Digital Data Collection in Paleoanthropology

DENNÉ REED, W. ANDREW BARR, SHANNON P. MCPHERRON, RENÉ BOBE, DENIS GERAADS, JONATHAN G. WYNN, AND ZERESENAY ALEMSEGED

Understanding patterns of human evolution across space and time requires synthesizing data collected by independent research teams, and this effort is part of a larger trend to develop cyber infrastructure and e-science initiatives.[1] At present, paleoanthropology cannot easily answer basic questions about the total number of fossils and artifacts that have been discovered, or exactly how those items were collected. In this paper, we examine the methodological challenges to data integration, with the hope that mitigating the technical obstacles will further promote data sharing. At a minimum, data integration efforts must document what data exist and how the data were collected (discovery), after which we can begin standardizing data collection practices with the aim of achieving combined analyses (synthesis). This paper outlines a digital data collection system for paleoanthropology. We review the relevant data management principles for a general audience and supplement this with technical details drawn from over 15 years of paleontological and archeological field experience in Africa and Europe. The system outlined here emphasizes free open-source software (FOSS) solutions that work on multiple computer platforms; it builds on recent advances in open-source geospatial software and mobile computing.

The past 20 years have seen a dramatic increase in the productivity of paleoanthropology *sensu lato* (that is, human paleontology and paleolithic archaeology). This increase (Fig. 1) comes from the discovery of new fossil and archeological sites,[2–4] increased collection efforts across all sites, novel analytical methods, and a wealth of new paleoclimate, geological, geochronological, and paleoenvironmental data that can be synthesized to evaluate richer hypotheses of human origins and adaptation.[5–9]

Museums and research institutions, including those in developing countries

Denné Reed is a paleoanthropologist at the University of Texas at Austin studying the influences of ecology and environment on human adaptation. He has conducted field research in eastern Africa, southern Africa, and Morocco. As the director of the PaleoCore website and data repository, Dr. Reed is working to integrate paleoanthropological data in order to address broad-scale questions about human evolution and environmental change.

W. Andrew Barr is a paleoecologist and paleoanthropologist at the George Washington University in Washington, D.C., whose research focuses on understanding the environmental and ecological context of early human evolution. He conducts field research in Ethiopia with the Mille-Logya Project, and in addition, he is a member of the core development team of the PaleoCore project.

Shannon McPherron is a Paleolithic archaeologist at the Max Planck Institute for Evolutionary Anthropology in Leipzig. He is primarily interested in the evolution of hominin cultural abilities from the origins of stone tool use through to the dispersal of modern humans. He has excavated a number of Paleolithic sites and is currently involved in field projects in southwest France (Abri Peyrony and La Ferrassie), in Morocco (Jebel Irhoud and Rhafas), and in Ethiopia (Dikika and Mille-Logya).

René Bobe is at the Institute of Cognitive & Evolutionary Anthropology at Oxford (UK). He studies the ecological and environmental context of early hominins in Africa, and the relationship between climatic change and evolutionary processes. His current field projects include work in the Afar Triangle (Ethiopia), the lower Omo Valley (Ethiopia), and the Lake Turkana Basin (Kenya). In parallel to his work in eastern Africa, Dr Bobe is establishing a new research project in the Andes of Patagonia, where he studies the origins and evolution of platyrrhine primates and other South American mammals.

Denis Geraads is at the Muséum National d'Histoire Naturelle in Paris. He is interested mainly in the paleontology and evolution of large, Old World mammals, including ruminants, rhinos, and carnivores. He has worked in France, Bulgaria, Greece, Turkey, Morocco, Algeria, Tunisia, Chad, Djibouti, Ethiopia, and Tanzania.

Jonathan Wynn is a geologist and stable isotope geochemist at the University of South Florida. Dr. Wynn has worked on the geological context of human origins throughout the rift sedimentary basins eastern Africa. He studies sedimentary and pedogenic environments and uses isotopic studies of soil carbonates, lake sediments and tooth enamel to understand the role of environmental change in the course of hominin evolution.

Zeresenay Alemseged is at the California Academy of Sciences. He studies the origin and evolution of early human ancestors, and the environmental factors that influence their evolution. His objective: to unearth and analyze clues to their biology and behavior and to identify milestone evolutionary events that ultimately led to the emergence of modern Homo sapiens. To this end he leads international and multidisciplinary field and lab work including the Dikika and Mille-Logya Research projects and focuses on the study of hominins and other primates.

Key words: mobile device; spatial database; paleoanthropology; collection methods; Ethiopia; Dikika; Mille-Logya

DOI: 10.1002/evan.21466
Published online in Wiley Online Library (wileyonlinelibrary.com).

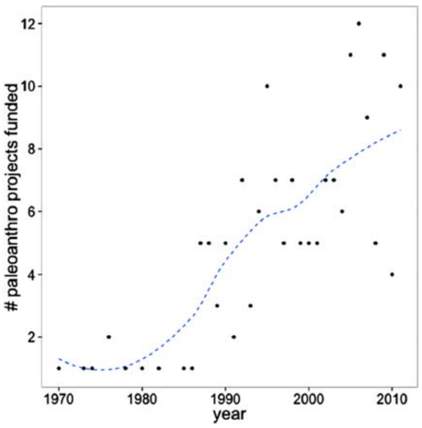


Figure 1. The number of paleoanthropology projects funded by the National Science Foundation over the past 40 years has increased. Data from http://nsf.gov. (Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.)

where much paleoanthropological research is conducted, are struggling to keep pace with the discoveries. Several symposia and workshops have addressed this problem.[10,11] Consequently, a consensus has emerged that sharing data is necessary and that data discovery and integration are desirable. It is now clear that paleoanthropology will advance more rapidly if teams can achieve consensus on important research objectives, collaboratively build on each other's findings, and develop standardized protocols for data collection and analysis; in other words, paleoanthropology must operate more like a "high-consensus, rapid-discovery" science.[12] This vision is consistent with cyberinformatic initiatives to link isolated information silos and make innovative use of Big Data.[13–15]

Furthermore, research teams have an ethical and professional obligation to safeguard the cultural heritage and scientific value of the materials they collect.[16] Preserving cultural heritage data requires investing in digital data management, which can greatly increase the quality and longevity of field data. Research improves when basic information (for example, location, date, time, collector) can be recorded automatically and attention can be focused on fewer important tasks. Digital data also facilitate discovery and reuse by other researchers, giving greater returns on the investment made in originally collecting the data.[17,18] However, a general, widely adopted solution has yet to emerge. The problem continues to grow as new data are collected but not digitized. This paper reviews available solutions to the problem of standardizing field data collection by outlining a system that is capable of handling spatial data in a way that increases analytical flexibility and promotes data integration.

LOCATION, LOCATION, LOCATION. . .

Paleoanthropological data are inherently spatial. The provenience of each fossil, artifact, and sample is fundamental to understanding its age, taphonomy, ecological context, cultural context, and behavioral meaning;[19] in this way, the spatial provenience of an item is often as important as the item itself. However, in the process of collecting an item, its context is destroyed.[20] For this reason, it is especially important that systems handle spatial and contextual data effectively.

Fortunately, global navigation satellite systems (GNSS) such as the U.S. global positioning system (GPS), the European Galileo system, and the Russian GLONASS system make it possible to record the real-world coordinates of any point to centimeter or even millimeter accuracy, depending on the effort invested. The question then shifts to whether it is practical to digitally record the location of all finds as they are collected.

RETHINKING THE PALEONTOLOGICAL LOCALITY AND THE ARCHEOLOGICAL SITE

Traditionally, paleontological data are collected by means of surface prospecting, using paleontological localities as the smallest unit of collection and analysis. It is important to note that the word "locality" has varying usage. In the tradition of North American paleontology followed here, a site contains localities, which are geographically and stratigraphically delimited. In many archeological contexts, the meanings of these words are reversed. For example, in Koobi Fora and West Turkana, Kenya, sites occur within localities, and the latter refers to a large area. In the context of European Paleolithic archeology, the word "locality" generally is not used at all.

In paleontology, the practice of using localities as the fundamental units of spatial and stratigraphic provenience speeds up fossil collection, but poses problems when localities vary in extent from a square meter to a square kilometer or vary greatly in the amount of time represented.[21] One solution is to record the location and stratigraphic position of each item; that is, piece-provenience. Under a traditional scheme of paleontological localities defined *a priori*, localities become *de facto* units of analysis, whereas piece-proveniencing allows items to be aggregated *a posteriori* into groupings suitable for a given analysis. If localities are desired for the sake of consistency or research tradition, one can always reaggregate piece-provenienced finds.

The collection and analysis of archeological data faces similar issues. Here the spatial unit typically is an excavation unit within a site with rectangular, horizontal (XY) units and a vertical (Z) unit. Units serve many practical purposes, but also help ensure commensurability between projects and between field seasons within a project. However, once spatial or chronological units are created, it may be difficult to restructure the data around newly defined units.[22,23]

Once defined, localities, sites, and stratigraphic units quickly take on an elevated status that becomes equated with the taxa or behaviors documented at that location, as if they took place exactly at the scale of the site, locality, or stratigraphic unit.[24] Redefining the site or the horizontal and vertical units within it to sample behaviors at a different scale can be very difficult given how data are typically collected and recorded. This is a significant problem because there is no way to guarantee that units defined during excavation or during fossil surveys will correspond with analytical units that are appropriate for the relevant research questions during analysis.[25]

The main challenge to the piece-proveniencing approach is the added time and energy needed to record information for each item. However, technology has increased the efficiency of piece-proveniencing to a point at which it is efficient and feasible. Also, the up-front cost of piece-proveniencing is offset by the

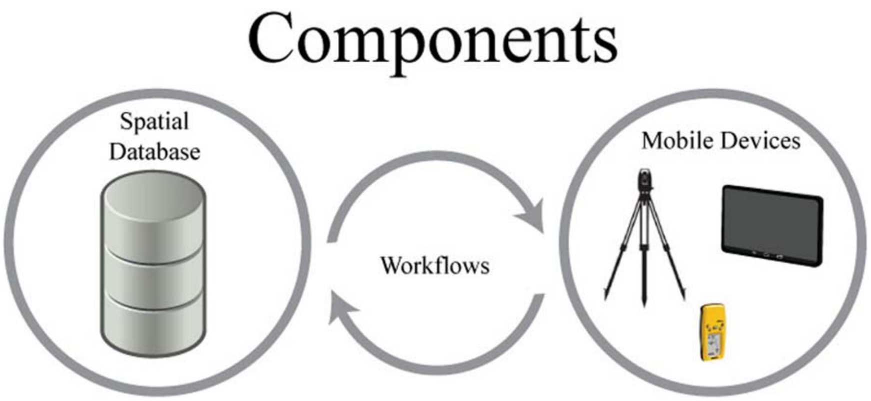


Figure 2. Schematic illustration of the three main components of the data collection system described here: the spatial database, the mobile devices for collecting data in the field, and the workflows that govern how data are collected and archived in the spatial database. (Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.)

immediate digitization of the data and reduced data entry later on.[26,27]

OVERVIEW OF THE DIGITAL DATA COLLECTION SYSTEM

The system outlined in this paper uses mobile devices to record the location and attribute data about each item and automatically transfers these data to a spatial database that serves as a digital catalog. The term "mobile device" refers broadly to any human-portable equipment capable of recording positional information and associated item-attribute data, including smartphones, tablets, GPS data loggers, total stations, and digital cameras. The system (Fig. 2) takes advantage of three innovations: the advent of spatial databases for storing and managing data, the ubiquity of inexpensive mobile devices with built-in GPS receivers, and software and programming tools for designing comprehensive and reliable data management workflows, which are systematized sets of actions and routines designed to reliably accomplish a task, such as a data migration workflow that governs how data are transferred from a mobile device to a database.

At the heart of the system, the spatial database serves as the main repository, a digital catalog for the project. New data are collected in the field on mobile devices and then, daily or at the end of a field campaign, are imported into the spatial database. A series of workflows dictates the processes by which data are collected and managed.

DATABASES

Since the 1970s, relational database systems have grown to become the most prevalent platform for managing information (see Box 1 for a primer on database terms and concepts). More recently, nonrelational data models, also called NoSQL for Not-only SQL database models, have gained ground because of their ability to handle massive datasets and semi-structured data in which the information that is present varies from one record to the next.[28,29] We focus on relational databases because most paleoanthropological research produces structured data.

Spatial Databases

A spatial database,[30] also known as a geographic database, geospatial database, or just a geodatabase, is an extension of a relational database, which is capable of storing a rich array of spatial information inside the same database tables that store nonspatial information.[31–33] For example, the spatial database design we describe includes an "Occurrence" table that stores basic information on every find (that is, fossil, artifact, geological sample, and such). The table includes columns for the collection date, item type (categorized as archeological, geological, or biological), collector, and other such information. In addition, a single column records the precise geographical location of the item in a format that a GIS program (or any other spatial analysis software) can read and map directly (Fig. 4).

The same database may contain a "Hydrology" table, which stores information on rivers and drainages, with the names of rivers, their size, and other relevant information alongside a column that stores the geographical location of a river segment, including the coordinates for all the vertices that comprise the river segment.

Each of these two tables represents a unique spatial dataset (Occurrence and Hydrology); both are stored inside the same database. Within each table, spatial data are stored alongside other attribute data. The database as a whole can contain many spatial and nonspatial datasets, such as a polygon spatial dataset representing permit areas, a nonspatial table storing taxonomic names and ranks, or a polygon spatial dataset storing the layout of an excavation grid.

Each feature (for example, fossil, artifact, drainage, collection area) occupies a row in a database table. The location of each feature is stored in the spatial column, using the geometry data type (Fig. 4). This column encapsulates all coordinates that make up a feature, such as the paired coordinates of a point feature, or sets of coordinate pairs describing the vertices of a line or polygon, along with information about the geodetic datum, projection, and coordinate system. In this way, the spatial database system is an improvement over storing coordinates in separate latitude and longitude columns. The binary data stored in the spatial column can be read directly by a GIS program without having to import or convert the data and can be represented in a standard, human-readable format such as Well-Known-Text (WKT).

Spatial analysis is nothing new to paleoanthropology or archeology. GIS programs are commonly used to analyze paleoanthropological data[1,34–38] and spatial databases are an integral part of contemporary GIS software. Yet spatial databases present an

Box 1. Database Primer

Relational Database Systems. Databases are systems for digitally managing, querying, and manipulating information. There is a rich assortment of database systems, but the most ubiquitous is the relational database system wherein a database consists of tables that store similar kinds of data (for example, an occurrence table or a lithics table). Using shared columns called keys, these tables are interrelated to one another. The software that manages a relational database is called a relational database management system (RDBMS).

Database Components. Each table in the database has an explicit structure or schema (Fig. 3). Tables comprise rows (also known as records or tuples) and columns (fields or attributes). In this way, database tables superficially resemble spreadsheets; however, a big difference is that databases require that columns contain information of a single data type, whereas spreadsheets allow mixtures of different data types in a column. Common data types for columns include alphanumeric characters, integers, decimal numbers (also called floating point numbers or floats), dates, boolean values (yes/no), and binary data such as images. The major feature of spatial databases is the inclusion of a geometry data type for geographical features such as point locations and polygon features.

Relationships. The tables in a relational database usually contain information about a single class of thing, such as fossil and artifact occurrences, localities, or collectors working on the project. The information in these different tables is interlinked through the use of keys (Fig. 3). Each table has a primary key, which is a column (or set of columns) containing a unique value for each row in the table. This primary key is then used by other tables in the database to relate records in tables to one another. When the columns of a table contain keys to other tables, these are called foreign keys. For example, an occurrence table may include a foreign key to the locality table to relate the find to its locality. When, for example, multiple finds come from the same locality, the relationship between the tables is called a many-to-one relationship. Similarly, there will be a many-to-one relationship between occurrences and an individual collector, because one person may find many fossils and artifacts. In addition, tables may also have one-to-one and many-to-many relationships. These relationships allow data to be managed in separate tables. They also enable rich queries to find related information, such as all fossils from a particular locality or the localities associated with fossils found by a particular person. The process of partitioning data across related tables is called database normalization; the process of querying data from related tables is called joining the tables.

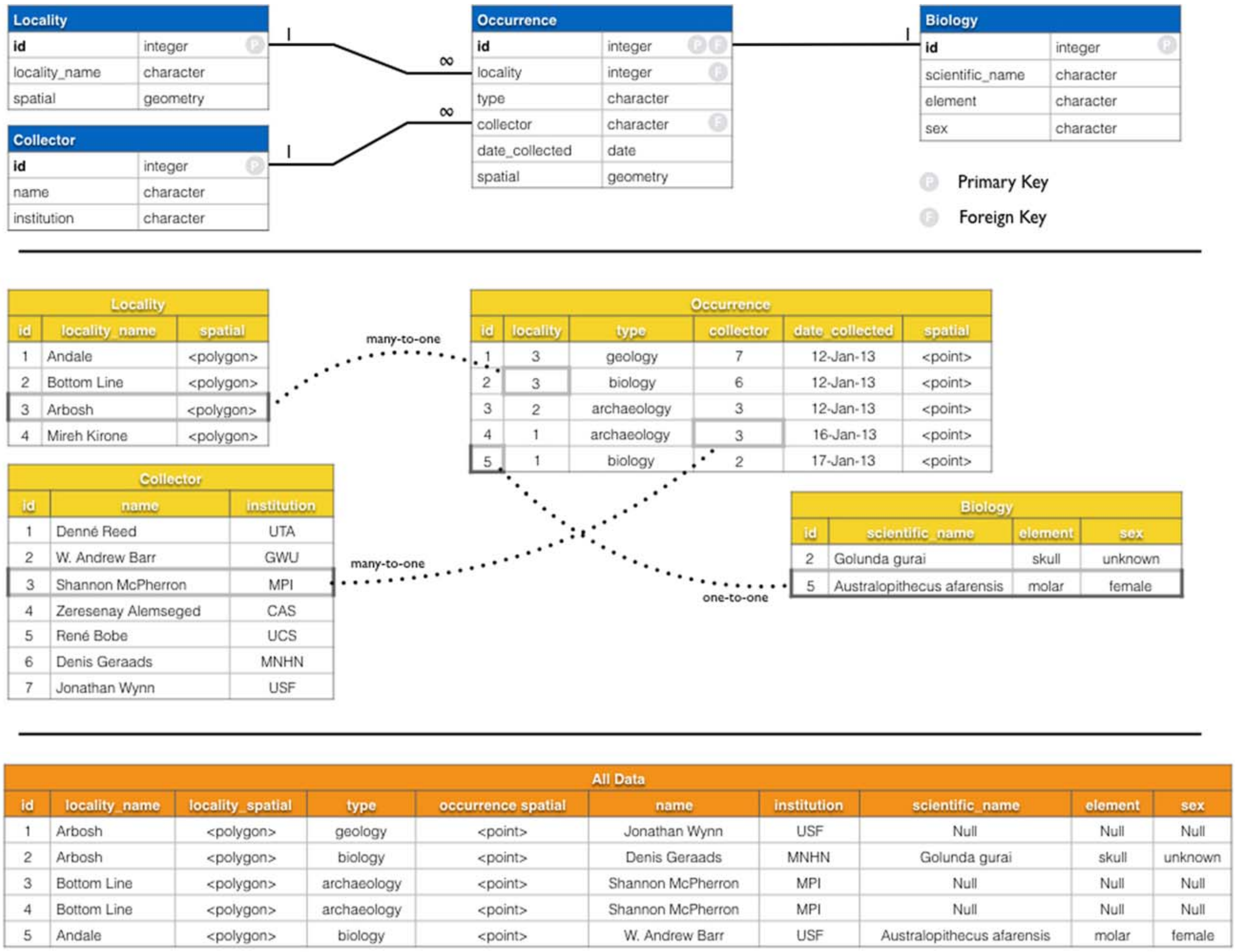
Structured Query Language. RDBM systems use a common programming language called structured query language (SQL). This programming language allows users to query, create, delete, and otherwise manipulate the data in the database and to modify whole databases, tables, users, and other entities in the database. Graphical user interfaces provide more intuitive means of manipulating data, but they, too, translate the point-and-click actions of the user into SQL commands.

opportunity to go beyond specialized GIS software, to operate as part of a spatial data infrastructure (SDI) that serves information to a variety of software applications (Fig. 5).[32] Decoupling the spatial database from a GIS application allows us to select the application most appropriate to a specific task. For example, research teams can collaboratively edit the data (spatial and nonspatial) through a web interface or perform analyses by connecting directly to the database with statistical software such as R.[39,40] An SDI is an extension of a spatial database system that includes the ability to connect to a spatial database in a variety of ways. This requires integrating the spatial database into a larger software system that can manage multiple types of connections. Table 1 provides a listing of open-source software packages that work well together in an SDI system. The PaleoCore website (http://paleocore.org) is an example of an SDI implementation for paleoanthropology that allows researchers to set up project databases, collaboratively manage them via the web, and connect to their data via GIS programs and R.

Database Schemas and Standards

The design and structure of spatial databases are important considerations when creating them. What information needs to be recorded in the database and how should that information be stored? One approach to good database design is to draw from existing examples, preferably ones that reference a standard, so that data can be synthesized later. Presently, however, there is no data **standard** for paleoanthropology, although there are published data structures that provide guidance on what data should be recorded and how.

Many paleontological data structures currently in use are derived, one way or another, from the seminal efforts of John Damuth,[41] who designed the Evolution of Terrestrial Ecosystems (ETE) Database to record occurrences of fossil species. The New and Old World (NOW) database[42] started as a clone of the ETE database. The Paleobiology database also was largely inspired by ETE. The ETE data structure is referenced in subsequent efforts by Gilbert and Carlson,[43] who published a data dictionary defining key data



Locality	
id	integer
locality_name	character
spatial	geometry

Collector	
id	integer
name	character
institution	character

Occurrence	
id	integer
locality	integer
type	character
collector	character
date_collected	date
spatial	geometry

Biology	
id	integer
scientific_name	character
element	character
sex	character

Locality

id	locality_name	spatial
1	Andale	<polygon>
2	Bottom Line	<polygon>
3	Arbosh	<polygon>
4	Mireh Kirone	<polygon>

Collector

id	name	institution
1	Denné Reed	UTA
2	W. Andrew Barr	GWU
3	Shannon McPherron	MPI
4	Zeresenay Alemseged	CAS
5	René Bobe	UCS
6	Denis Geraads	MNHN
7	Jonathan Wynn	USF

Occurrence

id	locality	type	collector	date_collected	spatial
1	3	geology	7	12-Jan-13	<point>
2	3	biology	6	12-Jan-13	<point>
3	2	archaeology	3	12-Jan-13	<point>
4	1	archaeology	3	16-Jan-13	<point>
5	1	biology	2	17-Jan-13	<point>

Biology

id	scientific_name	element	sex
2	Golunda gurai	skull	unknown
5	Australopithecus afarensis	molar	female

All Data

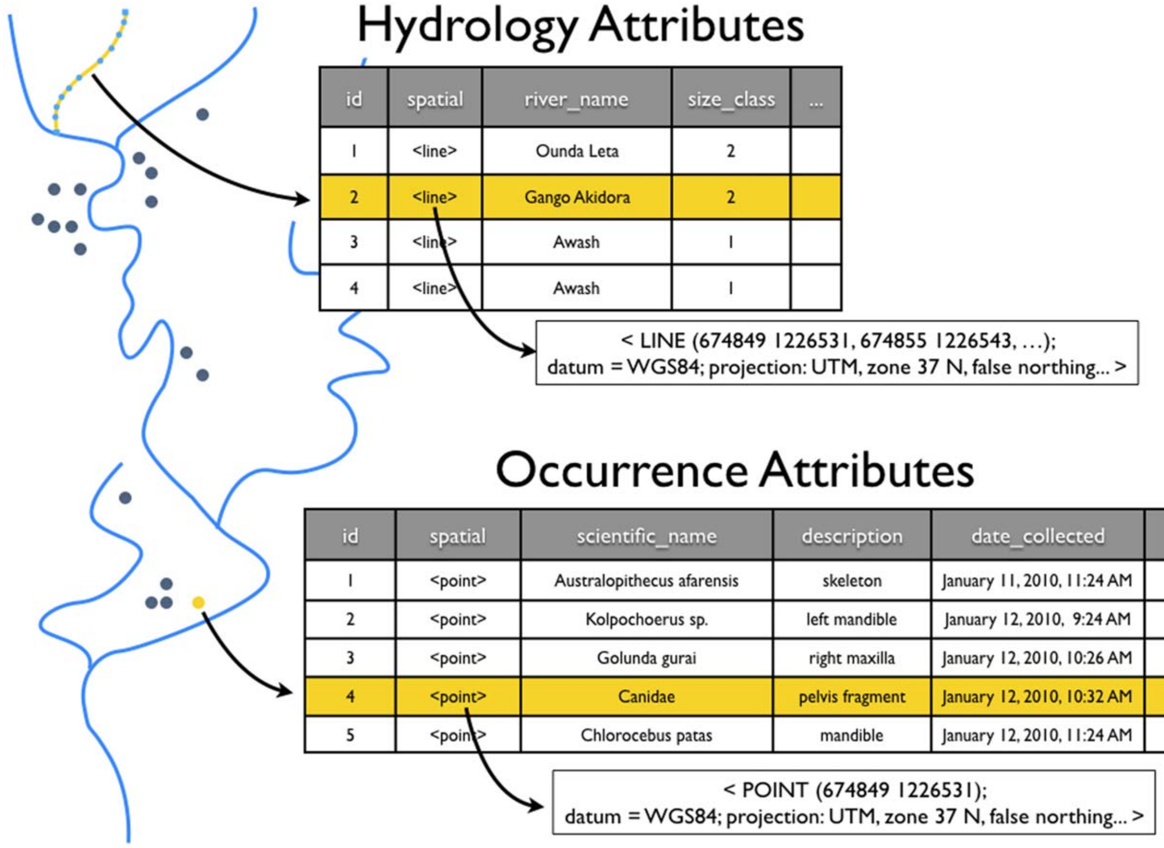
id	locality_name	locality_spatial	type	occurrence spatial	name	institution	scientific_name	element	sex
1	Arbosh	<polygon>	geology	<point>	Jonathan Wynn	USF	Null	Null	Null
2	Arbosh	<polygon>	biology	<point>	Denis Geraads	MNHN	Golunda gurai	skull	unknown
3	Bottom Line	<polygon>	archaeology	<point>	Shannon McPherron	MPI	Null	Null	Null
4	Bottom Line	<polygon>	archaeology	<point>	Shannon McPherron	MPI	Null	Null	Null
5	Andale	<polygon>	biology	<point>	W. Andrew Barr	USF	Australopithecus afarensis	molar	female

Figure 3. Example of a database structure. The top pane shows an example of a database schema using an Entity-Relationship Diagram (ERD). Each box represents a table in the database with the title across the top. The elements in each box represent columns. In this example, each table includes an id column that serves as the primary key for the table. The Occurrence table also contains several foreign keys with the lines indicating how the columns are related. The middle pane illustrates the tables with example data; the dotted lines indicate the related columns and the types of relations they represent. The bottom pane shows the same data in a nonrelational format. Note that data are unnecessarily duplicated and that this one table will quickly grow to have a cumbersome number of columns. This table also presents a risk of data loss. For example, if specimen 2 were deleted for some reason, all information about the collector, Denis Geraads, would also be lost. In the relational format, specimen 2 can be deleted without having any effect on information in the Collector table. These are just a few examples of the problems solved by using related tables. However, this also adds complexity because there are more tables to manage and the relationships between tables must be maintained. (Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.)

elements for paleoanthropology based on the data template developed for the Revealing Hominid Origins Initiative (RHOI). Independently, the FaunMap and MioMap Projects offer other examples of established data structures for paleontology.[44,45]

The ETE and related data structures were designed to track the occurrence and attributes of species, as opposed to specimens. Several software packages, each with its own data structure, exist for maintaining natural history collections. A popular example is Specify (http://specifyx.specifysoftware.org/), a free open-source, taxon-independent collection management system. Similarly, archeological data repositories such as Open Context, the Online Cultural and Historical Research Environment (OCHRE)[46] and the Digital Archaeological Record (tDAR) have established data recommendations, but none has reached the formal status of a data standard. In archeology there is a more energetic debate surrounding the usefulness of standards in that domain.[47]

Paleoanthropology can develop standards of its own by adopting elements from biodiversity informatics, which has mature data standards. A data standard comprises a list of terms or concepts with explicit definitions. The terms in a standard can be used to describe information in a domain. For example, the Darwin Core and Access to Biological Collections Data (ABCD), biodiversity standards maintained by the Biodiversity Information Standards Group (TDWG),[48,49] define terms that are useful for describing data on biological occurrences. Similarly, the Dublin Core Metadata Initiative (DCMI) maintains terms to describe books and other



id	spatial	river_name	size_class	...
1	<line>	Ounda Leta	2	
2	<line>	Gango Akidora	2	
3	<line>	Awash	1	
4	<line>	Awash	1	

< LINE (674849 1226531, 674855 1226543, ...);
datum = WGS84; projection: UTM, zone 37 N, false northing... >

id	spatial	scientific_name	description	date_collected	...
1	<point>	Australopithecus afarensis	skeleton	January 11, 2010, 11:24 AM	
2	<point>	Kolpochoerus sp.	left mandible	January 12, 2010, 9:24 AM	
3	<point>	Golunda gurai	right maxilla	January 12, 2010, 10:26 AM	
4	<point>	Canidae	pelvis fragment	January 12, 2010, 10:32 AM	
5	<point>	Chlorocebus patas	mandible	January 12, 2010, 11:24 AM	

< POINT (674849 1226531);
datum = WGS84; projection: UTM, zone 37 N, false northing... >

Figure 4. View of data in a spatial database. One hydrology and one occurrence feature are highlighted. Each occupies a row in its respective table. The coordinates for each feature are stored as geometry data in the spatial column of the data table. (Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.)

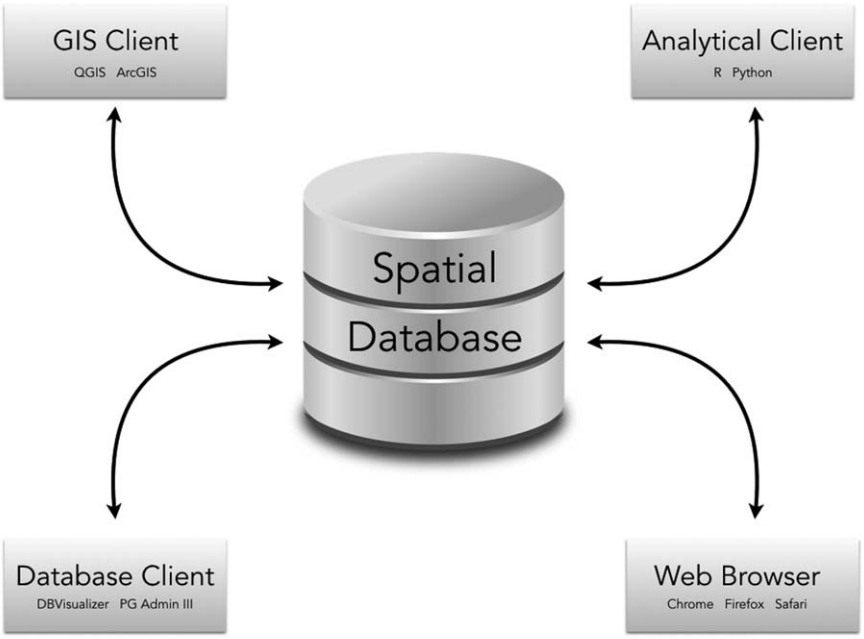


Figure 5. Illustration of the flexibility introduced by a spatial data infrastructure. Users may interact with a single spatial database through a variety of applications such as database clients, web browsers, or GIS applications.

library resources. Standards may also use terms from each other, as Darwin Core borrows several terms from Dublin Core. Conveniently, the Darwin Core and ABCD standards have been mapped to the data structures of some specimen management systems, including Specify. The PaleoCore project (http://paleocore.org/standard/) is currently developing an implementation using elements from Dublin Core, Darwin Core, ABCD, and others to create a set of standard terms and concepts suitable for paleoanthropology. Without standards, data remain isolated, or worse, present the risk of inappropriate data synthesis.[50,51]

MOBILE DEVICES

Mobile devices with GPS receivers, including smartphones, tablets, and total stations comprise the second key component in a digital data collection system. Choosing the appropriate

TABLE 1. Software Components for Digital Data Collection and Analysis

Software	Location	Description	License
PostgreSQL and PostGIS	http://www.postgresql.org/ http://postgis.net/	A full-featured, enterprise relational database management system. Spatial databases require installation of the PostGIS extension for PostgreSQL.	Free Open Source
SQLite and SpatiaLite	https://www.sqlite.org/ http://www.gaia-gis.it/gaia-sins/	A lightweight relational database management system. Spatial databases require installation of the Spatialite extension.	Free Open Source
PG Admin III	http://www.pgadmin.org/	A database management utility for PostgeSQL	Free Open Source
DBVisualizer	http://www.dbvis.com/	A universal graphical database management tool that works with PostgreSQL, SQLite, and most other relational database management systems.	Free Closed Source
Quantum GIS	http://www.qgis.org/	A GIS software package that can connect directly to PostgreSQL and SQLite spatial databases (that is, SpatiaLite).	Free Open Source
R with RODC, sp, rgdal, rgeos	https://www.r-project.org/ https://cran.r-project.org/	A statistical programming language. R libraries such as RODBC, sp, and rgdal allow connections to spatial databases and the manipulation of spatial data.	Free Open Source
R Studio	https://www.rstudio.com/	A graphical integrated development environment (IDE) for coding in R.	Free Closed Source
Python	https://www.python.org/	A programming language for scientific computing. It is also the main scripting language for QGIS and other GIS software packages.	Free Open Source

hardware requires balancing many competing factors, including price, weight, durability, and battery life. The built-in GPS receiver in most smartphones is capable of 3-5 meter precision, which is sufficient for surface survey. This can be improved by linking the device to an external GPS with differential correction to obtain submeter precision. Where centimeter or millimeter precision is needed, such as in excavation contexts, archeologists have, for many years, been using total stations for precise 3D coordinate data collection.[22,52–54] Whether by GPS receiver, total station, or a combination of the two, these devices make it feasible and efficient to piece-provenience items as they are discovered.

The operating system and software running on the mobile device have a big impact on the efficiency of data collection, and there is a fast-growing set of options. Table 2 lists some of the available apps designed for spatial data collection. Stand-alone apps such as GIS Pro are better suited to offline field collection, whereas the cloud-based solutions may be better in field areas where wifi or cellular network connectivity is available.

WORKFLOWS

A well-developed set of workflows is the third and perhaps most important component of the system. Workflows are the procedures and protocols that govern the data collection process. We focus on two main categories: collection workflows, which govern how researchers locate, document, and collect specimens in the field and data transfer workflows, which govern how data are migrated among mobile data collection devices, laptops, and servers. Additional workflows might govern how items are curated after collection, how measurements are made, or how images or 3D models are generated.

Collection Workflows

Paleoanthropologists use a wide variety of data collection methods in the field, ranging from *ad hoc* surface surveys to detailed excavations. As a first step, documenting the various collection methodologies is important as the key to unraveling collection biases and taphonomic processes.[55,56] Table 3 provides a hierarchical classification of collection methods and protocols arranged on a continuum of collection intensity. At each level, the collection protocol describes the minimum criteria for collection. For example, under Surface Survey Standard, all complete bovid femora must be collected while incomplete bovid femora could be optionally collected. If necessary material collected

TABLE 2. Spatial Data Collection Software for Mobile Devices

Software	Location	Description	License
geoODK	http://geoodk.com/index.php	Mobile spatial-data collection platform.	Free Open Source
iGIS	http://www.geometryit.com/	Mobile GIS application.	Free/Paid Closed Source
GIS cloud	http://www.giscloud.com/	Cloud-based data collection tools with offline capabilities.	Free/Paid Closed Source
Fulcrum	http://www.fulcrumapp.com/	Cloud-based data collection tools with offline capabilities	Paid Closed Source
GIS Pro	http://garafa.com/wordpress/all-apps/gis-pro	Fully featured GIS app for mobile devices. Includes built-in form- building tools for spatial data collection.	Paid Closed Source
Rhino Spect	https://rhinospect.com/	A cloud-based data collection app with offline functionality, emphasizing simplicity and ease of use in the field.	Paid Closed Source

above and beyond the protocol can be excluded to obtain comparability between collections during analysis. Adopting standard collection practices is necessary for building data sets suitable for synthetic analyses.

Similarly, in archeology there are no standards or best-practice guidelines for deciding which artifacts are piece-provenienced in an excavation and which are not. Most excavations will have different rules for different kinds of objects (for example, bones versus stones) and different rules for varieties of these objects, such as teeth versus bone splinters or tools versus unretouched flakes. One of the most basic rules, the size cut-off for piece proveniencing, varies widely between excavations and is seldom reported even in publications directly dealing with the amount of material recovered from the site. Unfortunately, some excavations do not even apply a size cut-off, instead reporting that all artifacts are piece-provenienced, which, in fact, means that the size cut-off varies by excavator within the site.

Transfer Workflows

A major challenge for digital data collection is developing efficient workflows for transferring data from the mobile device to the database. Two widely used data exchange formats are eXtensible Markup Language (XML) and JavaScript Object Notation (JSON). In the same way that spatial databases are extensions of relational databases, so are Keyhole Markup Language (KML) and geoJSON spatially extended versions, respectively, of XML and JSON. A third proprietary spatial data format, Shapefile, also is used widely for storing and exchanging spatial data.

The data transfer workflow proceeds by generating an export file on the mobile device and moving that file to the computer with the database where it can be processed and the records imported. Many GIS systems offer data migration workflows, but they can be complicated and unreliable under field conditions. Because there are no convenient off-the-shelf open-source solutions, we created a customized data migration workflow built into the PaleoCore website that allows projects to upload and download data files in KML format.

Automation is a key factor for a good data-migration workflow. Data transfers should require as little direct user engagement as possible, so that the process is quick and reliable. Automated import and export tools make this possible and reduce manual moving, copying, and renaming of files, hence avoiding accidental data loss.

IMPLEMENTING A DIGITAL DATA COLLECTION SYSTEM

The authors have been experimenting for years with digital data collection techniques in a wide variety of contexts, including ecological and taphonomic data collection in Serengeti and Amboseli national parks, numerous archeology projects in Europe and Africa and, most recently, paleoanthropological field work in the Afar region of Ethiopia.

Spatial Database

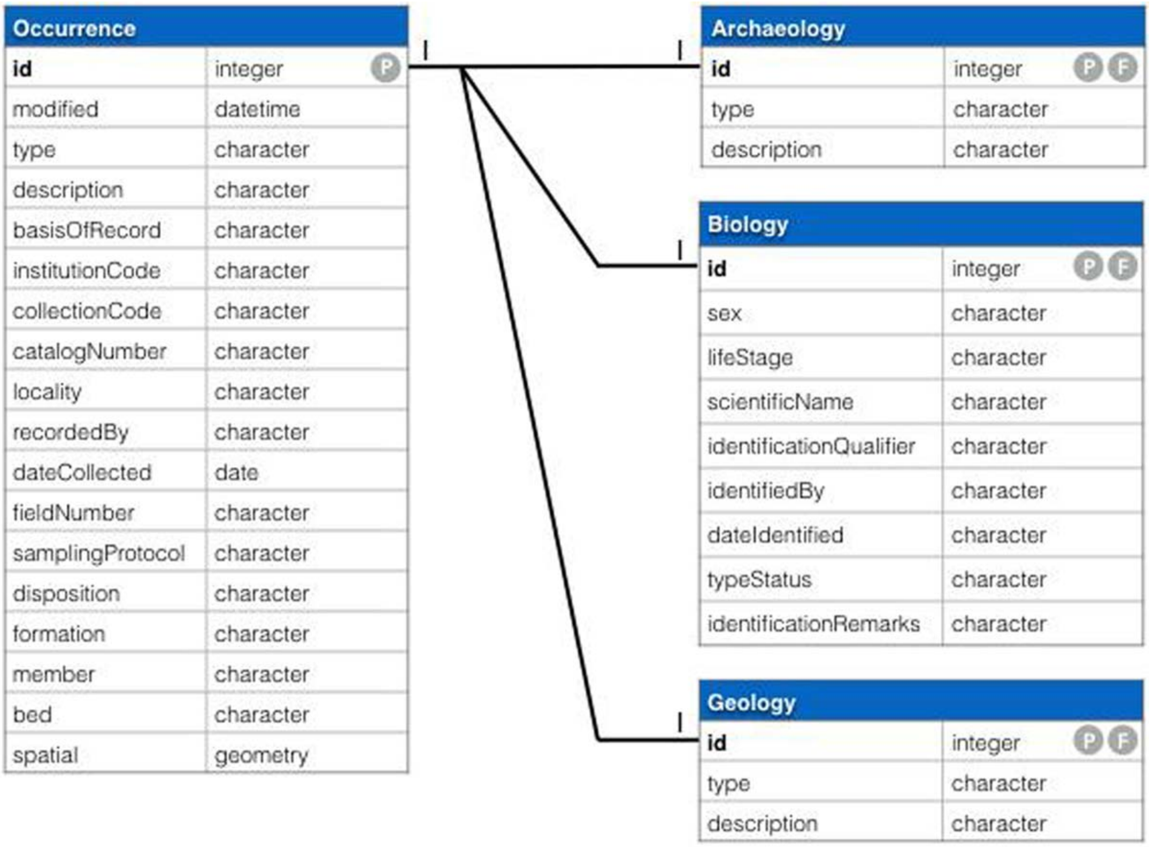
The current iteration of the system we have developed for field work deploys a PostgreSQL spatial database. PostgreSQL is a feature-rich, free, open-source database system that is easy to install on all platforms (http://postgresapp.com). PostgreSQL is powerful but can be complicated to manage. An easier alternative is SQLite, which has the advantage of being a native data format for QGIS. It is possible to migrate data between various open-source database systems using tools such as GeoKettle (http://www.spatialytics.org/projects/geokettle/). This is harder to accomplish when data are in closed-source database systems. We use the PaleoCore website as the main data repository and download data to a local instance of the database on a laptop for field campaigns. At the end of the campaign, we upload the data back to the repository.

The basic database structure that we use is diagrammed in Figure 6. The Occurrence table hosts basic information common to all items such as the collector, time and date of collection, spatial location, and a digital image. Items in the Occurrence table are divided into three basic subclasses, as recorded in the “type” column: archeological, biological, and geological. Three additional tables

TABLE 3. Data Collection Methods and Protocols

Name	Method	Collection Protocol
Surface Collection Exploratory	On-foot surveys moving quickly to reconnoiter new areas	Collection is ad hoc, focusing on finds of taxa not already recovered or previously known from the study area. Very large taxa such as Hippopotamidae and Elephantidae are recorded but not collected.
Surface Collection Standard	On-foot surveys moving deliberately at a pace to maximize the location and the recovery of all fossils visible from standing height.	At a minimum the following items are collected: All primate and carnivore remains regardless of preservation. All mammalian cranial remains identifiable to Family level except very large taxa such as Hippopotamidae and Elephantidae. These are recorded but not collected. All horn core and ossicone fragments. All mammalian isolated teeth that are at least half preserved. All complete crocodilian teeth, collected in bulk if there are many. Nonmammalian crania that are relatively complete. All astragali at least half complete. All complete or nearly complete calcanei, scapulae, humeri, radii, ulnae, femura, tibiae, fibulae, and metapodials -- except for very large taxa. Lizard and Snake vertebrae Significant uncollected finds are recorded and photographed.
Surface Collection Transect	Same method as Surface Collection Standard, but the path taken, distance, and duration are recorded. Transects are used to standardize collection effort.	Collection is the same as Surface Collection Standard.
Surface Collection Intensive	Same method as Surface Collection Standard.	Everything in Surface Collection Standard All fragments with at least one articular surface preserved.
Surface Crawl	Deliberate crawling or crouching to recover all items visible from kneeling height in areas of elevated significance and known fossil/artifact occurrence	All items are collected.
Dry Screening	Sediment is scraped from the surface to a depth 5–10 cm and shaken through sieves with 5 mm to 2 mm mesh. The area scraped should be documented.	All items are collected from the sieves.
Wet Screening	Sediment is dry-screened as described. The material that passes through the dry screens is retained and passed through finer < 2- mm mesh sieves with water. The concentrate retained in the screens is dried and sorted under magnification.	All items are collected from the sieves.
Excavation	Stationary *in-situ* recovery of material with careful digging, dry screening, and wet screening.	All items greater than the screen size are collected, although analysis may later be limited to items greater than 2.5 cm.

store information unique to each subclass; these have a one-to-one relationship with rows in the Occurrence table. This arrangement serves as a basic framework that is partially normalized, but robust enough under field conditions because all the critical information is stored in the Occurrence table. Ancillary information, such as taxonomic identifications, stored in the linked tables, can be recovered by referring to original fossils in the event that the ancillary tables are lost or the links to the Occurrence table are corrupted.



Occurrence		
id	integer	P
modified	datetime	
type	character	
description	character	
basisOfRecord	character	
institutionCode	character	
collectionCode	character	
catalogNumber	character	
locality	character	
recordedBy	character	
dateCollected	date	
fieldNumber	character	
samplingProtocol	character	
disposition	character	
formation	character	
member	character	
bed	character	
spatial	geometry	

Archaeology		
id	integer	P F
type	character	
description	character	

Biology		
id	integer	P F
sex	character	
lifeStage	character	
scientificName	character	
identificationQualifier	character	
identifiedBy	character	
dateIdentified	character	
typeStatus	character	
identificationRemarks	character	

Geology		
id	integer	P F
type	character	
description	character	

Figure 6. An entity-relationship diagram (ERD) illustrating the basic structure of the Occurrence table and related tables. Symbols and abbreviations are as in Figure 3. The column names appearing in each table are standard terms; their definitions are available online (http://paleocore.org/standard/). Additional nonstandard terms could be added as necessary. (Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.)

The column names shown in Figure 6 are standard terms drawn from Darwin Core and Dublin Core. Definitions and details for each are available on the PaleoCore website (htttp://paleocore.org/standard/). While it is useful to employ standard terms, doing so is not a requirement. It is most important that the database design fits the needs of the project, using whatever column names and concepts necessary. It is important to have in mind how the project's information maps to existing standards and to document the mapping. For example, the Dikika Research Project database uses some nonstandard names for columns that map directly to standard terms, as well as some columns that do not map well at all, such as "stratigraphic marker" and "distance from marker," which track the stratigraphic position of a fossil where it was found. These terms are not part of any standard, but they are vital to how data are collected at Dikika. Likewise, we use the "description" term (from Dublin Core) to record the anatomical element for fossils, but it would also make sense to create a dedicated column in the Biology table for this purpose. This illustrates the important point that the use of standards facilitates rather than dictates database design. The PaleoCore website includes a meta-database of project terms and how they map to standard terms and to each other. This mapping is the most important facet for future data synthesis.

Mobile Data Collection

In the field, we recently used iOS smartphones and tablets with a closed-source GIS app (GIS Pro by Garafa Software) because we preferred an iOS interface for familiarity and ease-of-use and needed a mobile app that can edit data offline and create new datasets on the device. When we designed the system, there were no open-source mobile solutions capable of handling spatial data. Now there is a fast-growing set of mobile data collection apps, from offline fully featured GIS tools to cloud-based solutions (Table 2). We have posted specific hardware recommendations online (http://paleocore.org/help/).

Workflows

During collection bouts, as workers find fossils and artifacts they mark them with survey flags; the data collector visits each, logs it on the mobile device and takes a photograph. Generally, logging an item takes about 30 seconds. If the item meets the criteria for collection, as listed in Table 3, then in addition to logging and photographing it, the collector completes a handwritten collection card that duplicates the data entered on the mobile device and bags the item with the collection card in a sample bag. If the density of items precludes individual piece-proveniencing, items that are close together are collected in bulk and separated later in the database. Duplicating data on the collection card provides an important fail-safe against data loss if the mobile devices were to fail. It also provides a written record that accompanies and identifies each fossil until a permanent catalog number can be written directly onto it, usually later that same day at the field lab. Photographing fossils at the time of collection is a safeguard for matching digital records to the actual fossil in the event that a collection card is lost; this is facilitated by including the written collection card (which includes a printed centimeter scale) in the photograph.

At the end of a collecting session (for example, at the end of a day in the field) the data on the mobile devices is copied to a laptop in KMZ format, which includes the data (in a KML file) plus all the images together in a zipped archive. We keep the individual data files from each device for each day as a backup. We upload the individual data files to the project's database on the PaleoCore data repository, using the KMZ import utility available on the website. PaleoCore is an open-source project and the source code for the entire site, including the Python code for importing and exporting data, is available on the GitHub source code repository (https://github.com/paleocore/paleocore).

Once data are migrated to the catalog, we review each of the collected fossils, update taxonomic identifications, and write catalog numbers directly onto the fossil with permanent ink. At the end of each day, all the data are digitized and ready for further analysis; at the end of the season, we can immediately generate an inventory of all collected items.

CONCLUSIONS

Our earlier field experiences were marred by the frustration of having to manually record fossil locations on GPS receivers with awkward user interfaces, then later hand-enter data into spreadsheets and GIS databases. The process of aggregating data from multiple GPS devices was cumbersome and prone to error. During analyses, it was difficult to track changes and versions of the datasets between members of the research team.

We experimented for years with closed-source solutions, but these too were expensive and complicated, as well as prone to errors. The system we have outlined here evolved as the best solution to the problem because it allows us to evaluate fossil densities and to plan field logistics better in real time. This system also facilitates the creation of more uniform paleontological localities *a posteriori* in areas with a tradition of localities (for example, Dikika[56–59]) and allows us to simplify and collect fossils without reference to localities as we do at our newest research area in Mille-Logya, Ethiopia. Using this system, we have standardized data structures and data-collection protocols with neighboring projects (such as Hadar), so that we can begin synthesizing information. Similarly, developing standardized data collection protocols was critical for conducting systematic taphonomic analyses, as we did recently in an investigation of Dikika cut-marked bone.[56] Developing a system with FOSS components helps reduce cost and makes it possible to customize features specifically for paleoanthropology. Because the tools do not require expensive licenses and are broadly available to the international scientific community, this system also fosters international collaboration.

Another advantage of digital data collection is that in designing the systems we must contemplate exactly what data are needed and how data should be collected. Our system compels us to articulate exactly which pieces of information are worth the time and effort of collecting in the field and explicitly to model how that information will be coded and documented. Along the way, it forces us to think about the things we collect: What is a fossil occurrence? When can we safely assume that two fossils cannot represent the same individual? What is an artifact? What constitutes an archeological site? What fossils should be collected and which ones should be left in the field? The answers to these questions reflect our research priorities and our conception of these issues. Digital data collection forces us to articulate our conception of the data. When combined with data standards it allows us to discover and combine information from multiple sites and multiple lines of evidence. By going a step further and adopting shared data collection methods we can begin conducting synthetic analyses, which pave the way to addressing complex and long-standing questions in paleoanthropology.

ACKNOWLEDGEMENTS

We thank the Authority for Research and Conservation of Cultural Heritage (ARCCH) and the Afar Regional State in Ethiopia for permission to conduct fieldwork. Funding for fieldwork and equipment was partially provided by Margaret and William Hearst. Support for PaleoCore was provided by the US National Science Foundation under award number 1244735, the Texas Advanced Computing Center in Austin and JetBrains Software. We also wish to thank the five anonymous reviewers for their constructive critique of our original submission.

REFERENCES

1 Atkins, D. E., Droegemeier, K. K., Feldman, S. I., et al. 2003 Revolutionizing science and engineering through cyberinfrastructure. Report of the National Science Foundation blue-ribbon advisory panel on cyberinfrastructure. Arlington, VA: US National Science Foundation..

2 Marean CW, Bar-Matthews M, Bernatchez J, et al. 2007. Early human use of marine resources and pigment in South Africa during the Middle Pleistocene. Nature 449:905–908.

3 Haile-Selassie Y, Gibert L, Melillo SM, et al. 2015. New species from Ethiopia further expands Middle Pliocene hominin diversity. Nature 521:483–488.

4 Villmoare B, Kimbel WH, Seyoum C, et al. 2015. Early *Homo* at 2.8 Ma from Ledi-Geraru, Afar, Ethiopia. Science 347:1352–1355.

5 deMenocal P, Bloemendal J. 1995. Plio-Pleistocene climatic variability in subtropical Africa and the paleoenvironment of hominid evolution: a combined data-model approach. In: Vrba E, Denton G, Partridge T, et al., editors. Paleoclimate and evolution with emphasis on human origins. New Haven:Yale University Press. p 262–288.

6 Potts R. 1998. Variability selection in hominid evolution. Evol Anthropol 7:81–96.

7 Behrensmeyer AK, Todd NE, Potts R, et al. 1997. Late Pliocene faunal turnover in the

Turkana Basin, Kenya and Ethiopia. Science 278:1589–1594.

8 Cerling TE, Wynn JG, Andanje SA, et al. 2011. Woody cover and hominin environments in the past 6 million years. Nature 476:51–56.

9 Kingston J. 2007. Shifting adaptive landscapes: progress and challenges in reconstructing early hominid environments. Yearbk Phys Anthropol 50:20–58.

10 Delson E, Harcourt-Smith WEH, Frost SR, et al. 2007. Databases, data access, and data sharing in paleoanthropology: first steps. Evol Anthropol 16:161–163.

11 Reed DN, McPherron S, Barr WA, et al. 2011. GPS data collection methods for paleoanthropology: examples from the Dikika Research Project geodatabase. TDWG 2011. TDWG.

12 Collins R. 1994. Why the social sciences won't become high-consensus: rapid-discovery science. Sociol Forum 9:155–177.

13 Stein LD. 2008. Towards a cyberinfrastructure for the biological sciences: progress, visions and challenges. Nat Rev Genet 9:678–688.

14 Stein LD. 2003. Integrating biological databases. Nat Rev Genet 4:337–345.

15 Goff SA, Vaughn M, McKay S, et al. 2011. The iPlant collaborative: cyberinfrastructure for plant biology. Front Plant Sci 2:1–16.

16 Sanz N. 2012. Heads 2: Human Origin Sites and the World Heritage Convention in Africa. World Heritage Center Papers Number 33. Paris: UNESCO.

17 Wallis JC, Borgman CL, Mayernik MS, et al. 2007. Know thy sensor: trust, data quality, and data integrity in scientific digital libraries. In: Kovács L, Fuhr N, Meghini C, editors. Research and advanced technology for digital libraries. Springer: Berlin Heidelberg. p 380–391.

18 Wallis JC, Rolando E, Borgman CL. 2013. If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. PLoS One 8:1–17.

19 Snow DR, Gahegan M, Giles CL, et al. 2006. Cybertools and archaeology. Science 311:958–959.

20 Flannery KV. 1982. The golden Marshalltown: a parable for the archeology of the 1980s. Am Anthropol 84:265–278.

21 Chew A, Oheim K. 2009. Using GIS to determine the effects of two common taphonomic biases on vertebrate fossil assemblages. Palaios 24:367–376.

22 McPherron SP, Dibble HA, Olszewski D. 2008. GPS surveying and on-site stone tool analysis: equipping teams for landscape analysis in the Egyptian high desert. In: Posluschny P, Lambers K, Herzog I, editors. Layers of perception: Proceedings of the 35th International Conference on Computer Applications and Quantitative Methods in Archaeology (CAA), Berlin, Germany, April 2–6, 2007. Dr. Rudolf Habelt GmbH: Bonn p 1–6.

23 Olszewski DI, Dibble HL, Schurmans UA, et al. 2010. Middle Paleolithic settlement systems: theoretical and modeling frameworks using high desert survey data from Abydos, Egypt. Settlement dynamics of the Middle Paleolithic and Middle Stone Age, Vol III. Tübingen: Kerns Verlag. p 81–101.

24 Binford L. 1987. Searching for camps and missing the evidence? Another look at the lower Paleolithic. The Pleistocene Old World: regional perspectives. New York: Plenum Press. p 17–31.

25 Mcpherron SJP, Dibble HL, Goldberg P. 2005. Z. Geoarchaeology 20:343–362.

26 Dibble H, McPherron S. 1988. On the computerization of archaeological projects. J Field Archaeol 15:431–440.

27 McPherron SP, Dibble HL. 1987. Hardware and software complexity in computerizing archaeological projects. Adv Computer Archaeol 4:25–40.

28 Grolinger K, Higashino WA, Tiwari A, et al. 2013. Data management in cloud environments: NoSQL and NewSQL data stores. J Cloud Comput Adv Syst Appl 2:22.

29 Pokorny J. 2013. NoSQL databases: a step to database scalability in web environment. Int J Web Information Syst 9:69–82.

30 Güting RH. 1994. An introduction to spatial database systems. The VLDB Journal 3:357–399.

31 Arctur D, Zeiler M. 2004. Designing geodatabases: case studies in GIS data modeling. Redlands, CA: ESRI Press.

32 Frehner M, Brändli M. 2006. Virtual database: spatial analysis in a web-based data management system for distributed ecological data. Environ Model Software 21:1544–1554.

33 Obe R, Hsu L. 2011. PostGIS in action. Greenwich: Manning Publications.

34 Thomas J, Potts R, Cole D. 1996. The role of GIS in the interdisciplinary investigation at Olorgesailie, Kenya, a Pleistocene archaeological locality. In: Aldenderfer M, Maschner H, editors. Anthropology, space and geographical information systems. New York: Oxford University Press. p 202–213.

35 Nigro JD, Ungar PS, de Ruiter DJ, et al. 2003. Developing a geographic information system (GIS) for mapping and analysing fossil deposits at Swartkrans, Gauteng Province, South Africa. J Archaeol Sci 30:317–324.

36 Conroy GC. 2006. Creating, displaying, and querying interactive paleoanthropological maps using GIS: an example from the Uinta Basin, Utah. Evol Anthropol 15:217–223.

37 Anemone RL, Conroy GC, Emerson CW. 2011. GIS and paleoanthropology: incorporating new approaches from the geospatial sciences in the analysis of primate and human evolution. Am J Phys Anthropol 146:19–46.

38 Conroy G, Anemone R, Regenmorter JV. 2008. Google earth, GIS, and the great divide: a new and simple method for sharing paleontological data. J Hum Evol 55:751–755.

39 Ihaka R, Gentleman R. 1996. R: a language for data analysis and graphics. J Comput Graph Stat 5:299–314.

40 R Core Team. 2014. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

41 Damuth JD. 1997. ETE database manual. Evolution of Terrestrial Ecosystems Consortium, Washington D.C.:Dept. of Paleobiology, Smithsonian Institution.

42 Fortelius M. 2008. Neogene of the old world database of fossil mammals (NOW). University of Helsinki, http://www helsinki fi/science/now.

43 Gilbert H, Carlson J. 2011. Data models and global data integration in paleoanthropology: a plea for specimen-based data collection and management. In: Macchiarelli R, Weniger GC, editors. Pleistocene databases: acquisition, storing, sharing. Mettmann: Neanderthal Museum. p 111–121.

44 Graham RW, Lundelius EL Jr. 2010. Faunmap. http://www.ucmp.berkeley.edu/faunmap/about/datastructure.html.

45 Graham RW, Lundelius EL. 1994. FAUNMAP: A database documenting late quaternary distributions of mammal species in the United States. Springfield: Illinois State Museum Papers.

46 David Schloen SS. 2012. OCHRE: an online cultural and historical research environment. Winona Lake, IN: Eisenbrauns.

47 Huggett J. 2012. Lost in information? Ways of knowing and modes of representation in e-archaeology. World Archaeol 44:538–552.

48 Robertson T, Döring M, Guralnick R, et al. 2014. The GBIF integrated publishing toolkit: facilitating the efficient publishing of biodiversity data on the internet. PLoS One 9: e102623.

49 Wieczorek J, Bloom D, Guralnick R, et al. 2012. Darwin Core: an evolving community-developed biodiversity data standard. PLoS One 7:1–8.

50 Borries C, Gordon AD, Koenig A. 2013. Beware of primate life history data: a plea for data standards and a repository. PLoS One 8:1–12.

51 Edwards PN, Mayernik MS, Batcheller AL, et al. 2011. Science friction: data, metadata, and collaboration. Soc Stud Sci 41:667–690.

52 Dibble HL. 1987. Measurement of artifact provenience with an electronic theodolite. J Field Archaeol 14:249–254.

53 McPherron S, Dibble H. 2002. Using computers in archaeology. New York: McGraw-Hill.

54 McPherron SJP. 2005. Artifact orientations and site formation processes from total station proveniences. J Archaeol Sci 32:1003–1014.

55 Alemseged Z, Bobe R, Geraads D. 2007. Comparability of fossil data and its significance for the interpretation of hominin environments: a case study in the lower Omo Valley, Ethiopia. Hominin environments in the East African Pliocene: an assessment of the faunal evidence. Dordrecht, The Netherlands: Springer. p 159–181.

56 Thompson JC, McPherron SP, Bobe R, et al. 2015. Taphonomy of fossils from the hominin-bearing deposits at Dikika. J Hum Evol. 86:112–135

57 Alemseged Z, Spoor F, Kimbel WH, et al. 2006. A juvenile early hominin skeleton from Dikika, Ethiopia. Nature 443:296–301.

58 Wynn JG, Roman DC, Alemseged Z, et al. 2008. Stratigraphy, depositional environments, and basin structure of the Hadar and Busidima Formations at Dikika, Ethiopia. In: Quade J, Wynn J, editors. The geology of early humans in the Horn of Africa. Boulder: Geological Society of America. p 87–118.

59 McPherron SP, Alemseged Z, Marean CW, et al. 2010. Evidence for stone-tool-assisted consumption of animal tissues before 3.39 million years ago at Dikika, Ethiopia. Nature 466: 857–860.