AMERICAN ANTHROPOLOGICAL ASSOCIATION
Advancing Knowledge, Solving Human Problems

Bringing Digital Data Management Training into Methods Courses for Anthropology

Cultural Anthropology: Principles and Practices of Digital Data Management

Kathryn Oths

2016

Recommended citation:
Oths, Kathryn. "Cultural Anthropology: Principles and Practices of Digital Data Management." In *Bringing Digital Data Management Training into Methods Courses for Anthropology*, edited by Blenda Femenías. Arlington, VA: American Anthropological Association, 2016.
http://www.americananthro.org/methods

Bringing Digital Data Management Training into Methods Courses for Anthropology
is a set of five modules:

General Principles and Practices of Digital Data Management
Archaeology: Principles and Practices of Digital Data Management
Biological Anthropology: Principles and Practices of Digital Data Management
Cultural Anthropology: Principles and Practices of Digital Data Management
Linguistic Anthropology: Principles and Practices of Digital Data Management

Project support: National Science Foundation, Workshop Grant 1529315; Jeffrey Mantz, Program Director, Cultural Anthropology

Organization

I. Review of material from “General principles and practices” module
II. Why is it important to preserve and protect your data?
III. Data types in cultural anthropology
IV. Managing data
V. Software
VI. Data archiving
VII. Exercises
VIII. References
IX. Acknowledgments

Review of material from “General principles and practices” module

- What are data?
- What is data management?
- What are the advantages of making data accessible?
- What are ethical dimensions of data management?
- What is a data management plan?

Why is it important to preserve and protect your data?

- Data collected in cultural anthropology research represents the cultural expressions and diversity of a people.
  - Because all cultures always change, the anthropologist is capturing a unique moment in time.
- Collecting data is a tremendous privilege.
- Anthropologists have an ethical obligation to protect the data they collect.
- Without advance planning to preserve and protect, valuable data may be lost.

In one sad case, 40 years of ethnographic research notes ended up in a dumpster upon the anthropologist's death because no provision had been made to archive them.

Photograph by Christine O. Masson and Tracy Jaeger. Used with permission

Ethical dimensions of cultural anthropology data collection and management

- Data collected today may be all that is available to future anthropologists.
- Data preservation and protecting the confidentiality of respondents are equally important.
- Anthropologists must negotiate shared access to the products of their research.
- Research participants must be informed about data archiving and access, and about participant identification.
- Areas of interaction with Institutional Review Board (IRB) and community of study at the earliest stages of research design:
  - Informed consent for archiving and sharing of data
  - Negotiation of access and availability of data with research subjects and/or community
  - Issues of intellectual property

Data types in cultural anthropology

- What is the nature of data?
  - Office of Management and Budget definition of research data: "the recorded factual material commonly accepted in the scientific community as necessary to validate research findings." https://www.whitehouse.gov/omb/circulars_a110#36, Subpart B .36(d)(2)(i)
  - In general, this means visual, textual and numerical data.

- What is metadata?
  - "Metadata is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource" (NISO 2004).
  - In short, it is data about data.
  - Metadata is necessary for a third party to make sense of your body of work: the overall organizational structure of, and relationships among, the various data files, data forms, and data sets you have created.

Data types in cultural anthropology

- Field notes
- Interview transcripts
- Audio recordings (tapes, discs)
- Visual recordings (films, photographs, videotapes)
- Letters
- PDFs, other document forms
- Drawings

Margaret Mead and Gregory Bateson working in the mosquito room, Tambunam, 1938

http://www.loc.gov/exhibits/mead/images/mm0211bs.jpg
Photograph by Gregory Bateson

Data types: “Born digital” compared to “made digital”

- Made digital: data that were collected in other than digital formats.
  - Made digital data require labor-intensive conversion into digital formats.
  - Specialists at the National Anthropological Archives devote considerable time converting hand-written data of earlier scholars to PDF and creating metadata forms
  - Photographs are converted using standard formats; see IPTC Metadata Standards.
- Born digital: With the new technologies available, research can be designed as digital from the very beginning of the process.
- Get in the habit of backing up: No matter what information you collect, be sure there is a secure, second copy somewhere.
- No fear: Sharing data may seem like a new concept, but it was common to include data appendices with publications in the early twentieth century.

Managing data

Data management provides necessary ways to make data:

- Interpretable
- Systematic
- Codified
- Portable
- Durable
- Retrievable
- Perpetual
- Sharable

"But I don't do data!"

- This is a common misconception of cultural anthropologists.
- If you gather any information, such as fieldnotes or photographs, while doing fieldwork, then you have data.

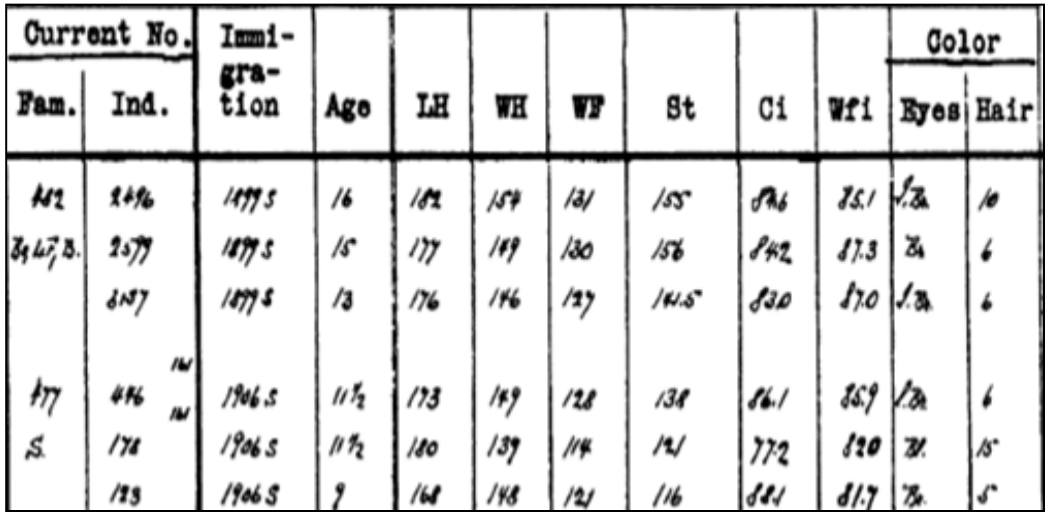
http://www.dreamstime.com/photos-images/open-laptop-computer.html

Managing data:
Case study of re-use

- Clarence Gravlee re-used data to revisit earlier, hotly contested findings of Franz Boas (1912) and Marvin Harris (1970) about race.
  - Boas measured cephalic index of immigrant groups to dispel notions of racial heredity.
  - Harris demonstrated the ambiguity and cultural construction of Brazilian racial classifications.
- While the original works were innovative and carefully done, the analytic methods needed to answer the questions fully did not yet exist.
  - One method Gravlee used was analysis of variance.

Managing data:

Case study of re-use



Current No.		Immi-								Color	
Fam.	Ind.	gra- tion	Age	LH	WH	WF	St	Ci	Wfi	Eyes	Hair
442	2496	1877 S	16	182	154	131	155	84.6	85.1	1.3a	10
3, 4, 7, 8.	2577	1877 S	15	177	149	130	156	84.2	87.3	3a	6
	2587	1877 S	13	176	146	127	141.5	83.0	87.0	1.3a	6
477	496 IW IW	1906 S	11½	173	149	128	138	86.1	85.9	1.3a	6
S.	178	1906 S	11½	180	139	114	121	77.2	82.0	3a	15
	123	1906 S	9	168	148	121	116	88.1	81.7	3a	5

Detail of a page of Boas's data in *Materials for the Study of Inheritance in Man*.

In Gravlee et al. 2003. Used with permission of the American Anthropological Association.

- Gravlee and his co-authors (2003, 2005), using modern statistical methods, both substantiated and refined the original findings.
- Boas's reanalyzed data were in raw form; Harris's original stimuli were replicated. Both are digitized and available online (www.gravlee.org/research), and have been used by other researchers.

Managing data:

Basic steps

- Think about ways to make data legible and meaningful to others beyond yourself and/or your research group.
- Anonymize the data:
  - Two sets of notes, one anonymized, may be required, one with sensitive data, one without.
  - Implement a system from the start to anonymize data.
    - Use a separate key that links names with ID numbers.
- Include contextual information:
  - Where, when, who, and how data were recorded
  - Demographics, e.g., neighborhood, gender, age
- Provide links between qualitative and quantitative data:
  - Often can use the same basic sampling parameters in both qualitative and quantitative data sets.

Software:
Text analysis

Functions of text analysis software

- Aids in the interpretation and management of large amounts of textual, graphical, audio, or video data
- Aids in identifying and using data
  - Any passage—from a word to a full section—can be tagged with a code.
  - Coding facilitates accurate and easy retrieval and/or comparison to other passages.

[In-class exercise: Discussion of data collection and analysis]

Software:
Text analysis

Common packages used by anthropologists:

Open source:

- AnSWR (CDC), AQUAD 7, CATMA, ELAN, EZText (CDC), QCAmap, QDA Miner Lite

Proprietary: all have graphical user interface

- Atlas.ti, Dedoose, DiscoverText, Ethnograph, HyperRESEARCH, Maxqda, Nvivo, QDAMiner, Quirkos

Data files in all programs are exportable to portable file types such as .txt and XML formats to ensure cross-platform readability.

Numerical data

Types of data that can be input:

- Field notes (content-analyzed)
- Interview transcripts (content-analyzed)
- Case studies (content-analyzed)
- Survey data
- Anthropometric data, e.g., height, weight, body fat
- Biomarker measurements, e.g., salivary cortisol, blood pressure

It is important to learn how to code, enter, and clean numerical (and some text) data using a standard statistical package designed for the social sciences.

Content analysis:

- A research technique for making replicable and valid inferences by interpreting and coding textual material
- This allows qualitative data to be converted to quantitative form.

Software:

Numerical data analysis

Common packages used by anthropologists

Open source:

- PSPP: graphical user interface
- R: syntax-driven

Proprietary:

- Anthropac, SPSS, SAS, STATA, SYSTAT, UCINET

While proprietary statistical packages may have more advanced features, open source packages are no cost and work well.

Data files in all programs are exportable to portable Excel-friendly formats such as .rtf or .txt to ensure cross-platform readability.

Steps to creating digital data:
The codebook

One way to store data in digital form is by numerical codes.

- A code is a symbolic (typically, numerical) representation of a bit of meaningful information.
- Note carefully that coding does not diminish, destroy, or dehumanize the original data set.
- Coding simply stores the data in an accessible, and ultimately manipulable, in an alternate form.

Steps to creating digital data:
The codebook

A codebook

- is a guide to the chosen codes that is created after they are assigned.
- allows for share these newly assigned meanings with others.

A codebook is an efficient tool

- for organizing the information needed to properly record the data codes.
- for your immediate use.
- to make data intelligible to others who may wish to share them, now and in the future.

[Outside-class exercise: Creating a codebook]

Data archiving

What to do with data once it is manageable?

Types of archives:

- Private: Secure multi-media backup in a protected environment, especially for the original raw data set
- Public: Digital archiving
  - Digital archive within a U.S. university, such as http://guides.lib.ua.edu/c.php?g=39901&p=3334457
  - A national archive
  - Registry of Anthropological Data Wiki http://anthroregistry.wikia.com/wiki/Wiki_Content

Data archiving

Why archive?

- Private: For security
- Public: Sharing of data to
  - Enhance open scientific inquiry
  - Promote new research
  - Encourage diversity of approaches to data analysis
  - Allow others to test new or alternative hypotheses
- Archiving helps move us from lone-wolf researchers to a community of social scientists.

In-class exercise: Discussion of data collection

1. What types of data have you generated
   - in the field?
   - at your office or home base?
2. For each situation:
   - How did/do you protect your data?
   - What methods of back-up did/do you use?
3. Are there any types of data for which you currently do not have a back-up plan?
4. With all candor, describe a time when you lost data due to insufficient protection.
5. Think about your most recent data collection instrument.
   - What identifying information do you have on it?
   - If you drop some field notes walking back from the library, will someone be able to return them to you?
6. Picture your most recent data files.
   - What identifying information exists on them?
   - If someone finds your data files 100 years later, will that person know what they are and how to interpret them?

Learn about and discuss one text analysis software: Click here for a brief tutorial of CATMA

Outside-class exercise: Creating a codebook

- At the top of the codebook, be sure to include general info identifying the project.
- Variable names are in CAPS, and each is ideally from 2–8 characters.
- Missing data is coded as a series of 9's that exceeds the highest value possible.
  - Example for Age: since someone could be 99 years old, the missing value for Age would be 999.
- The first variable is always "CASEID," the unique identifying number that each case carries.

Outside-class exercise: Creating a codebook

Example to use as a template:

CODEBOOK FOR ANTHROPOLOGY STUDENT SURVEY

Jane Dost, PhD, Researcher
1350 Doster Hall

January 1, 2016
University of Alaska

Variable #	Variable Description	Variable Name	Values	Format
1	Case ID Number	CASEID	Continuous	F2.0
2	Transfer Student	TRANSFER	1. Yes 2. No 9. Missing	F1.0
3	Transfer from where	WHERE	text	A35
4	Area of Concentration	AREACONC	1. cultural 2. biological 3. archaeological 4. linguistics	F1.0
5	Accessibility of Faculty	ACCESANT	Continuous, Scale of 1-10	F2.0

Note: The codebook can be created using PSPP; many tutorials are available online.

Outside-class exercise: Creating a codebook

- Each bit of data is a variable (i.e., the information will vary in value across your cases) and will need to be defined.
- The 3 types of variable are:
  - Nominal: the least complex; it names something using categories, with no logical order; e.g., Gender.
  - Ordinal: a measurement using categories, in which the categories are logically ordered though not necessarily of equal value; e.g., Illness Gravity.
  - Continuous (aka interval): the most complex; measurement on a continuous scale, with each interval of equal value; e.g., Age.
- Text variables are not coded, and may be entered as is ("string").
- A format indicates in what form a variable is coded.
  - F for Numerical Data: Fx.y, where F means numerical, X the maximum # of digits the highest value can have, and Y the # of decimal points. If the highest age possible (in round years, no fractions) is 99, the format will be F2.0.
  - A for Text Data: A#, where A stands for text, and # indicates the number of characters (including spaces) allotted to that variable. Allow for the longest possible answer, e.g., Teodolinda: A10

Outside-class exercise: Creating a codebook

Andean Highlander Demographics and Recent Illness History:
Using these sample data, create your own codebook for the variables Name, Age, Gender and Gravity of Illness.

1	At 32 years of age, Teodolinda is still living at home, nearly despairing of finding a husband. Her mother is worried that her *pena* (sadness) is to the point she cannot function well, and would like her to see a curandero for healing.
2	Daniel is 19, single, and the best soccer player his community has ever produced. As long as he stays healthy, everyone thinks he has a shot at playing for the national team.
3	Fidelita and her husband Raul would prefer to remain in the highlands and tend their crops and sheep, despite the pleas of their kids to come live with them on the coast, where they promise to get her treatment for her occasional skin allergies.
4	Azucena, 60 and recently widowed, is accompanied by two of her young grandchildren while their parents work in the city. She has dizzy spells that the doctor has said is due to extremely high blood pressure, though she thinks it is caused by *mal viento* (evil wind).
5	Since Jorge's wife died last year, there is no one to help around the house. Despite his advanced age of 99, he must ride to the market town each Sunday to get supplies. The last time, he fell off his burro and hurt his back, and is now bedridden with no family to care for him.
6	Lucía had a daughter, Claudia, with her childhood sweetheart. Her parents disapproved of the union, so at 27 years of age she gave birth at her sister's house without proper care from a midwife, which has led to the herbalist's diagnosis of a bit of *debilidad* (debility, exhaustion).
7	Tomás, divorced from his first wife for several years, has just moved in with a woman who is also 33. She has 2 teenage children from a previous union. The family would have planted their spring potato crop last week if he hadn't been bedridden with a case of the flu.
8	When Eustacia, 47, saw her son slip off the cliff during a storm, she suffered a tremendous *susto* (fright illness) that did not go away even though he lived. Her husband was powerless to make her feel better and was worried her illness was so severe she might die from it.

Outside-class exercise: Creating a codebook

All done? Your codebook should look something like this:

CODEBOOK FOR ANDEAN HIGHLAND ILLNESS RESEARCH PROJECT

Your Name, Degree, Role **Date**
Your Work Address **Institution/Company**

Variable #	Variable Description	Variable Name	Values	Format
1	Case ID Number	CASEID	continuous	F2.0
2	First name of participant	NAME	text	A20
3	Age in years	AGE	continuous	F3.0
4	Gender of participant	GENDER	1. female 2. male 9. missing	F1.0
5	Gravity of current illness	ILLNESS	1. none 2. mild 3. moderate 4. serious 9. missing	F1.0

Optional: Take a brief tour of how to manage data in PSPP:
https://www.youtube.com/watch?v=-ZRxpp1y4BY

References

Boas, Franz. “Changes in the Bodily Forms of Descendants of Immigrants.” *American Anthropologist* 14 (1912): 530-62. http://onlinelibrary.wiley.com/doi/10.1525/aa.1912.14.3.02a00080/full

Gravlee, Clarence C., H. Russell Bernard, and William R. Leonard. “Boas's Changes in Bodily Form: The Immigrant Study, Cranial Plasticity, and Boas's Physical Anthropology.” *American Anthropologist* 105(2) (2003): 326-32. http://onlinelibrary.wiley.com/doi/10.1525/aa.2003.105.2.326/full

Gravlee, Clarence C. “Ethnic Classification in Southeastern Puerto Rico: The Cultural Model of ‘Color.’” *Social Forces* 83(3) (2005): 949-70. http://www.jstor.org/stable/3598265

“Gravlee, Clarence C. - Research.” Accessed July 20, 2016. www.gravlee.org/research

Harris, Marvin. ”Referential Ambiguity in the Calculus of Brazilian Racial Identity.” *Southwestern Journal of Anthropology* 26(1) (1970): 1–14. http://www.jstor.org/stable/3629265

Leopold, Robert. “The Second Life of Ethnographic Fieldnotes.” *Ateliers du LESC* 32 (2008). **http://ateliers.revues.org/3132**. DOI: 10.4000/ateliers.3132

National Information Standards Organization (NISO). *Understanding Metadata,* Bethesda: NISO Press, 2004. http://www.niso.org/publications/press/UnderstandingMetadata.pdf

Ruel, Erin, William Edward Wagner III, and Brian Joseph Gillespie. “Data Archiving.” In *The Practice of Survey Research: Theory and Applications*, 305-12. London: SAGE Publications, 2015.

Silver, Christina, and Ann Lewins. *Using Software in Qualitative Analysis: A Step-by-Step Guide*. London: SAGE Publications, 2014.

AMERICAN ANTHROPOLOGICAL ASSOCIATION
Advancing Knowledge, Solving Human Problems

Acknowledgments

Modules: *Writers,* Arienne M. Dwyer, Blenda Femenías, Lindsay Lloyd-Smith, Kathryn Oths, George H. Perry; *Editor,* Blenda Femenías

Discussants: *Workshop One, February 12, 2016:* Andrew Asher, Candace Greene, Lori Jahnke, Jared Lyle, Stephanie Simms

Workshop Two, May 13, 2016: Phillip Cash Cash, Jenny Cashman, Ricardo B. Contreras, Sara Gonzalez, Candace Greene, Christine Mallinson, Ricky Punzalan, Thurka Sangaramoorthy, Darlene Smucny, Natalie Underberg-Goode, Fatimah Williams Castro, Amber Wutich

American Anthropological Association:
Executive Director, Edward Liebow
Project Manager, Blenda Femenías
Research Assistant, Brittany Mistretta
Executive Assistant, Dexter Allen
Professional Fellow, Daniel Ginsberg
Web Services Administrator, Vernon Horn
Director, Publishing, Janine Chiappa McKenna