DATA MODELS AND GLOBAL DATA INTEGRATION IN PALEOANTHROPOLOGY: A PLEA FOR SPECIMEN-BASED DATA COLLECTION AND MANAGEMENT

W. Henry Gilbert and Joshua P. Carlson

Abstract
The first decade of the 21st century has witnessed a dizzying proliferation of digital data and platforms. At the onset of this data revolution, romantic exaltation of global sharing was common, but this sentiment has rapidly faded into the realities of data architecture and database construction, and the next challenge, standardization, is coming rapidly into focus. Paleontology and paleoanthropology exemplify this trend. Many specimen catalogs are hosted online, but accessibility is significantly hindered by disjunction and ultimately by the inherent complexities of querying structurally different datasets. The way forward is the definition and use of standard objects, relationships, and fields. The institutions most directly responsible for the acquisition and management of physical specimens are best equipped to initiate this, especially projects collecting specimens and artifacts in the field. In this paper we present a basic data structure for paleoanthropology projects that integrates the logistics of specimen collection and curation with the ultimate goal of broad digital dissemination.

Keywords: database management, data collection, curation, digital dissemination

Introduction

Sharing data is good, but building and managing large datasets is costly (*Nature* 2007; "ETE Database Manual" 2010). Paleontology and archaeology are collections-based disciplines; specimens are collected, cataloged, and plotted in the field. Catalogs are revised thoroughly as curation, study, and finally publication, ensue. This is nothing new. From long before the age of the personal computer, investigators have archived project data in databases and sporadically published meaningful derivative subsets. This dissemination model continues to dominate scholarly communication. Increasingly, however, databases themselves are viewed as stable, citable repositories of knowledge, and database development is increasingly accorded appropriate credit.

Concern that selfish behavior was the primary hurdle to a world of unfettered data sharing makes less sense as the incentives for database publication converge on the rewards that stimulate publication of any research results. Straw men and stereotypes notwithstanding, current barriers to data distribution in paleoanthropology are less about selfishness than they are about sociology, logistics, and a lack of standardized data.

Any institution that hosts a paleontology database online has, by necessity, developed its own data standards. Several functional multi-institution paleontology databases currently exist, and these enact standardization across institutions. The most thorough provide extensive data dictionaries and instructions for how collaborating repositories should structure the data they provide. But each of these multi-institution databases is an organizational island, and no uniform standards exist. Additionally, all are designed from the perspective of the data aggregator, not the project. This is unfortunate, because projects collect, record, analyze, and curate the physical specimens, and data archiving is more efficiently handled at the project level.

This paper outlines a standard data dictionary (a list of fields and a specification of the format of their contents) for field-based paleoanthropology projects that is designed to follow all of the phases of data generation and dissemination, including database publication. It is based on the collective field experience of many paleoanthropologists and also on a rapidly intensifying dependence on publicly available databases for citable secondary data. Managing the transformation of data collected in the field into archived, accessible knowledge requires organization and foresight. The collection of more and more standardized data occurs when field results are recorded with an awareness of final data organizational structure. Also, public databases are better engineered when their developers understand the logistics of fieldwork and museum curation. The data dictionary in this paper aims to promote reconciliation of these perspectives on paleoanthropological data.

The database schema presented here has deep roots. Its digital ancestry can be directly traced to Middle Awash fieldwork in 1981 and the Paleoanthropology Inventory of Ethiopia in 1988 and 1989. The Middle Awash project has developed the data model extensively since. The initial plan of the Revealing Hominid Origins Initiative (RHOI) in 2003 was to require sharing of basic specimen data for participation and funding eligibility (Black et al. 2007), but a survey of participating projects turned up an astonishing diversity of platforms, data schema, and basic field definitions that rendered this intended requirement untenable (Black et al. 2007). Responding to this, several associates of the Revealing Hominid Origins Initiative collaborated to transform the Middle Awash model into a relational database template, and first distributed it publicly in 2007 (Black et al. 2007; Gilbert & Black 2007, 2010). The deliverable, a FileMaker-based relational database template, was freely distributed as shareware, along with a detailed user manual (Black et al. 2007). Several RHOI-affiliated projects have adapted versions of the template to their needs. The data definitions presented in this paper are based almost exactly on the RHOI database template fields and the Middle Awash project database format, with a few minor amendments.

PALEONTOLOGY DATABASES AND DATA SHARING

To understand the broader nature and structure of publicly available paleontology data, we conducted a survey of some of the most utilized online databases and noted all epistemological discussion of data definitions, metadata, or data standardization published with these datasets. Depending on criteria, there are hundreds to thousands of digitally published vertebrate paleontology databases and spreadsheets, but we looked at only the most accessible of these. Our survey revealed a diversity of data structures, most of which were associated with single-institution collections.

A number of major research-oriented paleontology specimen collections serve some form of data publicly, but they vary considerably in data structure and depth. Catalog number, taxon, and location are the only universal fields. The minimalist American Museum of Natural History (AMNH) Division of Paleontology database and the Omo Kibish Faunal Database serve only 3 specimen data fields, whereas the National Museum of Natural History (NMNH) paleobiology collections database serves 34 fields ("Division of Paleontology Search the Database" 2007; "Search NMNH Collections" 2010; "Omo-Kibish Faunal Database" 2010). Other university research collections hosting online databases fall somewhere in between in terms of field number, including the Yale Peabody Museum (6 fields), the Florida Museum of Natural History at the University of Florida (17 fields), the University of Wyoming Department of Geology and Geophysics Collection of Fossil Vertebrates (22 fields), the Texas Natural Science Center at the University of Texas (12 fields), and the Middle Awash project specimen database (14 fields) ("Yale Vertebrate Paleontology - Online Catalog" 2007; "Collection of Fossil Vertebrates Database" 2010; "Search the Vertebrate Paleontology Master Database" 2010; "Vertebrate Paleontology Laboratory " 2010; Gilbert 2010). Only the University of Wyoming Collection of Fossil Vertebrates Database provided data definitions.

Some online resources aggregate data from numerous institutions and localities, and thus engage in some degree of data standardization. Among resources surveyed, the Paleontology Portal ("Paleoportal"), MIOMAP, FAUNMAP, the NMNH Evolution of Terrestrial Ecosystems (ETE)/Paleobiology Database and Neogene of the Old World (NOW) Database of Fossil Mammals are aggregators of collection and locality information (Kaufman & Passarotti 2003a; Carrasco et al. 2005; „Search NMNH Collections" 2010; Fortelius 2010; Graham & Lundelius Jr. 2010; Alroy et al. 2010a). With the exception of Paleoportal, all of the above standardize metadata via either the ETE/Paleobiology Database or FAUNMAP/MIOMAP data conventions (Carrasco et al. 2005; Graham & Lundelius Jr. 2010). Both conventions provide data dictionaries (Kaufman & Passarotti 2003b; „ETE Database Manual" 2010; Alroy et al. 2010b). While the ETE architecture has a more extensive array of fields for locality than we present here, many of which are useful, they omit several topographic and logistical fields that are relevant to paleoanthropology fieldwork, including lack of any reference to archaeology. Also, they include many fields that will be used only very rarely or overlap with other fields and are thus inefficient. More importantly, neither the ETE database nor FAUNMAP/MIOMAP manage specimen data, but rather manage localities and faunal lists. While localities are logistically and organizationally important, they are, by definition, inconsistently bounded, and locality-derived faunal lists are invariably differently collected aggregates of specimens. Localities and faunal lists are therefore poor choices for the fundamental units of any research or database project.

Finally, although neither a collections database nor a specifically designed vertebrate paleontology database, Darwin Core ("Darwin Core" 2009) deserves mention in any discussion of standardization of biological specimen collections. The initiative provides a series of standards by which bioinformatics and collection data are recorded, presented, and shared at the specimen level. It is essentially a global, malleable data schema. The initiative recommends that its standards be modified to suit any biological sub-discipline, vaguely defining a customizable set of 'terms.' Darwin Core is wiki-based, and allows users to develop new terms and modify old ones. It requires that any new metadata terms consider the taxonomy of previously established terms, amending them to encompass the new term's description without changing the meaning of the existing term, or when this is not possible, by establishing a new term. Updates are moderated by the Technical Architecture Group, the Darwin Core Project's arbiter of official term standards ("Darwin Core" 2009). The broad, malleable approach taken by the Darwin Core Project provides meaningful guidance on large-scale data integration.

It is relatively straightforward to develop applications or websites that query multiple collections, so long as data structure is standardized. While there are numerous fossil databases online, few take on issues of data standardization, and none address the efficiency increases brought about by initiating standardization with collection. The inherent data aggregator bias we observed is an effect of history, not holistic wisdom. Projects are the entities best equipped to generate and maintain data. Here we define standard fields and data definitions suited to the needs of field paleoanthropology projects, such that they can easily be transformed into query-ready datasets.

Data structure in paleoanthropology

Many aspects of paleoanthropological fieldwork are universal. Exploration and discovery are the first phases. Geology specimens are often taken along the first transects in order to begin building a chronostratigraphic framework. Geomorphological analysis is simultaneous because erosion patterns the landforms available for collection, as well as the specimens themselves. Additionally, geological specimens are often taken from places with no fossils or artifacts.

Paleoanthropology projects usually collect geological specimens, soil specimens, microfossils, paleobotanical specimens, invertebrate fossils, vertebrate fossils, and archaeological remains. While individual collected objects retain specific position (GPS) information, clusters of specimens circumscribed by stratigraphy and/or exposure are common. These spatial and stratigraphic clusters range from enormous, as at Koobi Fora (Leakey & Leakey 1977), to restricted quarry sites, like Atapuerca Sima de los Huesos (Arsuaga et al. 1997), but are all usually designated as 'localities.' So, while very useful and often analytically meaningful, 'localities' are effectively administrative designations, and may be modified as work progresses. Without forethought, this evolutive process can lead to significant confusion and data damage. Worse, comparisons of locality-based sampling can be scientifically misleading.

Imprecise usage of the term 'site' for everything from large fossil- and artifact-rich regions to individual occurrences has traditionally been a source of even more confusion than variably defined localities. This problem is discussed in detail in Wolde Gabriel et al. (1992).The term 'area' is there suggested to be used to informally refer to any geographically bounded set of fossil and/or artifact-rich horizons. They define the term 'locality' as a "discrete paleontological or archaeological occurrence within an area of paleoanthropological significance. Localities are defined on the basis of geological, paleontological, and archaeological contents of delimited outcrops of one or a few stratigraphic horizons" (Wolde Gabriel et al. 1992, 474). In the terminology of Wolde Gabriel et al. (1992), localities are always within areas. This fundamental distinction is organizationally useful.

The boundaries of localities are stratigraphic and spatial. Therefore, localities are often named at natural, discrete unconformities and at the boundaries of physically linked outcrops. Several informal zones may exist within a single locality (excavated areas, sieve operations, etc.), and sometimes more than one distinguishable stratigraphic horizon may be exposed in a single locality. From a data management perspective, boundaries of localities are formally defined by a cluster of specimens. On the ground, however, they are bounded at depositional hiatuses and where either overburden or vegetation obscures traceable marker horizons. Some stratigraphic disconformities are more obvious than others, and often sustained stratigraphic work reveals pertinent information after the initial establishment of locality borders, so locality boundaries sometimes change (this is why it is advantageous to set the chronostratigraphic framework before sampling of fossil specimens and artifacts occurs, but this ideal is rarely met in real field situations). Regardless, care should be taken from the beginning to define and delimit localities precisely, and intense focus on individual geospatial data for each specimen collected is essential. For example, with precise spatial positioning, stratigraphic revisions can be applied to any collected specimen retroactively, should new geological information be gathered or the locality boundaries ever need to change, including mergers and subdivisions.

A data dictionary

The following data dictionary outlines a simple schema designed for efficient field entry, museum use, and queriability. Specimens are the fundamental unit of organization, and localities organize clusters of specimens (see Fig. 1). This basic organizational structure is common to many paleoanthropology projects.

Specimen table schema

- Project: Unique identifier of study area or project.
- Specimen type: Specimens will generally be of 6 types: Geology, vertebrate paleontology, archaeology, non-vertebrate paleontology, paleobotany, and microfossils. It is common, but not necessary, for projects to devise unique number schemes or abbreviations to distinguish among the types readily.
- Specimen number: Alphanumeric code that uniquely identifies specimen. Often, a specimen number contains metadata that allows for quick detection of the area of origin or current repository (e.g. KNM ER-1470). Avoiding characters that have a function in common query and markup languages (like quotes, percent signs, parentheses, semicolons, backslashes, etc.) is strongly advised. Although not absolutely necessary, a good rule of thumb is to use characters that would work in a MS-DOS filename. Care should be taken that no more than one individual is represented by each specimen number.
- Locality: Alphanumeric code that uniquely identifies locality.
- Elements preserved: List of skeletal elements preserved. Abbreviations are not recommended due to their incompatibility with keyword searches. Many projects use abbreviations, but there are not currently universal standards.
- Geological formation: Formal name of geological formation. Care should be taken to utilize the published geological formation name with established priority.
- Geological member: Formal name of geological member. Care should be taken to utilize the published geological member name with established priority.
- Stratigraphic horizon: Description of geological unit containing the specimen. Generally this description provides stratigraphic information beyond the geological

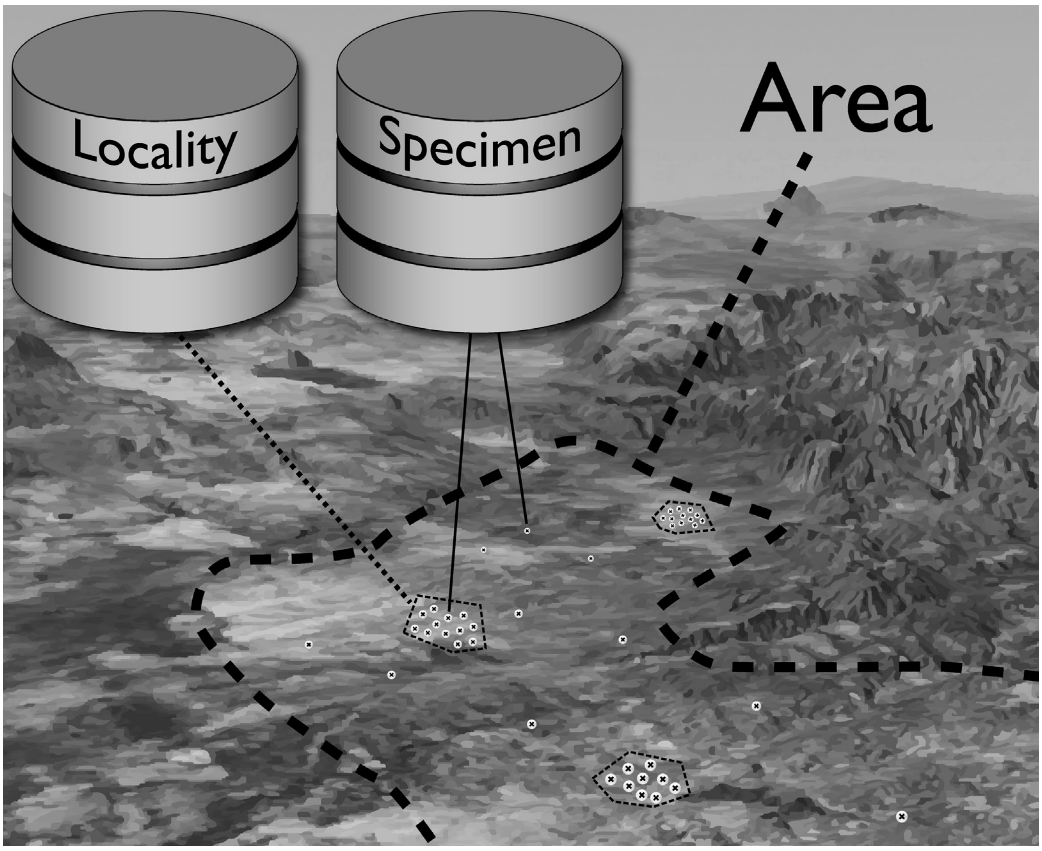

Fig. 1 General relationship of organized objects in a paleoanthropology project study area.

member, and care should be taken to provide as much detail as possible.

- In situ? (Y/N): Was the specimen recovered in situ? This is always a discrete yes or no answer.
- Estimated upper age limit (ma): Estimated specimen younger age limit in millions of years.
- Upper limit age basis: Basis for upper (younger) age limit estimate.
- Estimated lower age limit (ma): Estimated specimen older age limit in millions of years.
- Lower limit age basis: Basis for lower (older) age limit estimate.
- Sediment or matrix adhering?: Indication from adhering matrix or sediment that the specimen was extracted from an in situ position. This specifically does not apply to adherent recent residue.
- Repository: Institution formally recognized as owner of the specimen.
- Has a replica been made? (Y/N): Indication of whether a cast has been made of the specimen.
- Curatorial problem? (Y/Never/Corrected): Indication of occurrence of curatorial problem. See below for examples of curatorial problems. If a curatorial problem existed, but was later corrected, indication of this should be made such that the database becomes an archive of curatorial work with specimens through museum time.
- Curatorial problem description: Description of curatorial problem. Curatorial problems include lost specimens, corrected specimen numbers, specimens where MNI is found to be greater than one, specimens that have illegible numbers, specimens that have been loaned for which return is overdue, etc. Curatorial problem descriptions may indicate that the problem was corrected, but recorded problems should never be deleted and should be archived in curatorial notes upon correction of the curatorial problem.
- Curatorial notes: Curatorial notes include descriptions of corrected problems,

loans, movements, damage, and other phenomena affecting the physical disposition of the specimen.

- Taxonomic problem? (Y/N): Indication of occurrence of a taxonomic problem. Taxonomic problems generally result from discrepant identifications among specialists or from the inability of a collection manager to confidently identify a specimen to the level of precision perceived possible.
- Taxonomic identifier's names and dates: Names of identifiers and dates for identification of each taxonomic revision that triggers a change to the catalog.
- Taxonomic notes: Notes on taxonomic identification and revision. These are recorded by both specialists and catalog managers, and care should be taken by the catalog manager to acquire any notes from specialists.
- Taxon: It is advisable to establish a standardized taxonomic lookup table using a singular source to insure consistency in application of higher taxonomic nomenclature to identified specimens. It is useful to separate prefixes like cf., aff., and sp. from the Linnaean nomen to facilitate efficient data management: Class prefix, Class; Order prefix, Order; Suborder prefix, Suborder; Infraorder prefix, Infraorder; Superfamily prefix, Superfamily; Family prefix, Family; Subfamily prefix, Subfamily; Tribe prefix, Tribe; Genus prefix, Genus; Species prefix, Species; Subspecies prefix, Subspecies.
- Tool type/description: Type of artifact collected.
- Technique/industry: Name of industry or technique of manufacture. This field can be more verbose than 'Tool type.'
- Raw material: Raw material of collected artifact.
- Artifact identifier's name and dates: Names of identifiers and dates of identification of each archaeology revision that triggers a change to the catalog.
- Archaeology notes: Notes on artifact identification and analysis. These are recorded by both specialists and catalog managers, and care should be taken by the collection manager to acquire any notes from specialists.
- Formal excavation: Identifier of formal excavation from which specimen was derived.
- Collection procedure: Specific procedure used to obtain specimen.
- Reason for collection: Reason for collecting specimen.
- Collector's name: The collector is the person who found or encountered the specimen. While under some circumstances it may not be correct to assign discovery to a single person, as, for example, in excavations, it is not advisable to compromise this field by including more than one person's name. In our experience, it is either a single person or simply 'group.'
- Collection date: Date of collection of specimen.
- Latitude: Latitude in a standardized format (for example, decimal degrees).
- Longitude: Longitude in a standardized format (for example, decimal degrees).
- Geographic imagery reference: Reference to satellite or air photo imagery source that records geographic origin of the specimen.
- Specimen geographic location notes: Notes on geographic origin of specimen.
- Geography reference provider: Name of remote geographic information provider. For example, if a DGPS service is used, the name of the service provider and the correction signal type and frequency should be entered. If standard GPS is used, it should be noted here.
- Geography reference date: Date of geographic reference acquisition.
- Geography reference type: GPS, DGPS, total station, georeferenced image, etc.
- GPS unit (if applicable): Brand and model of GPS unit.
- Geo-reference notes: Notes on method used for georeferencing, including information on grid or alternative coordinate system.
- Elevation: Elevation in meters.
- Elevation type: Basis of elevation. Examples include GPS, altimeter, and topographic map.
- Elevation provider: Specific information on elevation provider. Examples include cartographer or DGPS service provider.

- Elevation date: Date elevation was recorded.

Locality table schema

- Project: Unique identifier of study area or project.
- Locality ID: Alphanumeric code that uniquely identifies locality. Avoid characters that have a function in common query and markup languages (like quotes, percent signs, parentheses, semicolons, backslashes, etc.). Although not absolutely necessary, a good rule of thumb is to use characters that would work in a MS-DOS filename.
- Locality common name: Name used to refer to locality. Often this is the local name of the vicinity of the locality.
- Geological member: Formal name of geological member to which the locality belongs. Care should be taken to utilize the published geological formation name with established priority. Only one member should be entered. If a locality exposes more than one formal member, the less common member(s) should be discussed in Stratigraphic interval.
- Geological formation: Formal name of geological formation. Care should be taken to utilize the published geological formation name with established priority (Salvador 1994)
- Stratigraphic interval: Detailed description of stratigraphic interval represented at the locality. Care should be taken to avoid redundancy between stratigraphic interval description and locality description, although some overlap is inevitable. This field has no length limitations and should include as much detail as possible.
- Estimated upper age limit (ma): Estimated locality younger age limit in millions of years.
- Upper limit age basis: Basis for upper (younger) age limit estimate.
- Estimated lower age limit (ma): Estimated locality older age limit in millions of years.
- Lower limit age basis: Basis for lower (older) age limit estimate.
- Locality description and boundaries: Description of locality using landscape features. Care should be taken to avoid botanical features and to minimize redundancy between locality description and stratigraphic interval.
- Locality dimensions (N/S): Approximate north to south locality dimensions in meters.
- Locality dimensions (E/W): Approximate east to west locality dimensions in meters.
- Uncollected taxa present: List of taxa represented among uncollected fossils. Taxonomic nomenclature should follow the taxonomic code adopted for the database.
- Archaeological evidence: Detailed description of archaeological evidence associated with the site.
- Macrobotanical evidence: Detailed description of non-microscopic botanical evidence associated with the site.
- Locality discoverer: Person or persons responsible for discovering locality. Unlike specimens, because it is potentially unrealistic to establish the first specimen found, localities may list the names of more than one discoverer.
- Locality discovery date: Date of locality discovery.
- Locality recorder: Person recording locality information.
- Locality recording date: Date of locality information recording.
- Latitude: Latitude of approximate center of locality.
- Longitude: Longitude of approximate center of locality.
- Geographic imagery reference: Reference to satellite or air photo imagery source that records geographic position of the locality.
- Geography reference provider: Name of remote geographic information provider. For example, if a DGPS service is used, the name of the service provider and the correction signal type and frequency should be entered. If standard GPS is used, it should be noted here.
- Geography reference date: Date of geographic reference acquisition.
- Geography reference type: GPS, DGPS, georeferenced image, etc.

- GPS unit (if applicable): Brand and model of GPS unit.
- Geo-reference notes: Notes on method used for georeferencing. Excavation grids, total station systems, and other sub-locality geo-spatial systems should be mentioned here.
- Elevation: Elevation in meters.
- Elevation type: Basis of elevation. Examples include GPS, altimeter, and topographic map.
- Elevation provider: Specific information about elevation information. Examples include topographic map cartographer and DGPS service provider.
- Elevation date: Date elevation was recorded.
- Field notes: Notes on locality taken in field.
- Repository notes: Notes on locality derived from museum work.
- Data notes: Notes about locality data handling.

Implementation

Fundamentally, the data dictionary presented above encompasses elements ('fields') that differ little from those traditionally utilized by field paleoanthropology projects. No attempt was made to include additional terms beyond those most efficiently employed in the field and laboratory. In an ideal world of universal data standardization, one in which every project or institution kept a digital database of all fossils, each project's responsibility would be to serve a queriable database of specimens with complete and accurate contextual, taxonomic, and curatorial data. The aggregator's role would be to design queries and interfaces that best compile, visualize, and relate data from the standardized global dataset. In this ideal situation, two citations would be generated when data was utilized, one for the aggregator and one for the project.

While this ideal will likely never be realized, it is clear that we are moving rapidly in the direction of increasing data integration. Under these conditions, data structure in paleoanthropology will inevitably converge on consistently utilized terms and a relational structure reflecting the real-world relationships of the organized entities. Anticipating these terms is a benefit to any project interested in preservation of the information it generates.

Conclusions

The most broadly utilized paleontology data aggregating initiatives, like the ETE Database or FAUNMAP/MIOMAP, are the outgrowth of a data integration movement that placed sharing at the top of the priorities list, one motivated by the promise of a linked, global paleontological dataset. Field projects do not have the same perspective. Each exists as a funded entity for its individual merit, operating under the rules and regulations of multiple institutions and governments and primarily focused on generating data to answer specific research questions. In the past, this has provided insufficient motivation for database efforts, particularly collaborative and standardized ones. Public databases, however, are increasingly citable, and thus even assessable for their scientific impact based on the number and profile of citations (see, for example, the Rat Genome Database (Twigger et al. 2007; Gilbert 2010)). There now exists a firm basis for assigning academic merit to people or projects electronically publishing data in relational tabular format. This fundamentally changes the scholarly rewards and economics of data sharing, making it more plausible for projects to invest in publishing datasets online. Of course, this economic shift is also a potential benefit to data aggregators. They can now start shifting the burden of data quality control and maintenance to projects and focus increasingly on development of queries, visualizations, and larger network topologies that stand at the heart of using the larger data sets to answer scientific questions.

More universal data standards directly benefit projects, data aggregators, and the larger paleoanthropological community. This paper presents a data dictionary and data schema that evolved in the face of strong field testing for over 20 years, then transformed into a publicly shared relational database template and user manual (RHOI), and finally successfully converted into the first publicly accessible database of hominids and hominid-associated vertebrate fossils from the African Rift (Twigg et al. 2007; Gilbert 2010). Most paleoanthropology projects will have very similar data structure needs, although most projects will have singularities that require customization. We offer the data schema and dictionary presented here as a humble step in the development global data standards in paleoanthropology.

Acknowledgements

The data dictionary outlined in this paper is based on the work of a large number of field paleoanthropologists working on the Middle Awash Project, the Paleoanthropology Inventory of Ethiopia, and a consortium of project leaders surveyed by the Revealing Hominid Origins Initiative. The authors would like to thank the following individuals for helpful comments and advice: Tim White, Michael Black, Ben Mersey, Jackson Njau, Kyle Brudvik, Socorro Baez, Raul Castillo, and Ray Bernor. This material is based upon work supported by the National Science Foundation under Grant Nos. SBR-9318698, SBR-9512534, SBR-9714432, BCS-9910344, BCS-0321893, and 0521538.

W. Henry Gilbert
Anthropology Department
3097 Meiklejohn Hall
California State University
Hayward, CA 94542

and

Human Evolution Research Center
3101 Valley Life Science Building
University of California
Berkeley, California
U.S.A.
henry.gilbert@fossilized.org
(510) 414-7239

Joshua P. Carlson
Human Evolution Research Center
Department of Integrative Biology
3101 Valley Life Sciences Building
University of California, Berkeley
Berkeley, CA 94720

References cited

Alroy, J., Kosnik, M. & Peters, S. (2010a). The Paleobiology Database Collection Search Form.
http://paleodb.org/cgi-bin/bridge.pl?user=Guest&action=displaySearchColls&type=view. 2010-10-23.

Alroy, J., Kosnik, M. & Peters, S. (2010b). The Paleobiology Database Frequently Asked Questions.
http://paleodb.org/cgi-bin/bridge.pl?user=Guest&action=displayPage&page=paleodbFAQ. 2010-10-23.

Arsuaga, J. L., Martínez, I., Gracia, A., Carretero, J. M., Lorenzo, C., García, N. & Ortega, A. I. (1997). Sima de los Huesos (Sierra de Atapuerca, Spain). The site. *Journal of Human Evolution* 33: 109-127.

Black, M., White, T., Brudvik, K., Su, D., Gilbert, H. & Boisserie, J. R. (2007). *RHOI Database Template User Manual*.

Carrasco, M. A., Kraatz, B. P., Davis, E. B. & Barnosky, A. D. (2005). Miocene Mammal Mapping Project (MIOMAP).
http://www.ucmp.berkeley.edu/miomap. 2010-10-27.

Collection of Fossil Vertebrates Database (2010).
http://paleo.gg.uwyo.edu/Search_DB.php. 2010-10-22.

Darwin Core (2009).
http://rs.tdwg.org/dwc/index.htm. 2010-10-05.

Division of Paleontology Search the Database (2007).
http://research.amnh.org/paleontology/search.php?search=Hominidae&media_only=-1&page=0. 2010-10-24.

ETE Database Manual (2010).
http://www.mnh.si.edu/ete/ETE_Database_Manual.html. 2010-10-22.

Fortelius, M. (2010). Database of Fossil Mammals.
http://www.helsinki.fi/science/now/index.html. 2010-11-12.

Gilbert, H. (2010). Middle Awash Specimen Database.
http://middleawash.berkeley.edu/middle_awash/specimen_db/query.php.
2010-10-22.

Gilbert, H. & Black, M. (2007). *The RHOI Database Template. Databases, Data Access, and Data Sharing in Paleoanthropology*. American Museum of Natural History, Wenner Gren Foundation, New York.

Gilbert, H. & Black, M. (2010). Data Models and Data Integration in Paleoanthropology. In: *Pleistocene Databases: Acquisition, Storing, Sharing. Proceedings of the workshop held at the Neanderthal Museum Mettmann, Germany from the 10th to 11th of July, 2010.* NESPOS Society and Neanderthal Museum Foundation, Mettmann.

Graham, R. W. & Lundelius Jr., E. L. (2010). FAUNMAP II: New data for North America with a temporal extension for the Blancan, Irvingtonian and early Rancholabrean. FAUNMAP II Database, version 1.0.
http://www.ucmp.berkeley.edu/faunmap/about/index.html. 2010-10-31.

Kaufman, S. & Passarotti, M. (2003a). PaleoPortal Collections Search.
http://www.paleoportal.org/portal/ 2010-10-23.

Kaufman, S. & Passarotti, M. (2003b). Collections Data Sharing and Data Use Agreements.
http://www.paleoportal.org/index.php?globalnav=footer_pages&sectionnav=data_agreements. 2010-10-23.

Leakey, M. G. & Leakey, R. E. (1977). *Koobi Fora Research Project, Volume 1: The Fossil Hominids and an Introduction to their Context 1968 - 1974*. Oxford University Press, Oxford.

Nature (2007). The Database Revolution. *Nature* 445: 229-230.
doi:10.1038/445229b

Omo-Kibish Faunal Database (2010).
http://turkanabasin.org/projects/omo-kibish/faunal-database. 2010-10-28.

Salvador, A. (1994). *International stratigraphic guide: a guide to stratigraphic classification, terminology, and procedure*. Geological Society of America, Trondheim, Boulder.

Search NMNH Collections (2010).
http://collections.nmnh.si.edu/search/paleo/. 2010-10-22.

Search the Vertebrate Paleontology Master Database (2010).
http://www.flmnh.ufl.edu/scripts/dbs/vp_uf_pub_proc.asp. 2010-10-22.

Twigger, S. N., Shimoyama, M., Bromberg, S., Kwitek, A. E. & Jacob, H. J. (2007). The Rat Genome Database, update 2007—Easing the path from disease to data and back again. *Nucleic Acids Research* 35: D658-D652.

Vertebrate Paleontology Laboratory (2010).
http://www.npl.utexas.edu/vpl/databases/index.php?mode=search&action=search. 2010-10-22.

Wolde Gabriel, G., White, T., Suwa, G., Semaw, S., Beyene, Y., Asfaw, B. & Walter, R. (1992). Kesem-Kebena: a newly discovered paleoanthropological research area in Ethiopia. *Journal of Field Archaeology* 19: 471-493.

Yale Vertebrate Paleontology - Online Catalog (2007).
http://peabody.research.yale.edu/COLLECTIONS/vp/. 2010-10-24.