once you have submitted your thesis.

Slide 12   Course Outline

This course is divided into four sections:

- Following this short introduction there's a very little bit of data management theory.

- Although this sounds boring - this is a useful introduction to concepts and jargon that digital repository people use.

- **The carrot** for doing a course in data management is some really useful information and tips on how to look after data, including Intellectual Property Rights, E-Theses and supplementary data, and how to archive data.

- However, you don't get a carrot without a **stick**. To make sure that you all continue to do what you say will, the stick to this course is producing a Data Management Plan for your post-doctoral research projects.

Presentation suggestion
If Module 8 Project and Professional Data is included add the following to the slide:

5. **The Future**
          Data management on post-doctoral projects and beyond.

Slide 13   Exercise 1: Defining Research Data – Template Form.

Before a bit of data theory, a little exercise to get the mind ticking over.

Thinking about the issues that have already been covered, each student will begin by simply defining the data they work with – both physical and digital.

First, spend a few minutes individually jotting down answers on the form.

You'll be referring to what you answer on this form at the end of the course when you start writing a full Data Management Plan.

Once you've filled in the form get into groups of between three and five and for ten minutes summarise for each other the data you will work with for your doctoral research, including:
- How, at the moment, you organise your data;
- Whether you've had or foresee any management problems over the coming years;
- And, not forgetting where you think your digital data will end up?

Presentation suggestions
1. **Move onto Slide 14** as the exercise forms are given out.
2. The course leader can move from group to group to answer any questions and encourage each student to contribute to the discussion.
3. **After 15 minutes open up the discussion across the groups – move on to Slide 15.**

Slide 14   Defining your Post-Graduate Research Data

Questions to help define post-graduate research data.

Slide 15 A   Digital Research Data in Archaeology – Discussion Points

Presentation suggestions:
These discussion points are based upon my own experience of what is important in data management. However, archaeology is such a vast discipline it is likely that there will be other points people wish to raise now in a general discussion.

Outputs: publication & digital dissemination
Research outputs are all, at least in part, digital.

The relationship between the physical and digital data
Documentation within the digital archives should point to the location of the original physical data.

Archiving data to achieve dissemination
If the digital data are archived at, or soon after, publication of the research this achieves the widest access to the data behind the interpretations.

Data management as a means to facilitate easy archiving
The best way to facilitate stress-free and easy archiving of data is good planning and data management from the beginning of the project.

[Press key for lower bullet point list]

Slide 15 B   Digital Research Data in Archaeology – Key points

Three key points to take forward are:
- Think of digital data early in project planning;
- Digital data management goes hand-in-hand with research outputs;
- Make your research data understandable to others.

Maybe we should add – talk to other researchers, learn from each other, and be flexible in your data management.

Slide 16   Acknowledgements

Module 1   Creating and Managing Digital Research Data in Archaeology:
                   An overview

Written by Lindsay Lloyd-Smith (2011)

Image credits
Slide 2 Anti-clockwise from the top left image:
David Robinson: Chumash pictograph, South Central California, USA.
Matthew Fitzjohn and Gianna Ayala: Map of Torina survey, Italy.
Neol Tan: Digital photography at Angkor Wat, Cambodia.
Cultured Rainforest Project: R. Ferraby carrying out geophysical survey in Kelabit
Highlands, Sarawak Photography: G. Barker.
Fraser Sturt: 3-D reconstruction.
Cultured Rainforest Project: Excavation at Ruma' Ma'on Dakah, Kelabit Highlands,
Sarawak. Photograph: B. Nyiri
Lindsay Lloyd-Smith: Iron Age burial at Trumpington Meadows, Cambridge, England.
Excavation by Cambridge Archaeology Unit.
Lindsay Lloyd-Smith: Henge-form at Old Wolverton quarry, Milton Keynes, England.
Excavation by Cambridge Archaeology Unit.