Advanced data management for social anthropology – an interactive workshop

Introduction

This is meant to run as a 1.5 hours (minimum) module, aimed mainly at pre-fieldwork PhD students but lending itself to participation by larger audiences (writing-p students, post-doctoral researchers…). It is designed as an interactive workshop, in which participants are encouraged to intervene and contribute their knowledge, and it presupposes basic data management skills (cf. basic module). The advanced module builds on the basic 'data management for social anthropology' course, which might be run separately or attached to the advanced course. Should the module be run on its own, rather than as a follow-up to the basic module for the same participants, a survey may 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 survey template). This will allow for better participation, and students could be asked to volunteer their practical skills or experiences 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.

This module 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, has also been 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 Content [slide 2]

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

1. Documenting your data [slides 4 and 5]

We deal with issues of **documentation**, i.e. the description of how data are created or digitised, what they mean, what their content and structure is, and whether/how they might have been modified.

This is important because of issues of long-term analysis and re-use, sharing, collaboration, and archiving.

Metadata are a subset of core data documentation, which provide standardised structured information on the purpose, origin, time references, geographic location, creator, access conditions and terms of use of data. Metadata are typically used for resource discovery, providing searchable information that helps users to find existing data, as a bibliographic record for citation, or for online data browsing.

Documentation and metadata are more easily applied if planned from the outset, although of course some interpretive categories are likely to become obsolete as your way of thinking changes. Metadata might thus be constantly changing.

Data documentation applies to both analogue and digital data. We begin by exploring some 'classical' ways of documenting data and archiving in anthropology, to then move to contemporary tools and techniques that apply to digital data.

i. Paper archives and indexes

Many among the great anthropologist of the 20th century have archived (at least part of) their data or have had their data archived by others, for public use. Some illustrious examples of anthropological paper archives (and their digital indexes):[1]

- Jack Goody's paper archive at St. John's College, Cambridge, 1950–2000 (circa) made up of c. 35 files, 50 boxes, containing notes, field notebooks, correspondence and offprints relating to Goody's researches on the Bagre, Gonja, the Lodagaa, Ghana, and Gujarat.[2]

- [slide 6] Bronislaw Malinowski's archive at the LSE. The Malinowski archive comprises a wide variety of material relating to Malinowski's life and work, including field notes; papers regarding his teaching at LSE, Yale and other universities; material relating to Malinowski's published and unpublished works; correspondence; printed material; and personal papers such as diaries and family correspondence. The archive also contains 1,000 photographs taken during Malinowski's fieldwork on the Trobriand Islands.[3]

---

[1] These can be shown to the class.
[2] http://janus.lib.cam.ac.uk/db/node.xsp?id=EAD%2FGBR%2F0275%2FGoody (retrieved 29 June 2011).
[3] http://www2.lse.ac.uk/library/archive/holdings/malinowski_bronislaw.aspx;
For a list of holdings,

Famously, his diaries were published posthumously[4] and caused great controversy for the vehement expression of his frustrations with the people and the place he was studying, which made some scholars doubt the validity of his 'scientific,' public writing. He had probably never meant for the diaries to be published – so beware!

More generally, making data public presents several challenges in ethical and methodological terms, which we will come back to.[5]

[slide 7]
[Participants discuss some issues relating to the publication of Malinowski's diaries, and the readings assigned. Possible questions for discussion: what kind of data is a private diary? Does it actually count as data, and how? Should it be archived and published? Does a diary like this undermine the rest of the research?]

- Annette Weiner also did a re-study on Trobriand Island exchange practices[6] which of course relied on Malinowski's own work (including the diaries, correspondence, and pictures). Weiner's archive (1970-1997) is held at NYU,[7] and contains administrative memoranda, correspondence, reports, field notebooks, publications, unpublished writings (reviews, course materials and dissertations), newspaper clippings and papers produced since Weiner's graduate years.

- SOAS-led research project on Indian village life, a 'restudy' of three villages in the Indian states of Gujarat, Madhya Pradesh and Orissa.[8]

- The London School of Economics Archive has an extensive anthropology section, including materials from Raymond Firth, Ernest Gellner, Phyllis Kaberry, Siegfried Nadel, Margaret Read, Audrey Richards, Charles Seligman and Isaac Schapera.[9]

---

http://archives.lse.ac.uk/TreeBrowse.aspx?src=CalmView.Catalog&field=RefNo&key=MALINOWSKI%2F3.

[4] Malinowski, B. 1967 *A Diary in the Strict Sense of the Term.* London: Routledge & Kegan Paul.

[5] This debate seems to be largely extraneous to anthropology as a discipline, whilst it has received some attention in other social sciences, most notably sociology and criminology. For some discussions in those areas, see Mauthner, N. S. and A. Doucet 2003 'Reflexive Accounts and Accounts of Reflexivity in Qualitative Data Analysis.' *Sociology*, vol. 37, 3: pp. 413-431; Ibid. 2008 '"Knowledge Once Divided Can Be Hard to Put Together Again": An Epistemological Critique of Collaborative and Team-Based Research Practices.' *Sociology*, vol. 42, 5: pp. 971-985; Parry, O. and N, S. Mauthner 2004 'Whose Data are They Anyway?: Practical, Legal and Ethical Issues in Archiving Qualitative Research Data.' *Sociology*, vol. 38, 1: pp. 139-152; Ibid., 2005 Back to Basics: Who Re-uses Qualitative Data and Why? *Sociology*, vol. 39, 2: pp. 337-342; Bishop, L. 2005 'Protecting Respondents and Enabling Data Sharing: Reply to Parry and Mauthner.' *Sociology*, vol. 39, 2: pp. 333-336; Broom, A., L. Cheshire and M. Emmison 2009 'Qualitative Researchers' Understandings of Their Practice and the Implications for Data Archiving and Sharing.' *Sociology*, vol. 43, 6: pp. 1163-1180.

[6] Weiner, A. 1976 *Women of Value, Men of Renown: New Perspectives in Trobriand* Exchange. Austin: University of Texas Press; 1986 Inalienable Wealth. *American Ethnologist* 12(2):178-183; 1988 *The Trobrianders of Papua New Guinea*. New York: Holt, Rinehart & Winston.

[7] http://dlib.nyu.edu/findingaids/html/archives/weiner.html

[8] http://www.soas.ac.uk/news/newsitem67996.html retrieved 6 July 2011.

[9] http://www2.lse.ac.uk/library/archive/holdings/anthropology.aspx

In these examples, indexes have been made in digital format or digitized, and presumably a number of people (and not necessarily the archives' authors themselves) contributed a lot of labour over many years.

[slide 8] What is an index, and how does it work?

[The floor is open for a brief discussion of what an index is]

Some examples

For personal, analogue archives, a common method of indexing is paper slips/index cards. On indexing methods for paper archives, see:

- Sir E. Frazer's index cards, as explained by Alan MacFarlane on video **[slide 9]**[10] - also an example of 'meta-digitisation.'

- Alan MacFarlane's own methodologies:
  o based on the 'one fact, one card' recording of data, which allows to store data according to a number of different criteria (chronological; according to location, or any other chosen one). Reshuffling is seen as an integral part of the analytical process.

  Similarly, Levi-Strauss' method for the structural analysis of myth[11] consisted in analysing each myth individually, breaking down its story into the shortest possible sentences, and writing each sentence on an index card bearing a number corresponding to the unfolding of the story. Each such card "will thus show that a certain function is, at a given time, linked to a given subject", that is, "each gross constituent unit will consist of a *relation*" (211). However, this definition remains unsatisfactory because the linguistic units of a lower order are also "made up of relations", and "we still find ourselves in the realm of a non-reversible time, since the numbers of the cards correspond to the unfolding of the narrative", while "mythological time" is both synchronic and diachronic.

  > 'The true constituent units of a myth are not the isolated relations but *bundles of such relations*, and it is only as bundles that these relations can be put to use and combined so as to produce a meaning. Relations pertaining to the same bundle may appear diachronically at remote intervals, but when we have succeeded in grouping them together we have reorganised our myth according to a time referent of a new nature… a two-dimensional time referent which is simultaneously diachronic and synchronic' (212).

  The "synchronic-diachronic structure of the myth permits us to organise it into diachronic sequences (the rows in our tables) which should be read synchronically (the columns)" (229). This "'slated' structure" (229), as he

---

[10] http://www.alanmacfarlane.com/FILES/frazer.htm; the clip was made as part of a class on film-making techniques: http://www.youtube.com/watch?v=gKag-d4FDTk (esp. 02'52'' to 07'18'').
[11] Levi-Strauss, Claude 1963 "The Structural Study of Myth." *Structural Anthropology. Vol. 1*. Trans. Clair Jacobson and Brooke Grundfest Shoepf. New York: Basic, 206-231.

terms it, "comes to the surface, so to speak, through the process of repetition," (229) as illustrated in the analysis of the Oedipus myth.

Here you also see how collection/capture is already a form of archiving, or at least that the two happen almost simultaneously.[12]

o a large index of slips on numerous topics arranged into a hierarchical index system, which MacFarlane started assembling since his undergraduate years, and continued with the help of his wife.[13] The hierarchical divisions were not worked out all at once, but gradually evolved, the main headings in the first five years, and then lower level ones over the next fifteen years. In this 'topics' database, the top level of the hierarchy had about twenty major headings such as:

Agriculture, Economy, Marriage, Politics, Capitalism

This slip index method was not just used for the 'Topics' database, but was later adapted for use in the study of particular communities, historical and anthropological. The top level then became:

Person, Place, Date, Source, Subject.

Under each heading would be the next level. For instance, under Capitalism would be:

'Causes of Capitalism', 'Consequences of Capitalism', 'Feudalism and Capitalism' and so on.

At the third level of the index, would be further sub-divisions, for instance under 'Feudalism and Capitalism' would be headings such as:

'Was England feudal?', 'Is European feudalism unique?', 'When did feudalism end?

Roughly there were some two thousand slips per unit of the top level, each of these divided into ten or more sub-headings, and each of these divided into ten or more with an average of twenty slips. In practice, MacFarlane explains that there is enormous variance in this four-level hierarchy.

The information was written on small slips. For reasons of economy of both space and expense, MacFarlane used reams of foolscap paper cut into blocks of rectangular slips. One ream of 500 sheets produced six thousand slips. There would be a heading, a short quotation or other

---

[12] MacFarlane, A. *Only Connect: Some reasons for using the 'one fact, one card' method*. Available Online at http://www.alanmacfarlane.com/TEXTS/CONNECT.pdf retrieved 19 April 2011.
[13] 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. As with Frazer, the archiving labour was done in collaboration with his spouse...Again a reminder of the costly, labour-intensive (and often under-valued) nature of the endeavour.

material in the centre, and the source from which the information came at the bottom.

For example, a general topic or heading might be 'Feudalism and Contract.' Then the abstracted text or quotation or idea was written in the middle of the slip, for instance:
"The master who taught us that 'the movement of the progressive societies has hitherto been a movement from Status to Contract' was quick to add that feudal society was governed by the law of contract. There is no paradox here..." (referring to Henry Maine)
There would then be the reference to the source of the idea or quotation:
author: Maitland
title: English Law, vol.1
page: 233
The small size of the slips (2.5''x 3'', or about 60x80 mm.) demanded brevity. The quotation above, with about forty words, is about the average; almost always the text would be between ten and fifty words long. The maximum would be about 100 words.

However, there are issues with these formats and techniques. Alan MacFarlane writes on 'The reasons why paper slip indexing systems collapse'[14] **[slide 10]**:
- Size (takes more time to file, the same card can go under many different headings)
- Difficult to remember the entire structure in hierarchical classification
- Shifting classification along with ideas
- Paper archives aren't portable and cannot be accessed remotely
- These issues make it more difficult to find cards in the dataset
- Rigidity of hierarchical classification
- 'By definition, all 'slips' contain a relation between at least two things, and probably many more. Hence each slip needs to go under at least two headings. In fact, it should probably go into three or four places in the filing system.' (Ibid. p. 4). This again makes it difficult to retrieve information.

Reflecting on his own card index, MacFarlane thus concludes:

The result of all these un-analysed pressures was that after assembling about 40,000 or so slips, I began to lose heart. It was usually quicker to remember an author and go to the book - to rely on human memory. I might then use the more usual method of indexing a book at the back. As long as I could remember the author, I could find the material. If I did not have the book, I abstracted it and did the same. (Ibid. p. 5)

ii. Digital archives **[slide 11]**
Some of the issues with paper/physical archives are solved by digital ones, especially those of size, portability, remote access, retrieval, rigidity. Indeed, these days most public paper archives (or at least their indexes) are digitized (as in some

---

[14]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.

of the examples cited before). While the larger the set of records existing in a database, the longer it takes the computer to index/store a new item, computers' ever-growing processing power easily keeps pace with, indeed exceeds, the needs of most social scientists - again from MacFarlane, whose 'topics' database has been converted into a web database which can be viewed at http://bamdemo.lemurconsulting.com/bamdemo/db/topics/. It contains 59326 records, based on MacFarlane's library (also indexed online at http://www.alanmacfarlane.com/FILES/library.htm).

Other examples of digital anthropological archives in the public domain:
- Pat Caplan's Concepts of Healthy Eating Food Research: Phases I and II, 1992-1996 (The Nation's Diet)[15] **[slide12]** *[instructors should show a few datasets and databases from the web]*
- P. Kaberry's Study of the Abelam of Papua New Guinea and the Nso of Cameroon, 1939-1963[16]
- Porter, G. and K. Hampshire Children, Transport and Mobility in Sub-Saharan Africa, 2007-2008[17]
- LSE anthropology photographs[18]
- Digital Himalaya: a project to develop digital collection, storage and distribution strategies for multimedia anthropological information from the Himalayan region[19]
- World oral literature project: a global initiative to document and make accessible endangered oral literatures before they disappear without record[20]

The first three items are from the Economic and Social Data Service, a national data archiving and dissemination service which came into operation in January 2003. The service is a jointly-funded initiative sponsored by the ESRC and JISC, and hosts, among others, the data produced by ESRC-funded research projects. Using the keywords 'social anthropology,' 'cultural anthropology,' 'ethnography,' their catalogue returns 64 records. The last two are hosted by DSpace@Cambridge, the institutional repository of the University of Cambridge, currently hosting the scholarly works of 17 authors and the PhD theses of 7 students in social anthropology; 83 occasional papers in sociology and anthropology; and several anthropology lectures.[21] [*should the institution at which the workshop is led have an equivalent repository or similar service provisions, these should be flagged.]*

[slide 13] For a comprehensive list of research data repositories, see http://www.datacite.org/repolist; http://oad.simmons.edu/oadwiki/Data_repositories

---

[15] http://www.esds.ac.uk/findingData/snDescription.asp?sn=5801&key=caplan+food; and http://www.esds.ac.uk/findingData/keywordListForSN.asp?sn=5801 (for keywords), retrieved 29 June 2011.
[16] http://www.esds.ac.uk/findingData/snDescription.asp?sn=5251&key=anthropology, retrieved 29 June 2011.
[17] http://www.esds.ac.uk/findingData/snDescription.asp?sn=6608, retrieved 29 June 2011.
[18] http://www2.lse.ac.uk/library/archive/anthropology_photographs.aspx
[19] http://www.digitalhimalaya.com/; http://www.dspace.cam.ac.uk/handle/1810/217843
[20] http://www.oralliterature.org/ (cf. also The Endangered Languages Archive at SOAS, London: http://elar.soas.ac.uk/)
[21] http://www.dspace.cam.ac.uk/simple-search?query=anthropology&x=0&y=0.

List of open-access digital repositories: http://www.opendoar.org/

We will come back to the issue of sharing data.

iii. Brainstorm **[slide14]**

[Participants are asked to briefly introduce their project and reflect on possible ways in which their data might be documented. Issues to bear in mind and discuss include data types and formats, different modes of documentation, software which might be helpful in cataloguing and retrieving data of different kinds, categories to be employed. The exercise can be done collectively or in groups, depending on size, and should last minimum 20 minutes. More detailed discussions and presentations to follow]

iv. Techniques **[slide 15]**

Some general points:
- techniques for documenting data can be
  - format-specific (depending on whether data is digital or not, and in what format and operating system it is created and stored, you might be able to link and back-link, search, have hierarchical or horizontal classification)
  - generic**:** things like colour-coding, categories, dates, place names…

- Reports to supervisors and/or funding bodies can act as data documentation or help in the process of documenting data.

- Keeping as much of your data and documentation in the same place helps too - so you might want to think about creating a sort of personal database.

- Specific software may prove useful (cf. handout and Software module).

- A quick run-through Irene Peano's [*can be replaced with instructor's own experience*] data archive and its evolutions:

  I started with a single word document in a Mac laptop: a journal divided into tabs, one for each month, with daily entries (and lots of gaps) whose headings were an attempt to index (mostly through names – of people, organisations or places). In the journal, I also listed any notes I might have taken in ink and paper, referring to the specific notebook and the themes/events; and any interviews, mentioning file name and location. I kept a paper folder with some fliers and other paper material, which I (sort of) indexed, not very systematically. I also tried to index ink and paper notes on-the-go, though not very systematically (by having a rough list of entries/themes on the first page of each notebook).

  I periodically revisited the journal entries - approximately every three/four months, i.e. each time I wrote a report - this helped refresh my memory and focalise issues. Reports also helped to summarise the research activities and the emerging themes.

Once back, I wrote some general reports for funding bodies which again helped summarise ideas and activities. I went through a few of the journal entries (especially the final few months) and began thinking about how to index not only fieldnotes but also bibliographic references and notes on literature. This is when I started experimenting with wikis as non-linear writing tools, and learnt more thoroughly about reference managers.

As a writing and storage/indexing tool, I finally landed on Scrivener, where I exported all the digital fieldnotes and literature. Whilst Scrivener gave me the possibility to tag, link and backlink, and to browse by category, I soon stopped worrying about those and relied on memory and search tools (in Scrivener and in OSX more generally) to retrieve information and descriptions whenever needed. The hierarchical classification of literature and notes also helped me navigate all this data.

[*run-through PhD project in Scrivener – this can be replaced by instructor's own archive run-through*]

As for reference managers, I started with EndNote and ended with Bookends, although I didn't update my library systematically until the very end, when I compiled the final draft. Both software tools connect with Scrivener. But I would not recommend Bookends or EndNote as they are not as reliable as it looks (though they do the job) - Mendeley seems like a better option. Zotero is also a good one.

At the moment, I think having everything on Scrivener helps me think about future projects, especially how to turn my PhD into a book. As time goes on, however, my memory is fading somewhat and it might be more difficult to retrieve information (especially from handwritten notes).

v. Categories and ontology **[slide 16]**

Some general points on cataloguing:
- Cataloguing is time-consuming, so evaluate the costs and benefits (as mentioned previously, many academics rely on assistants to do it if they can afford it).

- Creating metadata aids the analytical process - but you will need to keep the process going, to go back to the archive after you decide your collection phase is over.

- Categories are generally project- and focus-specific. Whilst there are systematic ontologies for the structuring of data, at most levels of research (at least in a subject like anthropology) it is easiest to devise whatever suits the researcher and the project rather than sticking to a predefined formalised scheme. As examples, take any of the online datasets cited in section ii. above,[22] or those of the Malinowski collection:[23] 41 main headings, e.g. 'Trobriand Islands fieldwork photographs,' then split into 35 sub-headings, e.g. 'sex,' which contains 16 items (mainly pictures of women…), each with a

---

[22] E.g. http://www.esds.ac.uk/findingData/keywordListForSN.asp?sn=5801

[23] http://archives.lse.ac.uk/TreeBrowse.aspx?src=CalmView.Catalog&field=RefNo&key=MALINOWSKI

description (e.g. 'Physical types [group of children]' and an individual reference number (e.g. MALINOWSKI/3/20/4).

- More importantly even, categories tend to be volatile and processual! As ideas and perspectives change…so for your own purposes it may be best to stick to generic descriptors (e.g. places and personal names, more easily remembered; dates; …).

- If you decide to catalogue your data, you will need to document and update your index: make a list of the categories you use, and systematically update as you add new ones.

- Spotlight or other search tools may be enough in some cases (long- vs. short-term perspective)

- Generally, specific ontologies are developed by data repositories and other archives. If you decided to deposit your data in a repository, they will catalogue it according to a specific system.

[slide 17] *[Participants are asked to brainstorm about how they might develop an index/ontology]*

vi. Sharing

Whilst data documentation is useful for personal access to data, it is even more relevant if data is to be shared or made public - so that other users can understand and re-use it correctly.

[slide 18] You might want to share fieldwork data with
- your supervisor;
- peers;
- other academics;
- research participants;
- wider audiences.

Of course, there are issues of privacy and intellectual property, which we will deal with in more detail later. Depending on who you are sharing with, different techniques may apply and different issues might arise.

As a general rule, if you know you are going to share large numbers of files, and especially if more than one person is likely to work on them, it is good to agree on certain rules and conventions as to the modality of sharing and the naming of files, and to find the most appropriate, effective, and supple way of sharing **[slide 19]**.

Tools useful for sharing and working collaboratively **[slide 20]**:

- Institutional networked storage

PWF:[24] public workstation facility, giving users (limited) storage space, can be accessed remotely by multiple users *[this is institution-specific - replace with appropriate, specific content, where available, and give practical examples/tours]*

- Virtual learning or research environments
CamTools:[25] online teaching and resource storage facility, managed by an administrator, with various levels of permission for different users. Has announcement, wiki and chat functions, among others *[this is institution-specific - replace with appropriate content, where available, and give practical examples/tours]*

- Dropbox:[26] up to two gigabytes of online file space free (and up to 300 gigabytes for a fee), increased periodically; ability to share folders with other users; ability to synchronise versions of your files between different devices (e.g. your laptop, your desktop, the online space); automatic backup.

- GoogleDocs:[27] especially helpful to edit documents collaboratively, to avoid using email attachments - but you need a web connection.

- Google+:[28] designed especially for online sharing, allows to create different circles

- Academic web networks: see especially the Open Anthropology Cooperative (OAC)[29] and Academia.edu.[30]

- Blogs:[31] also good for dissemination. Some famous anthropology blogs include Savage Minds,[32] Anthropologists for Justice and Peace,[33] Zero Anthropology.[34]

- Wikis (good for dissemination too). [35]

- Digital repositories: they maintain your digital files and ensure that they remain usable over time. Repositories also provide online access to papers and data for the research community. They can serve as a method of publishing files and data, making them more easily citable as well as accessible. Some communities are organised by subject (e.g. archaeological data, historical analyses, chemical data, etc.) while others are organised by institution (e.g. materials from members of a university, usually focused on publications and theses rather than data). Repositories also provide support for documenting

---

[24] http://www.cam.ac.uk/cs/pwf/
[25] https://camtools.cam.ac.uk/access/content/public/help.html
[26] https://www.dropbox.com/
[27] http://docs.google.com/
[28] https://plus.google.com
[29] http://openanthcoop.ning.com/
[30] http://www.academia.edu/
[31] http://en.wikipedia.org/wiki/Blog; for blog platform comparison, see http://www.weblogmatrix.org/
[32] http://savageminds.org/
[33] http://anthrojustpeace.blogspot.com/
[34] http://zeroanthropology.net/
[35] See fn. 20.

and annotating (metadata) and many provide additional services such as advice and assistance with data management, formats, security, and intellectual property rights concerns. Funding bodies' policies increasingly require the depositing of publications (and sometimes of data itself) in repositories such as the ESDS. Hence, if you think you are going to deposit all or part of your data, you might have to negotiate its sharing with your research subjects (seeking consent post-hoc is a real nightmare usually).

Higher Education institutions might provide digital repositories as part of their services. However, few Institutional repositories actually host research data.

The University of Cambridge's institutional repository is DSpace@Cambridge. It yields 1009 hits for a search with keyword 'anthropology' (some of them 'dark' items, not publicly available). You can deposit your thesis on DSpace, and the data supporting it (usually upon completion of your PhD). *[replace this with institution-specific information]*

However, while print theses deposited with the University Library are legally seen as unpublished manuscripts, when we put them on the web they are in fact legally published. This might pose problems in terms of negotiating commercial publishing, or the publication of sensitive/copyrighted data. More on copyright in the next section. (N.B. It may be possible to restrict access to theses for a period of time – as is the case with DSpace.)

Other examples of digital repositories **[slide 21]**: *[show as many repositories as you think fit]*
- ESDS (see above)
- UK Data Archive[36]
- EThOS: an online database of theses,[37] which provides a searchable gateway to new non-embargoed e-theses from contributing universities, a list of which is available on their website. They also scan existing paper theses manuscripts on request.

Posting things on CDs/DVDs might be a good idea for infrequent sharing of large amounts of data. Beware of security issues, which can be sidestepped by encryption (more later); and of decay/damage.

Whilst not ideal (attachments clog inboxes and may cause problems when downloading them), email is convenient. If you do use it, make sure one person is responsible for keeping track of the latest version of a file; beware of security issues; delete messages from your mailbox.

---

[36] http://www.data-archive.ac.uk/
[37] http://ethos.bl.uk/Home.do;jsessionid=4E4329AACA0E77CF0F0C390C3A8648C5

2. Ethical and legal issues [slides 22 and 23]

The dissemination of data comes with thorny ethical issues, and potentially relates to legal ones as well. A lot of the data produced through fieldwork is in one way or the other 'sensitive,' if only because it contains personal information concerning research subjects and the researcher her/himself. Therefore, when creating and storing data one needs to evaluate possible risks and consider taking appropriate measures. Ask yourself who might have access to your data (on your computer, USB stick, online cloud, notebooks…).

i. Risks and issues in data dissemination **[slide 24]**

Risks **[slide 25**]:

- Online storage: clouds and emails are handy, but potentially unsafe. Security for 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 (more later). 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 (see below).
- We also mentioned the issue of sharing and consent: if you plan to share data, you should inform research participants, ideally as early as possible, and get their permission.
- Crossing borders: under some legislation such as that of the US (of the 'sovereign exception' kind) electronic equipment can be randomly seized and searched at international borders, in view of 'terrorist threats.'[38] More widely applicable, international agreements that purportedly seek to combat 'counterfeiting' are being negotiated, which would have similar consequences.[39] Depending on which borders you cross, you might need to take this into consideration.

Different ethical issues push for sharing or withdrawing data. Here is what the Association of Social Anthropologists of the UK and Commonwealth (ASA) ethical guidelines say on the issue **[slide 26]**:[40]

(3) Sharing research materials: Anthropologists should give consideration to ways in which research data and findings can be shared with colleagues and with research participants:
(a) Research findings, publications and, where feasible, data should be made available in the country where the research took place. If necessary, it should be translated into the national or local language. Researchers should be alert, though,

---

[38] http://en.wikipedia.org/wiki/Border_search_exception;
[39] http://en.wikipedia.org/wiki/Anti-Counterfeiting_Trade_Agreement#Border_searches
[40] http://www.theasa.org/ethics/guidelines.shtml. Please note that at the time of writing this module, the guidelines are being updated, so refer to the website for the latest version. Also see the American Anthropological Association guidelines and resources on ethics at http://www.aaanet.org/committees/ethics/ethics.htm

to the harm to research participants, collaborators and local colleagues that might arise from total or even partial disclosure of raw or processed data or from revelations of their involvement in the research project;

 (b) Where the sharing with colleagues of raw, or even processed, data or their (voluntary or obligatory) deposition in data archives or libraries is envisaged, care should be taken not to breach privacy and guarantees of confidentiality and anonymity, and appropriate safeguards should be devised.

(4) Collaborative and team research: In some cases anthropologists will need to collaborate with researchers in other disciplines, as well as with research and field assistants, clerical staff, students etcetera. In such cases they should make clear their own ethical and professional obligations and similarly take account of the ethical principles of their collaborators. Care should be taken to clarify roles, rights and obligations of team members in relation to matters such as the division of labour, responsibilities, access to and rights in data and fieldnotes, publication, co-authorship, professional liability, etcetera.

So, to share or not to share? Ultimately, this is a context-specific choice which might have to be re-negotiated in time.

ii. Some tips and techniques to make data safer **[slide 27]**

- For important documents and to enter your hard-drives, restrict access by setting up passwords and access permissions (e.g. no access, read only, read and write, administrator-only permission).
- Make sure you log out of websites (e.g. social networks, online banking, email…) and of your hard -drives if you aren't using them
- Use firewalls and anti-virus software (esp. for Windows)
- Always keep multiple copies, in different locations and formats
- Destroy data when needed - both hard- and e-copies (n.b.: data deleted from a hard-drive might still be retrievable, as it is from emails etc. - which may be good or bad news)
- Encryption: this is a more secure way of both storing and transferring data. Software is available to perform this, both proprietary (e.g. PGP) and open source (e.g. GnuPG).[41] In transmission, a key and passphrase are used to digitally sign each encrypted file and thus allow the recipient to validate the sender's identity.
- Tiered consent: A good way to navigate the ethical complexities involved in ethnographic fieldwork is to negotiate a 'tiered' type of consent to participating in research, where possible. This involves asking subjects to consent to
    - participating in the study; having notes taken of their speech;
    - having speech recorded; having their names published;
    - having information concerning them anonymously published;
    - having information concerning them published with their names.

    Consent needn't be given to all levels - this might help you secure collaboration even when subjects may not be comfortable with use of personal information in some aspects of your research process. Of course, this is not

---

[41] http://en.wikipedia.org/wiki/Encryption_software

always and in all cases feasible or possible - once again, the evaluation of specific situations is up to you.

-   Anonymisation: another solution to ethical issues, especially where consent may be difficult to negotiate (as when 'data' is used from experience, from serendipitous/chance encounters, which were not formally envisaged as fieldwork in advance, for example - encounters on the bus, in bars, shops…you name it) or when participants specifically ask for it.

    A few tips:
    -   Replace names with pseudonyms
    -   Remove or alter other direct or indirect identifiers (e.g. date and place of birth and residence; age; occupation). You can often convey the sense of someone's life story, or of a situation, even if some details are altered (e.g. someone selling okra in the market might be presented as selling cooked food; a migrant who moved to Germany might be said to be working in the Netherlands; age can be kept vague, e.g. a woman in her 20s). It might be difficult to conceal some locations if the events taking place or the people involved are unique in some way (so in the case of my fieldwork, Benin City is the hub of trafficking in Nigeria and there would not be much point in withdrawing the place name. But other strategies could be used to conceal the identity of individual participants).
    -   Be consistent in the use of pseudonyms, and keep track of them (in separate place other than the data file itself)
    -   When possible, negotiate in advance with subjects what needs to be concealed/altered (such that perhaps names or other details might be omitted by subjects in interviews themselves, for example). However, you might want to consider long-term re-use of data: some contextual information might be useful in the future and best collected and kept, if in restricted access.
    -   If no other strategy works, consider restricting access to the data/thesis
    -   You may not need as many alterations as you think!

    For further advice, see http://www.data-archive.ac.uk/create-manage/consent-ethics/anonymisation

iii. Data Protection Act 1998 **[slide 28]**

The Data Protection Act 1998 was formulated in response to concerns about the amount of personal information - and the accuracy of such information - being stored, processed and passed on by organisations. It provides the legal basis for how organisations handle information relating to living people (personal data).[42]

A distinction is made between 'Personal Data' and 'Sensitive Personal Data' of living individuals.

---

[42] For further advice, see http://www.jisclegal.ac.uk/LegalAreas/DataProtection.aspx. For the full text of the Act, see http://www.legislation.gov.uk/ukpga/1998/29/contents. See also the more recent Data Sharing Review at http://www.justice.gov.uk/reviews/datasharing-intro.htm

- Personal Data relates to living individuals which identifies them: name, age, sex, address, etc.
- Sensitive Personal Data is data that may incriminate a person: race, ethnic origin, political opinion, religious beliefs, physical/mental health, sexual orientation, criminal proceedings or convictions.

Once a person is dead, the DPA does not affect their personal data.

A different level of security applies when dealing with Sensitive Personal Data: 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.

The Act establishes that:

- Data may only be used for the specific purposes for which it was collected.

- Data must not be disclosed to other parties without the consent of the individual whom it is about, unless there is legislation or other overriding legitimate reason to share the information (for example, the prevention or detection of crime). It is an offence for Other Parties to obtain this personal data without authorisation. Consent forms should be made and filled and signed participants for these types of personal data to be included in project archives.

- Individuals have a right of access to the information held about them, subject to certain exceptions (for example, information held for the prevention or detection of crime).

- Personal information may be kept for no longer than is necessary and must be kept up to date.

- Personal information may not be sent outside the European Economic Area (the 15 EU member states together with Norway, Iceland and Liechtenstein) unless the individual whom it is about has consented or adequate protection is in place, for example by the use of a prescribed form of contract to govern the transmission of the data, or if 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.

- Subject to some exceptions for organisations that only do very simple processing, and for domestic use, all entities that process personal information must register with the Information Commissioner's Office.

- The departments of a company that are holding personal information are required to have adequate security measures in place. Those include technical measures (such as firewalls) and organisational measures (such as staff training).

- Subjects have the right to have factually incorrect information corrected (note: this does not extend to matters of opinion)

Anonymised or aggregated data is not regulated by the Act, providing the anonymisation or aggregation has not been done in a reversible way. The Act applies only to data which is held, or intended to be held, on computers ('equipment operating automatically in response to instructions given for that purpose'), or held in a 'relevant

filing system'. The DPA may apply to the contents of an electronic address book, or email messages that have been backed up to a 'cloud' solution.[43]

iv. Freedom of Information Act 2000 **[slide 29]**

This act was established to increase transparency in the public sector. It gives people the right to request access to recorded information held by public sector organisations or be informed whether information is held. Research data can be requested under the Freedom of Information Act, but copyright to such data stays with the original researcher.[44]

Exceptions exist to the Act, such as:

- personal data cannot be requested
- information that is accessible by other means e.g. via a website
- information intended for future publication
- information that is subject to a confidentiality agreement, such as in a signed consent form or sensitive data held under restricted access by a data archive

Any person can request any data held by public authorities – including universities. The data does not have to have been produced by the university: it is the fact they hold the data that is important. A request must specify what data are sought.

This is a potential issue for collaborative projects where multiple copies of data are held in different institutions and countries.[45]

v. *Intellectual Property and copyright* **[slide 30]**

"Intellectual property rights… are rights granted to creators and owners of works that are the result of human intellectual creativity"[46]

Different forms of intellectual property are regulated by law:

- Copyright: this refers to creative works fixed in material form. Under the Copyright, Designs and Patents Act, 1988[47] copyright applies to (among other things):
  - original literary, dramatic, musical or artistic works
  - sound recordings, films, broadcasts or cable programmes
  - the typographical arrangement (layout) of publications
  - teaching materials and blogs

---

[43] For further guidance, see http://www.ico.gov.uk/for_organisations/data_protection/the_guide.aspx. Institutions should also be able to offer support in relation to this and other legal requirements.
[44] For the full text of the act, see http://www.legislation.gov.uk/ukpga/2000/36/contents
[45] For further guidance on FoI and research data, see http://www.jisc.ac.uk/publications/programmerelated/2010/foiresearchdata.aspx
[46] www.jisclegal.ac.uk
[47] For the full text of the legislation, see http://www.legislation.gov.uk/ukpga/1988/48/contents

Most research outputs such as spreadsheets, publications, reports and computer programs fall under literary work and are therefore protected by copyright. Facts, however, cannot be copyrighted. Data are not covered by copyright, but the arrangement of data in a spreadsheet or database is. More generally, copyright protects the expression of an idea, not the idea itself. Copyright does not require registration. In the act of creating a piece of work, writing something down, or recording an interview or song, the creator(s) of the work by default hold the right to copy the work in the future.

The ownership of copyright is not the same for all creators of work, it depends on their academic status (e.g. students or lecturers) and employment position. Students are not employees so they enjoy copyright in their own work, and some universities allow their academics and researchers the rights to their works.[48] Of course things can become messier when students are employed on projects and there are external funders or partners involved in projects. Different institutions have different copyright clauses in their employment contracts.

A copyright owner has the right to control the copying, adaptation, publishing, performance and broadcast of the work, and under what conditions this may be done. These conditions may involve payment of a royalty or licence fee. The owner may also give or sell some or all of the rights to others. In addition the author of a copyright work has certain "moral rights" that always remain with the author. These are the right to be identified as the author of the work, the right to object to derogatory treatment of their work and the right to object to false attribution of a work. However, these rights do not exist where copyright in a work has been originally owned by the author's employer. The onus of responsibility lies with the user of a work to get permission, even if the rights holder is unknown or cannot be traced.

The right to use copyright material is typically obtained:

- with the permission of the copyright owner, as set out in the terms of a licence (such as the work's terms of use), or through a licence issued by a collective licensing society which has authority to issue a licence on behalf of the copyright owner

- by seeking and obtaining permission directly from the copyright owner

- by means of an assignment (assignation in Scotland) of copyright in writing from the copyright owner.

A licence gives someone permission to do the acts which the copyright owner is entitled to authorise or prohibit without infringing copyright. This is how a great deal of material is lawfully used in the education context. In addition there are certain very specific situations where it may be permissible to make use of someone else's copyright protected works without seeking permission from the owner. For example, it is not necessary to get permission in order to use an insubstantial part of a copyright protected work.

Even if material is available on the Internet, permission will still be required in order to reuse the material (such as copying it, adapting it or dissemination of it by a different means or in different formats). Some websites may give information about

---

[48] This is the case for Cambridge. *[please replace with institution-specific information]*

the permissions (licence) which is granted to users, which will clarify what can and cannot be done with material.

There are a number of exceptions in copyright law which allow limited use of copyright works without the permission of the copyright owner. In the education context relevant exceptions include:

- fair dealing for non-commercial research and private study, criticism and review
- non-exact copies of works for teaching purposes in educational establishments (such as copying material by hand)

Use of a copyright protected work without its owner's permission may be a civil infringement and/or a criminal offence depending on the circumstances. Copyright is infringed if a person does (or authorises another to do) any of the exclusive acts restricted by copyright without the permission of the owner, in relation to the whole or a substantial part of a copyright work. What amounts to a substantial part is not defined in law but it is quite likely that even a small portion of the whole work will still be a substantial part.

Copyright is essentially a private right so it is generally for the rights holder to decide what to do when his or her copyright is infringed. The infringer could be taken to court and can run the risk of having to pay compensation to the copyright owner. They could also face:

- having an injunction (interdict in Scotland) taken out against them to stop use of the material
- being ordered to surrender the copyright material to the copyright owner
- an order requiring that infringing goods be destroyed or delivered up to the copyright owner, and that any resulting profits from the infringement are paid to the copyright owner.

Where deliberate infringement of copyright is undertaken as part of a trade or business, it may be a criminal offence, punishable by an unlimited fine and up to ten years' imprisonment.

The duration of copyright may depend on whether a work is published or unpublished, whether the creator is known or unknown, and whether transitional arrangements from previous copyright legislation apply. However, in general terms, following the end of the year of the death of the creator(s), copyright lasts:

- 70 years for literary, dramatic, artistic and musical works, films and video recordings
- 50 years for sound recordings and broadcasts
- 25 years for the typographical arrangement of published editions.

Other forms of intellectual property include:

- Designs: Appearance and shape of product
- Patents: Inventions – things that make things work
- Trade marks: Signs that distinguish goods and services
- Moral rights: Right to be attributed for your work or right to object to derogatory treatment of your work.

Intellectual Property Rights can be bought, sold, rented, gifted and bequeathed. Different countries have different copyright law.

[slide 31] *[In light of what has been discussed, participants are asked to go back to issues of sharing and publishing in ethical, political and analytical terms – cf. also the reading list, especially De Nicola 2011a,c]*