JASs *forum*
Journal of Anthropological Sciences
Vol. 86 (2008), pp. 209-212

Anthropology from the desk? The challenges of the emerging era of data sharing

Sarah Elton[1] & Andrea Cardini[1,2]

1) Hull York Medical School, The University of Hull, Cottingham Road, Hull HU6 7RX, UK
e-mail: sarah.elton@hyms.ac.uk

2) Museo di Paleobiologia e dell'Orto Botanico, Universitá di Modena e Reggio Emilia, via Università 4, 41100, Modena, Italy
e-mail: alcardini@interfree.it, cardini@unimo.it, andrea.cardini@hyms.ac.uk

The previous contributions in this forum highlight the immense potential for anthropological data transfer and sharing in the digital age. Not only is it easier than ever before to disseminate information and work collaboratively, the types of data being collected are also increasingly driven by information and computing technology (Kullmer, 2008). Simply sending an email with an attached file to someone on the other side of the world is now old fashioned. Sophisticated web-based programs allow not just the transfer of information but almost instantaneous collaborative document editing from remote locations. Several large-scale databases, covering material as diverse as the palaeoclimates of Neogene mammal localities (NOW: Fortelius, 2007) and 3D reconstructions of fossil hominoids (3D-archive of fossil hominoids: Weber, 2001) are available from your desktop. And if you need to speak to or even see your colleagues to discuss these data, voice over IP (VoIP) and similar technologies allow you to videocall the rest of your research team at a moment's notice, wherever they are.

Some anthropologists were quick to grasp the research opportunities offered by the digital age. The Centre for Social Anthropology and Computing (http://lucy.kent.ac.uk) had a web presence as early as 1986, and has used information technologies extensively ever since to host shared research resources such as open-access programs as well as to provide an online repository for the huge archives generated by anthropological fieldwork and ethnographic enquiry. The Human Relations Area Files (eHRAF: http://www.yale.edu/hraf/index.html) are another well-known source of data on human cultural variation. Biological anthropology tends to be even more dependent on digital data, from measurement collection and processor-heavy statistical analyses to the virtual anthropology techniques described by Kullmer (2008). Yet in many ways, biological anthropologists have been much slower to develop online databanks or even constructively debate how and where data could be shared.

This appears to be changing fast. Although there have been numerous high-profile arguments over access to fossil material (Gibbons, 2002; Dalton, 2004; Marfat, 2008), the past couple of years have seen some serious attempts not only to get people talking about sharing data but also putting their money where their mouth is. Notable examples include EVAN and NESPOS (both discussed in detail by Kullmer, 2008) and the Wenner-Gren/NSF-funded workshop on data sharing in palaeoanthropology (Delson *et al.*, 2007). Funding bodies are also placing more emphasis on how data will be stored and made available in the future. In addition, an increasing number of museums now ask visiting scientists to leave copies of electronic data in order to make them available to other researchers, partly to avoid over-use of delicate collections. The net result of these initiatives is that biological anthropologists can now browse through numerous and ever-expanding web-based digital resources to

help them in their research. This trend is likely to continue in the future, facilitated by the development of progressively cheaper and more powerful tools for collection of digital data,

Our interest is in the dissemination of quantitative data from large datasets, an area that is slowly unfolding with the promise of resources such as PRIMO (Delson et al., 2007), developed by the New York Consortium in Evolutionary Primatology (NYCEP). In theory, at least, it should be relatively straightforward for those of us working with quantitative or metrical data to share them. Although space constraints in articles and books limit the extent to which primary data can be published, datafiles can be easily distributed, measurements or landmarks described and analyses performed again. But is sharing a dataset simply about distributing measurements or is it also about passing on ideas that are not usually given freely? Initially, this might seem like a ridiculous notion: surely a dataset and the theoretical basis of its construction are revealed when published? This is certainly true for parts of the dataset, some of which – the landmark configuration, for example –might be hugely significant. However, many researchers will never publish their full dataset in a single place, choosing instead to 'mine' nuggets and publish sequentially. This then begs the question whether dissemination of a whole dataset stands apart from sharing other primary data such as a CT of an ape skull, since a dataset may well be more than the sum of its component parts.

Another question that we are currently grappling with is how we should present our data so that they can be used and evaluated by the maximum number of people. Traditional measurements (TM) have the benefit of being relatively straightforward to conceptualise and analyse, and are therefore of use to a wide range of researchers, including those who do not specialise in morphology, and undergraduate students doing general projects. Their reliability and validity can also be ascertained fairly easily and cheaply. Geometric morphometric (GMM) data, on the other hand, require a greater degree of specialist knowledge to be used effectively, but are becoming increasingly heavily employed in anthropology. They have several advantages over TM, which have been widely discussed in the literature. In the context of data sharing, one important consideration is that whereas TM are unlikely to give geometric data, GMM can be easily converted to linear measurements, provided that suitable landmarks are defined *a priori*.

In order to be most effective, a database destined for widespread distribution needs to be accessible, easily interpreted without reference to the originator, and work on several platforms. An important initial question is which measurements (TM) or landmarks (GMM) to include. On the one hand, removing data that appear to be redundant might make the dataset more practical as there will be less 'noise' for the user to contend with. On the other, data that today seem meaningless might have future importance. Other technical challenges might arise when datasets on the same configuration from different sources are merged, something that may need to be solved on a case-by-case basis. The format in which the data are presented is another important aspect of producing a dataset that is easily interpreted. Should anthropologists be striving to develop standard formats for databases, or is it safe to assume that the experienced user will be able to navigate numerous interfaces? What types of files (.csv, .txt and so on) are best? These questions also play to accessibility issues and utility on several platforms, as does whether single or multiple repositories should be used.

Data must also be stored in a 'future-proof' way, an issue that organisations like the Archaeological Data Service (ADS: http://ads.ahds.ac.uk/) continually debate. Twenty years ago, storing data on tapes or 12" floppy disks was the norm, but now few machines can read such media. We assume that online storage of data has largely solved such problems, but software evolves at least as fast as hardware. How do we know that the .csv and .txt files that we use today will be readable by programs ten years on? And is it practical to expect that we will regularly update our datafiles as software and platforms change? Online databases therefore need a policy of long term management. Besides keeping up with

changes in platforms and software, resources have to be found to allow periodic maintenance operations. Such tasks are fundamental to the prevention of degeneration of digital copies, to correct errors that may be reported by users and to update information on often unstable taxonomies. Indeed, issues of long term maintenance can be crucial for the success and survival of digital databases. The recent attempt to catalogue all living species online highlights how essential it is to secure long-term funding for digital archives: this high-profile project, headed by Edward O. Wilson, had to be halted after only a few years of operation because of a lack of grants and donations (Miller, 2005).

Our first attempt at making data widely available can be found at http://ads.ahds.ac.uk/catalogue/archive/cerco_lt_2007/. This dataset gives traditional linear measurements of guenon skulls, derived from a larger GMM database comprising hundreds of specimens of Old World monkeys. Initially described using a configuration of three dimensional landmarks (Cardini et al., 2007), the coordinates include many anatomical landmarks used in studies of primate skull morphology. Thus, sets of linear measurements employed in traditional morphometrics can be easily obtained. Where available, we also provide information about the provenance of the specimens. We are by no means the first people to make guenon data freely available: Verheyen's monograph (1962) included a large set of linear measurements which was fruitfully used in several studies after the original publication. For instance, Martin & MacLarnon (1988) reanalysed Verheyen's data to emphasize how taking allometry into account in comparisons of species spanning a large range of sizes may be crucial to reconstruct their relationships. Later, Shea (1992) compared measurements from *M. talapoin* and *C. cephus* and showed that differences in adult cranial proportions in these species, and indeed probably in most guenons, may be largely related to common patterns of ontogenetic scaling, a finding supported by our recent three-dimensional geometric morphometric analysis of this group (Cardini & Elton, 2008).

Datasets published in books and papers have thus already proved their potential utility. However, they generally require considerable effort to find then extract from the original publication in a form suitable for analysis. Digital data, in contrast, are easy to share using websites and can be made available in formats that allow them to be imported into almost any statistical software. When publishing our data, we were fortunate to be able to use the existing repository at the Archaeological Data Service, which provides an interface that is easily navigable and allows .csv file downloads. In the future, we aim to expand the dataset, first by increasing the number of specimens and later making the set of three-dimensional landmark coordinates available directly.

Such a process, which aims to make a large amount of data available, is bound to take some time. This is not only because of understandable desires to mine an extensive dataset, which in our case was designed for a specific and still ongoing large scale project exploring the potential of Old World monkey models to provide a contextual framework for human evolution. Large datasets also need to be 'cleaned' in order to find and correct errors (e.g., incidental misplacement of landmarks during digitization, incorrect species identification in museum catalogues, errors in conversions of file formats, and misspelled or outdated locality names), and their accuracy as well as their user-friendliness have to be tested. To aid this, it might be possible to collaborate with other scientists, either in cooperative projects or with them acting as 'beta-testers' who can use data for their own studies and help by reporting errors, inconsistencies and ambiguities.

Widespread institutional policies on online, open access data repositories of the type seen in molecular biology (e.g. EMBL-bank: http://www.ebi.ac.uk/embl/index.html) are still some way off in biological anthropology and morphology. However, it is highly probable that in the future it will be much easier and less expensive to get data from either fossils or modern specimens. The need for morphologists to do the 'grand tour' of museums is likely to be much reduced. This has obvious positive benefits not only for

researchers' bank balances and carbon footprints but also for the collections, which over time and under extensive use get worn and unintentionally damaged. Making data available may also allow a whole new population of researchers to be involved in analysis: scientists in regions rich in fossils but short of funds, for example, could mine online data for comparative samples without the need for extensive travel, and students on restricted budgets might have the chance to undertake studies as extensive and interesting as those done by their wealthier colleagues. Making the first steps towards online databases, as we and many other colleagues are doing, will help us to better understand the problems and benefits of electronic data sharing and to keep the momentum that will further encourage funding bodies to support pioneering initiatives to disseminate datasets. Maybe, however, 'anthropology from the desk' will deprive our job of that little bit of romance and appeal that one gets working in a dusty room of an old museum collection. It may also give us less chance to learn from curators and collection managers, who know the history of those specimens which they have dedicated part of their life to collect and preserve so well, and who generously provide help and advice to visiting scientists. Now that the data sharing ball is rolling in anthropology, we need to keep it moving whilst also ensuring that it takes the debates, careful observations, detailed comparative analyses and human interactions with it.

References

Cardini A. & Elton S. 2008. Variation in guenon skulls I: species divergence, ecological and genetic differences. *J. Hum. Evol.,* 54: 615-637.

Cardini A., Jansson A-U., Elton S. 2007. Ecomorphology of vervet monkeys: a geometric morphometric approach to the study of clinal variation. *J. Biogeogr.,* 34: 1663-1678.

Dalton R. 2004. Anthropologists rocked by fossil access row. *Nature,* 428: 881.

Delson E., Harcourt-Smith W.E.H., Frost S.R., Norris C.A. 2007. Databases, data access and data sharing in palaeoanthropology: first steps. *Evol. Anthropol.*, 16: 161-163.

Fortelius M. 2007. Neogene of the Old World Database of Fossil Mammals (NOW). University of Helsinki. http://www.helsinki.fi/science/now/.

Gibbons A. 2002. Glasnost for hominids: seeking access to fossils. *Science,* 297: 1464-1468.

Kullmer O. 2008. Benefits and Risks in Virtual Anthropology. *J. Anthropol. Sci.,* in press.

Mafart B. 2008. Human fossils and paleoanthropologists: a complex relation. *J. Anthropol. Sci.,* in press.

Martin R.D. & MacLarnon A.M. 1988. Quantitative comparisons of the skull and teeth in guenons. In A. Gaultier-Hion, F. Bourliere, J.P. Gautier, J. Kingdon (eds.): *A primate radiation: evolutionary biology of the african guenons. Cambridge University Press,* Cambridge, pp. 160-183.

Miller G. 2005. Taxonomy's Elusive Grail. *Science,* 307: 1038.

Shea B.T. 1992. Ontogenetic scaling of skeletal proportions in the talapoin monkey. *J. Hum. Evol.,* 23: 283-307.

Verheyen W.N. 1962. Contribution a la craniologie comparee des primates. Les genres Colobus Illiger 1811 et Cercopithecus Linne 1758. Musee Royal de l'Afrique Centrale, *Tervuren Annales,* 8 (105), 1-255.

Weber G. 2001. Virtual Anthropology (VA): calling for Glasnost in palaeoanthropology. *Anat. Rec.,* 265: 193-201.