Evolutionary Anthropology 16:161–163 (2007)

Databases, Data Access, and Data Sharing in Paleoanthropology: First Steps

Accessibility, sharing, and dissemination of large datasets are becoming important issues in paleoanthropology. Such access increasingly allows researchers to arrive at exciting new insights and findings while avoiding the needless repetition of existing work. New technologies, particularly computer power, database software, and networking capabilities, have made it possible for researchers to post large and complex datasets online. These can subsequently be mined in increasingly sophisticated ways. However, researchers in our field, as in others, use a multitude of different approaches to database organization, software design, and the level of access allowed to outside users. As a result, the time was ripe for a broad-based evaluation of the current state of databasing in paleoanthropology.

In April 2007, a diverse group of scientists met in New York at the American Museum of Natural History to discuss the major issues surrounding the access and dissemination of paleoanthropological data. The workshop's participants represented institutions from around the globe and included paleoanthropologists, paleontologists, archeologists, database specialists, collection managers, and representatives of major funding agencies (Fig. 1). The workshop, co-funded by the Wenner-Gren Foundation and the National Science Foundation, was divided into four sessions over two days.

The morning session of the first day started with comments from Leslie Aiello (Wenner-Gren Foundation) and Mark Weiss (National Science Foundation), followed by brief introductions from each participant. The rest of the morning focused on the general importance of sharing and disseminating paleoanthropological data. Although the views expressed were diverse, a broad consensus emerged that databasing is a critical issue for paleoanthropology and its allied fields. It was recognized that many essential topics needed to be addressed before any further initiative could be taken. For example, the idea was broached of developing a "portal" site through which a researcher could access numerous individual databases. The concept of "data" was explored, with alternative subdivisions into metadata, contextual, primary ("raw"), and derived data. In addition, the relationship between institutions and researchers who study their specimens was discussed.

The afternoon session was devoted to demonstrations of existing databases in order to acquaint workshop participants with the different ways paleoanthropological, paleontological, archeological, and institutional collection data are currently being collated and organized. Databases and their representatives included the Primate Morphology Online (PRIMO) database (Eric Delson, Will Harcourt-Smith, and Steve Frost, CUNY/AMNH and the University of Oregon: http://www.nycep.org/primo); the Human Origins Database (HOD) (Bernard Wood and Adam Gordon, George Washington University); the Smithsonian Paleoanthropology database (Rick Potts and Matt Tocheri, Smithsonian Institution); the Revealing Human Origins Initiative (RHOI: http://rhoi.berkeley.edu/) database (Henry Gilbert, California State University at East Bay); the Neanderthal Studies Professional Online System (NESPOS) (Gerd-Christian Weniger, Neanderthal Museum: https://nespos-live01.pxpgroup.com/display/openspace/Home); the Ancient Human Occupation of Britain database (AHOB) (David Polly, Indiana); the digital@rchive of Fossil Hominoids (http://www.virtual-anthropology.com/3d_data/3d-archive) and European Virtual Anthropology Network (EVAN) (both Gerhard Weber, University of Vienna); the Siwalik Database Project (John Barry, Harvard); the Neogene Old World Mammals (NOW) database (Mikael Fortelius, Helsinki: http://www.helsinki.fi/science/now/); the Knowledge-based Archaeological Data Integration System (KADIS) (Keith Kintigh, Arizona State); the Transvaal Museum database (Francis Thackery, Transvaal Museum); the National Museums of Kenya database (Frederick Kyalo Manthi, NMK); the Institute of Vertebrate Paleontology and Paleoanthropology site database (Wu Liu, IVPP, Beijing: http://mdata.ivpp.ac.cn:8080/ivppweb/enspecimensearch/); the AMNH Vertebrate Zoology Catalogue (Richard Monk, AMNH); and Paleoportal, presented by Chris Norris of AMNH (http://www.paleoportal.org). This last system exemplified a means of integrating several local independent (here institutional) databases via a central query-based web site. URLs are provided for those databases that currently are web-accessible.

On the second day, after a visit to the new AMNH Spitzer Hall of Human Origins, the participants discussed integration and dissemination of paleoanthropological data. Topics included the value of distributed databases such as Paleoportal, types of data, access to data, and ownership of data. A large part of this dis-

Figure 1. Workshop participants on the steps of the American Museum of Natural History.

cussion covered alternative concepts of "data," especially the differences between "primary" data (that is, direct representations of specimens such as CT or laser scans) and "derived" data (for example, measurement or character data, more commonly termed "raw" data). The former would allow a user to produce a duplicate specimen, either virtually or as a solid printout and thus avoid the need to return to the original to study it. Thus, institutions understandably wish to retain control of such data. Digital images also fall under the primary rubric. Of course such data are also in fact "derived" in some ways, as they often require expert interaction by the operator or editor before they can be used. Another concept that was much mentioned but used in different ways by various participants is "metadata." Many speakers used this term to refer to collection or contextual data about specimens. However, as Richard Monk and others clarified, any information about a specimen is "data," while "metadata" should only be used to refer to information about the data (for example, data collector, date of collection, instrumentation used to collect data, exactly what the data were, or the repository of data). Contextual or collection data might be of primary importance for some researchers, among them Potts, RHOI, AHOB, and NOW, but secondary for others such as HOD and PRIMO, depending on the goals of the database involved.

Access to and ownership of data were major topics. All participants agreed that the primary control of a database obviously rests with the institution, research group, or individuals that originally created it and, presumably, collected the included data. That collection is usually the result of many years of meticulous work and hard-earned funding. On one hand, such data collectors desire to retain the option of sole access until publication of the information. On the other hand, not only can they benefit from sharing data but, as some argued, data collected with the aid of governmental or private sources belongs to a wider community or public.

As John Yellen (NSF) discussed, several years ago the NSF sponsored an online questionnaire about access to fossils recovered with NSF or other public support. Details have not been disseminated, but Chris Norris summarized part of the results. Of 66 respondents, about one-fourth thought that despite its funding support NSF should have no say in the disposition of data, be it access to fossils or otherwise. Nearly 75% of the respondents disagreed, thinking that NSF oversight of some kind would be beneficial. All concerned accepted that fossils collected in a country formed part of that nation's patrimony. Although they felt that discoverers must be allowed "sufficient" time to publish their results, the extent of that time interval was highly variable. "Consumers" of fossil data were generally more willing than "producers" to accord NSF authority to ensure timely access to fossil material. All agreed that the rights of host institutions should not be infringed upon. Gerhard Weber reported on a similar survey he conducted with comparable results.

All institutional representatives at the workshop were asked to briefly state their policy on access to fossils and to the production and ownership of primary data. The institutions represented included the AMNH, Smithsonian, Musée de l'Homme (Martin Friess), Neanderthal Museum, National Museums of Kenya, Transvaal Museum, and the IVPP. Although there is much variation and flexibility, in general institutions allow fossil collectors to retain or control sole access until a "major publication" (undefined, but usually beyond the time when a preliminary announcement or description in *Nature*, *Science*, or similar journal) is produced. The question of access to a specimen named as a holotype by the discoverer was also discussed, given that the International Code of Zoological Nomenclature recommends, but does not mandate, that institutions make types "accessible for study." Institutional policy also varied in terms of CT scans and other types of primary data, but many museums now require that such images and, in some cases, photographs and derived data as well, be deposited with the institution, either as a copy or by claiming ownership of and copyright to the original. Susan Antón (NYU) urged that all institutions develop and clearly announce such a policy.

It is clear that museums and universities have much to gain from the continued use of their collections. A host country also gains from visits by researchers, who bring foreign

currencies. The potential loss of such revenue through access to primary data instead of originals worried several participants. One potential mechanism to address this was presented by Gerhard Weber. At the University of Vienna, Weber and colleagues have CT-scanned fossils from around the world. The resulting data are sold online and a portion of the sale price is directly returned to the institution that houses the specimen.

After two days of positive discussion and debate, concluding points were universally accepted. These conclusions were in no way intended as declared resolutions, but rather general points of agreement and suggestions for the best way to move forward. Furthermore, the group was not prescribing any restrictions on, or recommendations for, the standardization and formatting of data. The focus was only on the sharing of ideas. One major conclusion of the workshop was support for the creation of an online "portal" that would allow researchers to search many databases from different institutions using a single, agreed-upon interface at the front end. As a preliminary first step, participants urged the posting of a website collating links to existing databases, including but not limited to those demonstrated at the conference. The next step will be designing and implementing a framework for a portal, with database field equivalency being a major goal. The "Paleoportal" website linking a series of paleontological collections across multiple institutions via DiGIR software (http://digir.sourceforge.net/) was the primary example. To begin the process, the domains http://www.paleoanthportal.org and http://www.paleoanthportal.eu have been set aside for mirror sites independent from any other organization.

Another area of agreement was the importance of benefits to those who participate in database-related activities. On one hand, institutions housing specimens should receive some type of benefit or value from the study of those collections, particularly collection of primary data. Possible benefits could include financial compensation, help with the development of their own databases, and/or help with software and hardware for viewing or reading such data. It was stressed that this approach could apply not just to institutions in developing countries (for example, African and Asian museums), but to all museums around the world.

On the other hand, individual contributions to collective databases should also come with some type of benefit. For example, data contributed by an individual researcher that is publicly available could be considered a citable electronic publication. This consideration would be an incentive for researchers to contribute their data to such databases, which in turn would enhance the sustainability of these databases. Its value would hinge on recognition of such publications by academic departments and tenure committees.

The long-term survival of databases was a major concern of all participants. Databases focused on and maintained by specific museum collections are theoretically sustainable because such institutions commit to the care and maintenance of those collections. Databases constructed by individuals or research groups, however, are vulnerable in the long-term. This is particularly true if they represent the results of finished projects rather than ongoing research. Once funding and support for these projects ceases, the databases may be lost. Potential solutions include endowing such databases or placing them within the cyber-infrastructure of larger institutions.

Finally, funding agencies worldwide were urged to encourage researchers to disseminate data widely, whether individually or by deposit into existing databases. Continued communication among the participants and other colleagues is obviously expected to lead to further collaboration in this area. Given that their databases already overlap in content and design, Delson and Harcourt-Smith have planned to meet with Wood, Gordon, Potts, Tocheri, Gilbert, and Polly to discuss field equivalencies across their databases and ways of beginning portal-style integration. The database "skeletons" of the RHOI, PRIMO, HOD and other systems will be made available to act as templates or starting points for colleagues to house their respective data. Other widely used database systems and formatting protocols are being investigated as possible guidelines for the further development of paleoanthropological databases. These include, for example, GEON (http://geongrid.org) and the Darwin Core (http://wiki.tdwg.org.twiki/bin/view/DarwinCore/WebHome). A wide variety of other databases in cognate fields is posted on the PaleoanthPortal links page. It is clear that databasing and data sharing are major components of the transformation of cyber-infrastructure that will characterize scientific research in most fields for the foreseeable future.

Eric Delson
Department of Anthropology
Lehman College, City University of New York and Department of Vertebrate Paleontology
American Museum of Natural History
E-mail: eric.delson@lehman.cuny.edu

William E. H. Harcourt-Smith
Department of Vertebrate Paleontology
American Museum of Natural History
E-mail: willhs@amnh.org

Stephen R. Frost
Department of Anthropology
University of Oregon
E-mail: sfrost@uoregon.edu

Christopher A. Norris
Department of Vertebrate Paleontology
American Museum of Natural History
E-mail: norris@amnh.org

© 2007 Wiley-Liss, Inc.
Published online in Wiley InterScience (www.interscience.wiley.com).
DOI 10.1002/evan.20141