Data management for social anthropology – a basic course

Introduction

This is meant to run as a 1.5 hours (minimum) module, aimed mainly at pre-fieldwork PhD students. It is designed as an interactive workshop, in which participants are encouraged to intervene and contribute their knowledge. A survey might be distributed in advance of the session, to find out about students' skills, requirements and interests in relation to data management (see separate document for a possible template). This will allow for better participation, and students may be asked to volunteer their practical skills to the rest of the class. A reading list is also available, with key readings starred so that students might go to the workshop prepared for discussion on selected topics. The duration might therefore change depending on the expected level of participation.

Another, more advanced module is also available. This can be run separately or attached to the basic course and deals in more detail with issues of:

- metadata, ontology, and other data organisation, with examples from repositories;
- ethics, freedom of information, intellectual property and data protection;
- sharing and dissemination tools, techniques, and experiences.

A further module, aimed at PhD students in their writing-up stage, was also developed. This contains information on issues relating to:

- long-term digital preservation and data sharing;
- funding bodies' requirements in terms of data management and sharing;
- digital repositories;
- ethics, freedom of information, intellectual property and data protection.

The three modules can also be joined together to form a one-day workshop. A list of web-based resources forms part of the information on each course, and it can be handed out to participants for further reference (see separate file). All references have been included in the handout that accompanies these notes and the related slides. It is advised that at least some of the webpages are shown to students during the workshop, but please note the importance of verifying that links still work and, more generally, to ensure that the information provided is up to date! Finally, make sure that you substitute all institution-specific references as appropriate.

[Note: italic text between square brackets is intended as notes for the instructors.]

Course Outline [slide 2]

[slide 3] N.B.: this is a participatory exercise! Please feel free to interrupt, contribute, ask questions at any point...!

[slide 4] 1. What is data management, and why is it relevant?

i. *Funding and governance* **[slides 5 and 6]**

In recent years, there has been a shift in patterns of funding and in funding bodies' requirements. Funding bodies such as the AHRC and the ESRC now require that data produced through their funding be made available to the public, and therefore introduced the concept of a 'data management plan.' For these purposes, 'data' is defined as anything 'machine-readable' – so handwritten notes, tapes, printed images and other objects are not *per se* included, though of course they might be digitized. At present, this requirement only applies at post-doctoral level and above, although it may well change in the near future to include PhDs. Either way, the idea is that researchers should be trained early on in the art of managing their data for re-use, and students are 'strongly encouraged'[1] to deposit their data and make it public (it might become mandatory in the near future, but decisions are pending).

ii. *Personal organization and skills* **[slide 7]**

Regardless of 'Foucauldian-'type governance demands, planning in advance and with long-term vision is a good idea so as to be able to produce, safely store, re-use and access one's data over time. Academics' and students' personal research experience overwhelmingly tells that learning tools and techniques early on is a great asset.

So here are some issues to take into account, and some reasons to consider data management:

- Data is increasingly produced in digital formats, its volume is increasing, and legal/ethical issues are also multiplying.

- Avoiding loss of one's own data

- Improved efficiency – not to spend time locating files and information

- Re-purposing and re-use of one's own data – or sharing with colleagues privately or in the public domain (publications with shared data have in some disciplines higher citation rates, although this does not seem to have touched anthropology)

- Increasing skills base of researcher, career development

- Allowing verification of research/research integrity

- Fulfilling mandate(s) from funding bodies and institutions

[1] cf. ESRC postgraduate funding guide, http://www.esrc.ac.uk/funding-and-guidance/guidance/postgraduates/PFG.aspx [in it, it is stated that 'ESRC-funded students are required to formally offer any data created or repurposed during the lifetime of the award to the Economic and Social Data Service (ESDS) within three months of the end of the award. The ESRC-funded students are responsible for providing these data to the ESDS for assessment, and if accepted, to ensure that they meet the requirements of the ESDS for preservation and future re-use' (p. 24). However, the requirement should be modified shortly to read as 'strong encouragement.']. This is likely to be reviewed for studentships commencing in October 2011. [*Instructors should monitor any changes in data management requirements.*] For the general ESRC policy on data, cf. http://www.esrc.ac.uk/_images/Research_Data_Policy_2010_tcm8-4595.pdf.

[slide 8] But of course there are also complications:

[slide 9] Derrida famously described the archive as a paradox, working '*a priori* against itself.'[2] Thus, the 'archive fever' must come with the awareness that you cannot preserve and document everything, once and for all. Attributing descriptions is at best a provisional exercise (ideas about 'data,' however conceived, change with time), and things perish (that's the point of having archives…!).

Furthermore, throwing things away MIGHT be beneficial! **[slide 10]** (Though if in doubt, keep it!). A few classical ethnographers claimed to have lost their field notes (Edmund Leach, Evans-Pritchard, Max Gluckman), and it is not uncommon to hear from more contemporary academics that loss of data in some cases led to intellectual progress.

Alan MacFarlane, professor of anthropology with a background in history, has this to say on the issue of archiving and its limitations:

> the best article [anthropologist R.R.Marrett] ever wrote was when he was accidentally parted from the increasingly unwieldy monster [of his slips index] (Marett, Jerseyman at Oxford, pp. 117,156). I have temporarily in my possession some of the index books with perforated slips which Sir J.G.Frazer filled in order to write The Golden Bough. Many have not been detached for filing. The great historian Lord Acton assembled boxes of cards, which are now in the Cambridge University Library. As he read more and more and abstracted many thousands of references, he wrote less and less. Much of his best writing was done earlier in his career (see Butterfield, Man on his Past, 63; David Mathew, Lord Acton and His Times, p.104).[3]

[slide 11] Similar themes are explored in essays and works of fiction (most notably in the work of Argentinian writer J.L. Borges). **[slide 12]**

[Brief discussion of items on the reading lists between participants, for example: does more information equal less knowledge or meaning, as Baudrillard and others have claimed? Does a focus on data impair creative thinking?]

So one needs to evaluate the costs and benefits of so-called 'curation' **[slide 13]** - i.e. the selection, preservation, maintenance, collection and archiving of (digital) data: it's time consuming (in the short term) and often boring; but it helps - not only for future purposes but also in analytical terms.

The workshop will touch on the following processes:

- so-called sequential tasks, such as planning; creating/capturing data; selection; archiving;
- continual tasks: management and back-up;
- occasional tasks: disposing of unwanted data; reappraising and updating a data management plan; and migrating data.

---

[2] Derrida, J. 1995 'Archive Fever: A Freudian Impression.' *Diacritics* 25(2): 9-63, p. 14.

[3] MacFarlane, A. *Paper slips to computers: Notes on setting up the 'topics' database.* Available online on http://www.alanmacfarlane.com/TEXTS/Connect2.pdf, retrieved 19 April 2011.

[slide 14] 2. What is data?

Before we deal with issues of 'data curation' we should think about what might count as data - the issue is not as straightforward as it seems, and retrospectively one might come to realize that there are many more kinds of data than anticipated, which come in all sorts of formats. This is relevant also in thinking about data management plans and data sharing.

[slide 15] [*Participants are asked to reflect on the kinds of data they anticipate collecting, and on possible issues with collection/capture, storage, migration etc., each student briefly introducing their project and expectations. This can be done in groups or individually, depending on the size of the group and amount of time available.*]

Some possible questions:
- What do you expect to get out of this?
- What kinds of data do you intend to create and capture in the field? Who creates it? In what format/through what software?
- What do you think the main issues are going to be – e.g. with respect to the process of data collection, capture, storage, preservation, sharing, intellectual property, ethics?

A list may be compiled on a flipchart, slide or text file – cf. **[slide 16]** as example

Irene Peano's PhD experience - a case study of good and bad practices. *[can be replaced with instructor's own experience]*

> During my fieldwork, I tried to take notes daily on my computer, in a single Word document, noting down what had happened, the people I met, their reflections, our interactions etc. I divided the documents into different sections, one for each month, and gave each entry a heading (with names of people, organisations, places I had visited) which I thought would help me retrieve the information later on. There are many days with no entry! But I still wrote down a few keywords, even if I didn't write things down in full. I also collected fliers, books, films and photos (in analogue and digital format), which I (sort of) indexed, not very systematically; I audio-recorded some events and some conversations (which you might want to call 'interviews'); and took handwritten notes whenever that seemed convenient and appropriate. Most of these were also listed in the electronic notes, either as keywords or in the main text, referring to the specific notebook and the themes/events. I did the same with any interviews, mentioning file name and location. Handwritten notes were taken on 'moleskine'-type ruled notebooks - I always left the right handside blank to be able to write further notes, comments, keywords later, and sketched a very rough index of (some of) the contents of the notebook on the first page. But while the electronic notes are searchable with spotlight (mac search tool) and with the 'find' function in word, these obviously aren't. I went through the electronic notes every time I wrote a report to my supervisor - approximately every three months. Though as time progressed my zeal and consistency also withered...Once back, I made several attempts to index and catalogue all my material, which however I never accomplished in any thorough or systematic form. And, like many anthropologists, I hardly looked at my fieldnotes (only to give substance to some ethnographic sketches which I had in my head, to find quotes, or to verify details). After experimenting with several different kinds of software (mainly wikis), and looking for a platform that would allow me to link text in non-linear ways (to mirror analytical processes), I settled for a database called Scrivener, which allowed me to store different

kinds of data (audio, video, images, and text) in a single place; to link between different sets of data; to search, tag, code, link and back-link; and to connect to a reference manager.

[slide 17] 3. Data creation, capture and organisation

i. Introduction

While the processes of creation and capture (as well as preservation and cataloguing) never really stop, and data is not a unified corpus with a clear temporal beginning and end (you are the binding authority of your own archive!), here we focus especially on fieldwork.

Data is fundamentally a representation of experience – in a sense, all archives are an attempt to capture, evoke and represent experience, more or less directly and ever so imperfectly. Bear in mind that a lot of anthropologists consider their memory to be the key repository of data that they draw on in their writing **[slide 18]**! So you need to make sure that you have ways ('archives') to train and prop up your memory **[slide 19]**, especially in view of long-term data analysis (which of course not everybody will end up doing!) but also to be able to flesh out your ethnography and verify details. Of course, the degree to which you rely on memory will depend to a great extent on the type of data you use, and on your theoretical/epistemological standpoint.

The question is then: what techniques make this representation most effective and durable?

Ultimately, the choice of capturing and archiving methods will depend on the practicalities of one or the other in a particular setting/circumstance; on what works for you; and on the nature of your fieldwork (whether material artefacts are an important part of the collection process, for example, as opposed to information which is already in digital format, such as datasets).

Although research practices have partly changed (for example towards large collaborative grants rather than individual research over the course of a lifetime as used to be the case), some (personal or otherwise) re-use is still desirable – if only to turn a PhD into a book.

So some form of archiving (and conversion) on-the-go is recommended! **[slide 20]** It's much easier to do it on the go than post-hoc.

<u>Analogue and digital</u>

We all use notebooks…but are they documented (e.g. index on the first or last page, index cards)? And if not, will they be readable and decipherable in a few years?

Aside from 'analogue' material, at least some of the data is likely to be in digital form. Digital technology allows for more effective forms of archiving and retrieving data, e.g.:

- portability;
- size reduction;
- increased speed of input and retrieval;
- spotlight/search tools, also for categories/metadata (data about data);
- possibility to link across different files, especially in a database;
- tagging and backlinking;
- remote access

It also allows for less bulky and more easily done back-ups (more later). Given this, and despite the fact that digital data supports and formats may be less durable, it is wise to have at least some of the important data in digital format. **[slide 21]**

[Participants may be invited to share their thoughts on how the different formats in which data is collected and stored might also affect the ways in which it is analysed – cf. Derrida, Eco, Baudrillard, Turkle] **[slide 22]**

ii. Some tips for the conversion of data into different formats **[slide 23]**

[A question for participants: do you anticipate having to convert data into different formats? If so, which? Do you have suggestions on what tools and techniques to use?] **[slide 24]**

Analogue to digital

To convert handwritten notes, printed books or images into digital formats, the simplest option is to use a digital camera **[slide 25]**, though there are issues of space (which you might be able to partially overcome by zipping files and folders) and of file format for text. Optical character recognition software (e.g. Simple OCR - free; MS OneNote; ABBYY[4]) is available for turning digital images of type-written words back into text, but there are issues of accuracy, and most tools are proprietary. Other options include scanning (but it has the same issues as photos, it is more time consuming and less practical); typing (for short texts only).

Sound recording **[slide 26]**

Digital voice recordings are recommended as they last longer than tape, and connect to a computer.

Factors to consider:

- *Recording Time*. Recording time figures always have to be taken with a grain of salt. The higher the quality of the recording, the fewer hours of recording time will be available. Be sure to check how long you can record in high quality mode (this is the quality of recording voice-to-print or voice recognition software will need for transcription).
- *Data Storage*: it's handy to have a digital voice recorder that can handle more than just voice files. Digital voice recorders with USB mass storage class support can store other types of files, such as documents and images, as well as audio. Another aspect of data storage to look for is the ability to organize your files. For file management purposes, be sure that the digital voice recorder will store voice data in a minimum of three separate folders.
- *Recording Features*: Does the digital recorder have voice activation, for instance, which stops recording automatically whenever there's a period of silence, or include cue/review features that make it easier for you to find the section of the recording you want to play back? Does the digital recorder have an LCD display and what information is displayed?

[4] For a comparative table of different software tools, see http://en.wikipedia.org/wiki/List_of_optical_character_recognition_software

- *Connectivity*: Most digital voice recorders that do allow you to transfer voice data to a PC use a USB interface, either directly or through cables.
- *Transcription*: Some digital voice recorders come bundled with voice recognition software or their own proprietary transcription software; in other cases, you will have to purchase it separately. You can also use accessories such as foot pedals and headphones if you want to use manual transcription.
- *Special Transcription Features*: some recorders allow individual user IDs. Some transcription software includes management features, such as allowing you to assign different work types to dictation, or automatically sending all your downloaded transcription to a specific e-mail address.

You can find relatively inexpensive ones (under £100) that have quite good features (though nothing advanced)[5]

Speech recognition and transcription **[slide 27]**:
Speech recognition software is available on the market. However, it is effective only for one's own voice (as dictation), and with little background noise.

Transcription guidelines:
- Make some notes on interview/recording content as it happens, and archive it if possible (i.e. also note contextual information such as date, name of interviewee(s), keywords for content)
- Weigh the costs and benefits of transcribing
    - Compared with audio content, a text transcript is searchable, takes up less computer memory, and can be used as an alternate communication method, such as for closed captions.
    - But is it necessary to have everything written down?
    - Does it make sense to transcribe as you collect narratives, or would you wait until after the end of fieldwork? (there are benefits in both)
    - And so do you need (often proprietary and expensive) specialised software to do it? This of course depends on your project.
    - If you decide in favour of extensive transcription, you might consider relying on assistants (depending on budget and scale of recorded-speech material). If you do, then also think about the following:
        - confidentiality – make sure that if there are sensitive topics being discussed in the recording, your transcriber(s) are made aware of the necessity to privacy (confidentiality form; ask transcriber to destroy files once they have completed their task)
        - give transcribers guidelines (e.g. whether you want all sounds and pauses transcribed, including things like laughter, 'umm,' ah-ha; how to deal with changes in tone and volume). If professionals, they might already have their own, which you should discuss in advance. It is a good idea to be consistent in any case, so you might want to come up with a convention yourself.
        - context – transcribers might benefit from contextual information on the issues discussed

[5] Check up-to-date online reviews for more specific and current advice: e.g. http://www.digitalvoicerecorder.info/

- pay – find out what adequate standards might be! (the same holds for translators and research assistants more generally)
- acknowledgement

Translation
Keep originals! **[slide 28]**

If you rely on interpreters/translators, the same guidelines as for transcribers apply (confidentiality; context; guidelines and conventions; adequate pay; acknowledgement)

Other forms of digital file conversion
In most cases, you will need some form of specialised software in order to do this. For a comprehensive list, see

http://dataconv.org/ **[slide 29]**

Bear in mind that some data, as well as quality might be lost!

iii. File formats **[slide 30]**

A few general rules

Whenever possible, choose open-source or standard formats as they make access and long-term preservation easier.

Beware of proprietary formats with very limited range of use and distribution, especially if they lock data and don't allow exporting it easily. If you need to work with specific formats, always keep a copy of your data in widely accessible and recognised ones too.

Migrating data across operating systems is increasingly easy, but some issues of compatibility might nonetheless arise. (For example, different operating systems impose different restrictions on length and allowed characters on filenames).

Changing computers can result in loss of data, especially if the latter is stored in proprietary, obsolete and/or incompatible formats.

UK Data Archive (UKDA) table on recommended data formats for long-term preservation of research data **[slide 31]**:
http://www.data-archive.ac.uk/create-manage/format/formats-table

This is one recommendation, but others might recommend different formats.

cf. also Cambridge University Library data management pages: http://www.lib.cam.ac.uk/dataman/ *[this can be substituted with institution-specific resources]*

iv. File naming **[slide 32]**

- try to be consistent from the outset of your project (e.g. Author-date for articles and books; names and dates for interviews; chronological order for fieldnotes and pictures – for pictures, programs like iPhoto organise pictures for you, but make sure you have them organised in the original folder too)
- make names as self-explanatory and understandable as possible;
- bear in mind issues with certain characters (as a general rule, avoid punctuation – this is especially important if you are working across operating systems)
- If you want to be sure, follow the ISO convention (abridged):
  restrict filenames to upper case letters, digits, underscores ("_"), and a dot.
  file names shall not include spaces.
  file names shall not start or end with the dot character.
  file names shall not have more than one dot.
  directory names shall not use dots at all.

- 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 digits 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) at end of name orders files of same name by year.
- Year first is good for ordering files, e.g. publication pdfs
- / Slashes / in file names can cause problems.
- Capitals may be hard to read
- batch renaming tools to rename large numbers of files all at the same time

Version control: this applies to different versions of a same document (and it's particularly important when writing articles/chapters etc., and working collaboratively). Some useful tips:

- add numbers at the end of the file name (01, 02...) and include the file name itself within the document (e.g. in the header);
- regulate permissions and access rights to files (e.g. read-only tag);
- versioning tools, such as the 'track changes' option in a word document. More sophisticated options are also available, but not recommended.[6]

v. File structure **[slide 33]**

- be consistent
- use folders, but avoid over-use
- structure folders hierarchically
- separate ongoing and completed work
- regularly select (delete, archive, rename, move…)
- backup
- document your convention

vi. Email **[slide 34]**

[6] For a comparison of different revision-control software tools, see http://en.wikipedia.org/wiki/Comparison_of_revision_control_software.

It is easy to let email get out of control, especially if you are away from a computer for a while, or have only intermittent internet access. Mailing lists are especially bad for clogging up inboxes, so before leaving think about the following:

- email storage space - if your provider does not allow large amounts of storage, perhaps it's worth activating a forwarding service to a more spacious account. However, keep in mind issues of safety, privacy and data mining, especially salient with the main corporate email providers such as Gmail, Microsoft, Yahoo…Remember you can apply for space extensions on your Hermes *[please substitute with institution-specific references as appropriate, or simply advise students to check with the university's computing service about getting an extension on space allowance]*.
- Unsubscribe from unnecessary lists that you are unlikely to read regularly
- Use folders and make sure you regularly weed out unnecessary messages (depending on space and search facilities)
- Archive old emails (email tools allow archiving to local disks too)
- Always transfer important and sensitive information out of email
- 'Attachment is the source of all suffering' (old Buddhist teaching) - limit the use of attachment to transfer information (and to backup)
- Use filters

vii. Remote access **[slide 35]**

When in the field, you may benefit from remote access to files stored away from your physical location, especially if you have issues with security and logistics. Here are a few possibilities:

- Institutional networked storage

PWF: all Cambridge university members have rights to a Public Workstation Facility (PWF) account. Whilst space is limited to 500MB, you can apply for an extension. You can access PWF remotely from anywhere on the web, relatively easily and through a number of solutions. You can also give other users rights to view/update files or folders[7] *[please substitute with institution-specific references as appropriate]*

- Virtual learning or research environments

CamTools: although CamTools is designed for teaching, the administrators will give CamTools sites to research projects or groups, if you ask. This gives you remote access to folders and files (technically no space limit, but if a site gets above a few gigabytes of files, they may start reconsidering this). It has also e-mail 'announcements' capability (good for sharing), and a wiki function. You can control permissions for specific people to access specific folders[8] *[please substitute with institution-specific references as appropriate]*

- Online storage and backup service or 'cloud computing,' e.g. DropBox[9]:

- up to two gigabytes of online file space free (up to 300 gigabytes, for a fee), increased periodically
- ability to share folders with other users

[7] http://www.cam.ac.uk/cs/pwf/
[8] https://camtools.cam.ac.uk/access/content/public/CamTools%202.5%20Administrators%20Guide.pdf
[9] http://www.dropbox.com/features

- ability to synchronise versions of your files between different devices (e.g. your laptop, your desktop, the online space)
- automatic backup
- File synchronisation avoids a lot of problems in terms of version control, especially if you work on different computers/other devices. If you don't have web access, to avoid losing track of when you modified which files where make sure you have one main location where the most up-to-date version is, and stick to it.
- Security for these established online backup and sharing services is decent, but not guaranteed, and some of the intellectual property rights agreements for the sites are a bit vague; you should encrypt your files if they contain particularly sensitive data.
- There is always the possibility that your online service will go out of business, leaving you without your important files.
- Finally, beware of the Data Protection Act. For example, the DPA may apply to the contents of an electronic address book, or email messages that have been backed up to a 'cloud' solution. The Act specifically prohibits the transfer of any personal data to a country or territory outside the European Economic Area (the 15 EU member states together with Norway, Iceland and Liechtenstein) unless that country or territory ensures an adequate level of protection for the rights and freedoms of data subjects in relation to processing of personal data. The DPA also places additional restrictions on the processing of any 'sensitive personal data' which includes information relating to race or ethnic origin, political opinions, religious beliefs, physical/mental health, trade union membership, sexual life or criminal activities. This could potentially apply to the contents of email and research data in some fields; students and researchers in the medical, social sciences and allied subjects are particularly urged to be aware of this requirement.[10]

[question for participants: do you use any platform for remote access? Can you share your experience?] **[slide 36]**

10

4. Back-ups [slide 37]

Tips **[slide 38]**:

- Always have multiple backups on different platforms/formats/services/media
- Backup regularly (preferably at scheduled times - computers might do this automatically, e.g. Time Machine on Mac)
- you might consider backing up in different formats (e.g., for long-term preservation, PDF/A is recommended)
- Select and weed files regularly
- Test your back-up at regular intervals
- For further guidance, see http://www.lib.cam.ac.uk/dataman/pages/backup.html

Options **[slide 39]**:

- External hard-drives: portable, allow for relatively large amounts of data to be stored. But they perish in a few years or less. Some formatting issues across operating systems.
- Online storage: see above (Dropbox; but also Amazon S3). Might allow for automatic backup and synchronisation; remote access. But issues of space and costs, and of security (both in terms of sensitive data and of the stability of the service itself). Of course you need a reliable internet connection.
- Networked storage and virtual learning/research environments (PWF and CamTools *for Cambridge*): some space issues, need internet access, but safe and reliable
- USB sticks: more portable than external hard-drives, might have similar issues with formatting. Easier to lose and break.
- CD-roms and DVDs: even more portable than USB sticks, might allow for less storage space, more perishable but no formatting issues per se.
- Photocopies and printouts: not practical to carry and costly to you and the environment, but can be archived before travelling.
- digital pictures/scans: space issues, and time consuming, but perhaps the easiest way to backup/represent non-digital data
- email - as we said, not recommended long-term but it might be a practical short-term, one-off solution