Bringing Digital Data Management Training into Methods Courses for Anthropology

Biological Anthropology: Principles and Practices of Digital Data Management

George H. Perry

2016

Recommended citation:
Perry, George H. "Biological 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. Advantages for biological anthropology in data sharing

III. Challenges for biological anthropology in data sharing

IV. Databases and considerations for various types of data

V. Primary data compared to processed data

VI. Exercises

VII. References

VIII. 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 the ethical dimensions of data management?
- What is a data management plan?

Advantages for biological anthropology in data sharing

- Biological anthropology is a data-rich discipline.
- Primary data availability maximizes not only reproducibility but also impact.
- Data availability facilitates opportunities for the next generation of anthropologists.
- Obligations for data collected with taxpayer-supported funds include:
    - Increasing requirements from funding agencies
    - Evaluation of the data management plan as part of the grant review process
- Long-term experience in data sharing community standards and benefits:
    - Anthropological genetics/genomics

Biological anthropology as a data-rich discipline

A very partial list of biological anthropology data types:

Behavioral records, fossils, isotopic measurements, bones, hormone measurements, X-ray images, microscopic images, tissues, skeletal measurements, medical records, biomechanical models, cadavers, bioarchaeological assessments of age and sex, genetic/ genomic genotypes and sequences, volatile organic compound measurements, computed tomography images, geocoded sample information, environmental/ ecological data, food mechanical and nutritional properties, paleopathological differential diagnoses, histological data, energetics data

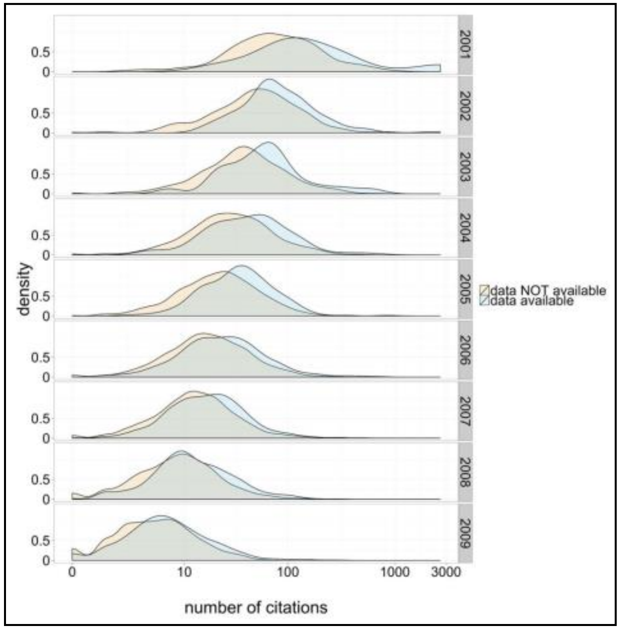
Biological anthropology data types

- There are differences among data types in regard to management:
  - Each type may have different repositories.
  - Each type may have different data curation needs.
- Consider these differences from the start of your project as part of creating a good data management plan!

[Outside-class exercise: Discuss data types]

Data sharing and scientific impact

- Greater availability of data tends to increase some measures of scientific impact.
- The number of citations, one measure of scientific impact, is typically greater for publications with full data availability.



The graphic shows the results of an analysis based on 10,555 studies (from 2001–2009) that generated gene expression microarray data.

Piwowar and Vision 2013. Published under prevailing CC BY license: https://peerj.com/about/FAQ/#license. DOI 10.7717/peerj.175/fig-1

Challenges for biological anthropology in data sharing

- Fast-moving technology and methods for data collection and analysis create the need to gain and maintain knowledge of evolving databases and standards.
- Large file formats are becoming more and more prominent.
  - Computational training and skill are increasingly needed to use and archive data.
- Varying availability of appropriate community-supported databases:
  - Not all such databases yet exist.
  - They require stable funding and management.
  - Databases with "too big to fail" status are defined as a critical mass of widely valuable research data, such that the long-term integrity of the data would likely be maintained by major funding bodies or governments, even if the database itself became outdated or was no longer maintained.

Challenges for biological anthropology in data sharing

- Data sharing may be at odds with historical operating procedures for many paleoanthropologists.
  - Even upon access to specimens, restrictions may be placed on use.
- Directors of long-term field ecology (e.g., primatology) studies may have concerns about fully open data accessibility.

Golden-crowned sifaka (*Propithecus tattersalli*) near Daraina, Madagascar.

Photograph by George Perry

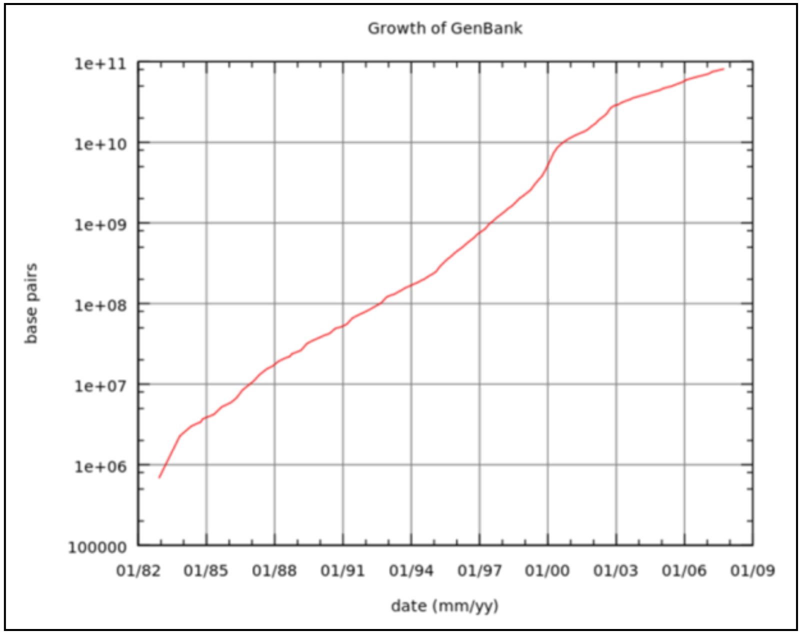
Challenges for biological anthropology in data sharing

- There are privacy and ethical considerations with some forms of human biological data that others could potentially associate with participants.
- Incorporate planning for maximal data sharing, given the privacy risks, into the data management plan.
- Address data sharing with participants in the informed consent process from the outset.

Databases: Anthropological genetics/genomics

GenBank: A database maintained at National Institutes of Health (NIH) since 1982 for depositing determined nucleotide sequences of a gene/ genomic region for specified individuals and organisms.

- Sequences deposited receive accession numbers and are cross-referenced with associated publications.
- Users can search by topics such as organism and genetic locus, or query directly against nucleotide sequences via various tools.



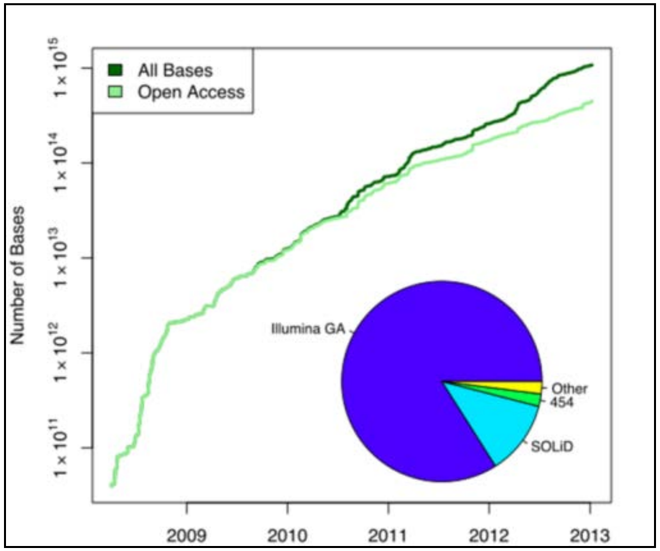
Wikimedia Commons.
https://commons.wikimedia.org/wiki/File:Growth_of_Genbank.svg

Databases: Anthropological genetics/genomics

Reference sequences for the human genome and for other organisms, including archaic hominins such as Neandertals, are now available.

- Sequence depositions from individual labs have collectively facilitated an otherwise impossible level of scientific advance.
- With new, massively parallel sequencing technology, data are now frequently deposited into other databases, such as the Sequence Read Archive (SRA).
  - Original sequence data for many thousands of studies are available on SRA.
- Numerous other databases also exist for different types of genomic data.

Exponential growth of the Sequence Read Archive, 2009–2013



Wikimedia Commons, by Ben Moore: https://commons.wikimedia.org/wiki/File:History_(and_predicted_future)_size_of_the_Sequence_Read_Archive.svg

Databases: Paleoanthropology/Skeletal biology

For some analyses, it is important to work with original fossil and skeletal material.

- Data such as measurements of individual specimens should be made available.
- There is limited access to materials, at minimum due to travel expenses.

Homo naledi mandible

Wikimedia Commons, by Patrick Randolph-Quinney.
https://commons.wikimedia.org/wiki/File:Homo_naledi_mandible_2.jpg

Databases: Paleoanthropology/Skeletal biology

- There have been many recent advances in digital imaging technologies.
  - Digital measurements can be more precise than those made by hand.
  - There are increased opportunities for automated, scalable, highly reproducible measurements.
  - Analyses of internal structure are only possible with imaging technology (e.g., CT).
- Stored image data facilitate subsequent, and not originally envisioned, analyses.

- Digital data can be shared!

Databases: Paleoanthropology/Skeletal biology

- CT scanning can be done of both external and internal surfaces.
  - The use of varying resolutions facilitates analyses of different scales of structure
- External laser scanning is now sufficient for high quality measurements and analyses of fine-scale shape.
- Some but not all external laser scanning methods provide fine-scale resolution sufficient for most external surface research purposes.

[Optional exercise: 3D scanning and printing of skeletal elements]

3D CT of bilateral mandible fracture

Wikimedia Commons, by Coronation Dental Specialty Group, https://commons.wikimedia.org/wiki/File:3D_CT_of_bilateral_mandible_fracture.jpg

Databases: Paleoanthropology/Skeletal biology

Resources available online:

- MorphoSource, Duke University:
  - A cost-free database supported by NSF
  - Provides open source storage and retrieval of imaging data files.
- Forensic Anthropology Data Bank (FDB), University of Tennessee:
  - Supported by National Institute of Justice (NIJ)
  - Database contains demographic information and skeletal information for thousands of cases.

Databases: Primatology

- Publications often incorporate analyses of long-term, high-investment, ongoing field data that are expected to be the basis of many subsequent publications.
- Researchers may be apprehensive about open sharing of all data underlying each paper.
  - The idea of sharing is often at odds with increasing expectations for publication and funding of data sharing.
  - There is some risk of reduced incentive for initiating or continuing long-term studies.
- In the absence of a current solution, some recent suggestions include (Mills et al. 2015, Whitlock et al. 2016)
  - The willingness from journals and funders for relatively long data embargos, such as a period of 5 years.
  - Increased data-tracking processes and communication with data generators.

Primary data compared to processed data

- Consider depositing beyond the minimum required by journals and funders.
- Processed data files, rather than only the raw data, and other information can greatly aid reproducibility of the work and maximize its impact and usefulness. Examples:
  - Sequence alignments rather than only raw reads
  - Both raw and processed image data
  - Code used for analyses
- Options include the Dryad Digital Repository, Figshare, and GitHub (for code).
- Personal or department websites are not acceptable options.
  - The reliability of a permanent hosting commitment and “too big to fail” status are needed.
  - Data generators and users both benefit from the functionality of community archives.

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

In-class exercise: Discussion of data collection and management

1. What types of data have you generated
   - In the field?
   - In the laboratory?
2. For each situation:
   - How did you plan your study
   - Did you plan for a data backup procedure?
   - Were there any issues of accidental data loss?
3. Imagine that you returned to your data one year after collection.
   - How permanent was your data storage solution?
   - Is there sufficient annotation of the data
     - for you to be able to understand what everything represented and how to analyze all the data?
     - for someone else to understand and analyze everything?

Outside-class exercise: Data types

Objectives: Identify data types used in research as published in peer-reviewed articles, and evaluate current and future access to the data.

You may select an article from journals published by the American Anthropological Association, such as *American Anthropologist*, or others that meet criteria for peer-reviewed journals. (Consult your university library's website for criteria.)

1. Select two data types used in biological anthropology studies. You may choose from the list provided on Slide 6, or suggest additional types and have your instructor confirm your selection.
2. Locate an article in a peer-reviewed journal in which each data type is used as the basis for analysis and interpretation.
3. For each data type in each article:
   - Does the author(s) provide information about the data's location, e.g., depository?
   - Could you readily access all the data cited in the article?
     - Today?
     - In the future?
     - If not all, how much?
   - If "no" to any of the above, what limits the access?

Optional exercise: 3D scanning and printing for data about skeletal elements

Instructor notes: The guidelines provided are for small groups to do the exercise in two class periods, with the actual printing between the periods likely to be done outside class at the printer's location. The exercise can also be done by individuals, and discussed afterward in class.

If the elements are available in your university's collection or a nearby museum, students can scan objects before printing. Another option is to use data in an existing database as a basis for printing.

1. Have a group discussion about hominin or non-human primate fossil or skeletal elements. These could be individual bones or cranial elements.
2. Each student should choose 4 elements that would be valuable in biological anthropological research. You will examine, 3D scan and print, and compare the resulting printed objects.
   - In what ways are the different individual elements interesting?
   - How can side-by-side comparison of multiple elements from the proposed set be valuable?
3. Discuss the value of 3D scanning and printing
   - for collaborative research.
   - for science education.
4. Decide on the 4 elements that the group will scan and print.
5. Have the elements printed.
6. In the next class, continue discussion with the specimens in hand.
   - Does examining the printed objects change your ideas about
     - the value of the elements?
     - of 3D printing?

3D print of human skull.

Photograph by Nevit Dilmen, Wikimedia Commons

References

Michener, William K. “Ten Simple Rules for Creating a Good Data Management Plan.” *PLoS Comput Biol* 11(10) (2016): e1004525. doi:10.1371/journal.pcbi.1004525

Mills, James A., et al. “Archiving Primary Data: Solutions for Long-term Studies.” *Trends in Ecology and Evolution* 30(10) (2015): 581–89. DOI:10.1016/j.tree.2015.07.006

Piwowar, Heather A., and Todd J. Vision. “Data Reuse and the Open Data Citation Advantage.” *PeerJ* 1:e175 (2013). https://doi.org/10.7717/peerj.175

Reed, Denne, et al. “*Digital Data Collection in Paleoanthropology.” Evolutionary Anthropology 24*(6) (2015): 238-49. DOI: 10.1002/evan.21466

Whitlock, Michael C., et al. “A Balanced Data Archiving Policy for Long-term Studies.” *Trends in Ecology and Evolution* 31(2) (2016): 84–85. DOI: 10.1016/j.tree.2015.12.001

Wilkinson, Mark D., et al. “The FAIR Guiding Principles for Scientific Data Management and Stewardship.” *Scientific Data* 3 (2016). doi:10.1038/sdata.2016.18

Web resources

Dryad Digital Repository. http://datadryad.org

Figshare. https://figshare.com

Forensic Anthropology Data Bank. https://fac.utk.edu/background/

GenBank. http://www.ncbi.nlm.nih.gov/genbank/

GitHub. https://github.com

MorphoSource. http://morphosource.org/

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