CHAPTER ELEVEN

# *PaleoCore*

## An Open-Source Platform for Geospatial Data Integration in Paleoanthropology

DENNÉ N. REED, W. ANDREW BARR, AND JOHN KAPPELMAN

### INTRODUCTION

*Who* we are is intimately connected to *where* we are and what is around us. Thus, it is hardly surprising that geographic location plays an important role in nearly every branch of anthropological research: sociocultural anthropologists map events and analyze geographic relationships between people and communities; primatologists track their study animals and map the resources they depend on; while archaeologists and paleoanthropologists maintain a nearly obsessive interest in the spatial provenience of artifacts and fossils. Twenty years ago, Mark Aldenderfer and Herbert Maschner (1996) reviewed the many ways that geographic information systems (GIS) were being adopted in anthropology. Since then, developments in GPS (or global positioning systems, as they are known more generally) and online tools and the growth of open-source-software communities have shaped new geospatial approaches in anthropology that are networked, collaborative, open, and less reliant on stand-alone desktop GIS software. The ability to (1) *collect* and combine large amounts of georeferenced information, (2) collaboratively *manage* and document that information, and (3) *share* it with others greatly expands opportunities to synthesize and reuse data, which in turn opens the way for researchers to address broader questions that are too big to tackle individually.

The other contributions to this volume document the diverse ways in which geospatial techniques are deployed in anthropology across a broad range of temporal and geographical contexts, from Andean archaeology to Eocene primatology. This chapter takes a slightly different approach. Rather than focusing on an analytical technique, geographic area, or temporal period, we look more generally at the challenges of geospatial data management, which we view as a strategic issue because addressing big-picture questions in anthropology requires synthesizing the collective efforts of multiple research teams. This trend crosses many branches of science, from astronomy to genetics to zoology, and thus a treatment of new geospatial approaches in anthropology should extend beyond the traditional discussions of geospatial analysis techniques to include the broader topics of geospatial data collection, management, and sharing, that is, spatial cyberinfrastructure more generally (Wright and Wang 2011).

In recent years, data collection across all branches of anthropology (and science generally) has expanded dramatically both in terms of the volume and variety of information collected, including the amount and breadth of geospatial information. This growth in part reflects the ubiquity of inexpensive and embedded GPS (global navigation satellite systems [GNSS]) receivers in mobile devices, cameras, and other instruments. The deluge of data has prompted new data-management requirements from the National Science Foundation (NSF) and other scientific funding agencies, as well as new academic norms about sharing data (Bell, Hey, and Szalay 2009; Tenopir et al. 2011). It has also encouraged new approaches to geospatial data management in anthropology, such as accommodating spatial data directly in relational databases to better handle the unique requirements of these data and to enable geospatial manipulations and queries outside of a GIS software package.

Likewise, data-management tools are improving to match the increased pace of collection and the increased complexity of modern research techniques. As our work becomes more sophisticated, anthropologists are working in teams rather than individually, which in turn creates demand for tools to support collaborative, team-based data management and analysis. Teams require the ability to edit and update their data online, download them for analyses, or connect directly through GIS software clients.

Collaboration within teams is connected to an expanding culture of sharing information between teams. Synthesis is critical for addressing big-picture questions in anthropology (Delson et al. 2007), but successfully synthesizing these data requires a much greater investment in standardizing data collection and management practices, lest we combine data that were collected in different and incompatible ways. Furthermore, online connectivity is changing how desktop GIS systems store data and connect to data sources and how people use GIS software generally, since a greater diversity of GIS and spatial analysis tools are available online and on mobile devices. As part of this evolution, GIS software is also becoming more modular and may offer separate tools for geospatial data collection, data storage and management, and data analysis. The free, open-source software for geospatial (FOSS4G) movement has played a key role in the evolution and modularization of GIS software by introducing standards for how GIS data are stored and exchanged. Free geospatial software also lowers the cost barrier to entry for GIS software, and it fosters communities in which researchers can contribute as users and developers.

This chapter outlines the history of FOSS4G and offers examples of its application to anthropology by focusing on how FOSS4G resources help anthropologists manage spatial data and share it more effectively. We introduce PaleoCore as an example of a FOSS4G online spatial data infrastructure (SDI) platform. We describe the evolution of FOSS4G and SDI, how these are implemented in the PaleoCore scientific database and website, and their role in anthropology. Currently, PaleoCore hosts data for paleontology and archaeology projects, reflecting the degree to which these domains of anthropology have embraced geospatial data analysis, but we also argue that this platform can be more broadly relevant. New approaches to geospatial analysis in anthropology should consider if and how innovations can meet a broader range of needs, and thus we conclude the chapter with ideas for deploying these systems in other anthropological subfields.

## THE EVOLUTION OF GEOSPATIAL SOFTWARE

Geographical or geospatial information systems have a history going back at least to the 1960s (Coppock and Rhind 1991; Foresman 1998). Desktop GIS software followed the advent of personal computing in the early 1980s (e.g., the 1981 release of Arc/Info by Environmental Systems Research Institute [ESRI]), and it continued to evolve into more sophisticated systems with graphical user interfaces (GUIs) in the early 1990s (e.g., ESRI's ArcView software suite). At the same time, GIS became widely adopted as a data management and analysis platform in anthropology.

The use of desktop GIS software is now commonplace in anthropology, especially in archaeology and paleoanthropology, where spatial provenience of artifacts and fossils is integral to analysis. The GIS systems of the 1980s and 1990s used specialized and proprietary data-storage formats, such as ESRI's "coverage" and "shapefile" formats. The latter became a standard that is still widely used today. By the mid-2000s spatial data were integrated into relational database management systems, giving rise to ESRI's ArcGIS personal geodatabases and the Postgres/PostGIS spatial database. This change marked a shift away from desktop GIS that focused on digital cartography and toward GIS as multicomponent spatial database systems in which the desktop GIS was a client software application (one of many possible clients) used to visualize and analyze data stored separately in a spatial database.

Spatial databases play an important role in the evolution of GIS, and it is worth summarizing the technology here. A spatial database builds on the relational database management systems (RDBMS) that became widespread beginning in the 1970s. These systems store data in sets of related tables linked together by shared columns. Spatial database systems are a refinement on traditional RDBMS that include the ability to store rich spatial objects such as points, lines, and polygons, along with all the other data, inside an RDBMS data table (Obe and Hsu 2011). The spatial data simply occupy a column beside all the other data, and this column contains bundled data objects that store all the information needed to represent spatial features, including all the coordinates, as well as projection and coordinate system information (Reed et al. 2015). Furthermore, the standards developed by the Open Geospatial Consortium (OGC) help ensure that spatial data are interoperable across a variety of server and client software systems (Dunfey et al. 2006).

The advent of spatial databases based on client-server architecture led to the development of networked systems for geospatial data management, called spatial data infrastructure. Stefan Steiniger and Andrew Hunter (2012) define an SDI as a system comprising data, technologies, policies, people, and standards that enable the discovery and use of geospatial data by many users and also the reuse of geospatial data for new purposes. In a later review, Steiniger (2013) identified nine categories of FOSS4G software (table 11.1).

SDIs are typically developed by large organizations to serve many people and groups simultaneously. In this way, an SDI is well suited for integrating information between multiple teams. For example, in an academic context (as opposed to a government data-management context) this effort can include many different research teams that wish to aggregate their data for the purpose of understanding biogeographic and temporal distributions of fossils and artifacts collected across separate sites and for conducting synthetic analyses. Whereas desktop GIS targeted single users, spatial databases precipitated the rise of collaborative SDIs that decomposed the software system into many constituent modules linked by Internet connections and that can

Table II.I. Categories of GIS software with FOSS4G examples for each category

| Categories | Examples |
|---|---|
| desktop GIS | QGIS, GRASS |
| spatial database | PostgreSQL, MySQL, SpatiaLite |
| Internet mapping | GeoServer, MapServer, Leaflet |
| server GIS and web-processing server (WPS) | PyWPS, 52°North, WPS |
| mobile | GeoODK, QGIS for Android |
| libraries | GDAL, GeoKettle |
| extensions, plugins, and APIs | OpenLayers plugin for QGIS |
| remote sensing | InterImage, OSSIM |
| exploratory | R, open GeoDA |

accommodate many users simultaneously. The PaleoCore SDI is an example of such a system.

## PALEOCORE: A FOSS4G SDI FOR ANTHROPOLOGY

PaleoCore (http://paleocore.org) is an NSF-funded cyberinformatics initiative hosted at the Texas Advanced Computing Center (TACC) in Austin. It is dedicated to maintaining a spatial data repository for anthropological data, developing tools for digital data collection in the field, and developing (meta) data standards for the discovery and exchange of information. PaleoCore's overarching goal is to provide the research community with the infrastructure it needs to digitally collect, manage, and preserve invaluable data and to allow easy discovery and reuse of those data by researchers, educators, students, and the public. A longer-term goal is to integrate data from multiple research teams for the purpose of addressing broad-scale questions about human origins and evolution, such as the origin of the genus *Homo* or the spread of modern humans out of Africa.

## PALEOCORE SOFTWARE ARCHITECTURE

The PaleoCore spatial data infrastructure is built on free open-source software for geospatial (FOSS4G) components. A fundamental principle embodied in the system is that data are stored and managed independently from the client software used for analyzing the data. The data can be accessed from multiple

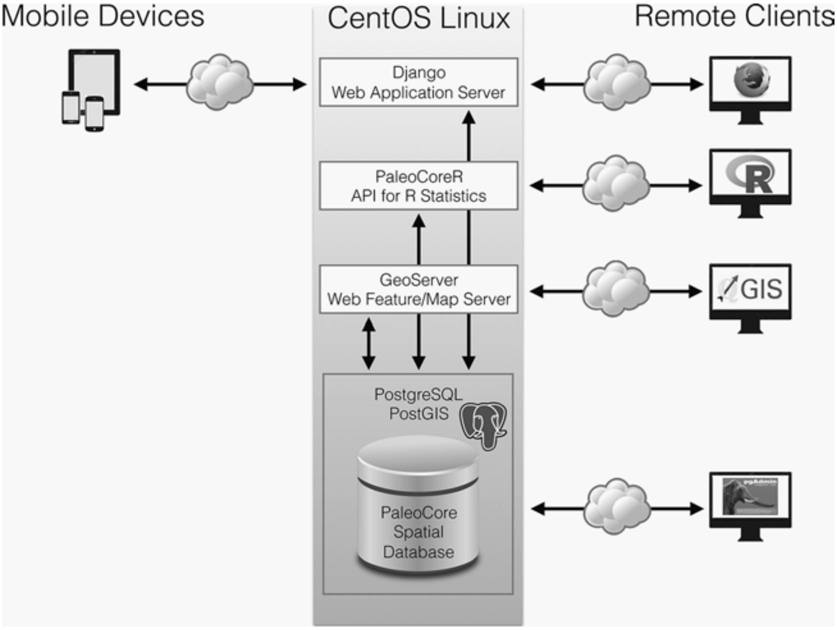

Figure 11.1. The PaleoCore open-source software stack. Mobile devices are used for field data collection. The field data are uploaded via the Internet to the PaleoCore data repository hosted at the TACC. The data-repository services (white boxes) allow connections from several remote-client software programs.

client programs depending on the needs of the analyst (fig. 11.1). PaleoCore uses PostgreSQL (http://www.postgresql.org/) as its relational database management system and the PostGIS extension to PostgreSQL to enable the storage of spatial data within the database. PostgreSQL and PostGIS together create a powerful spatial data storage facility, but it requires a user interface for working with the data. To this end, we employ the administrative interface that comes with the Django web application framework (http://django.org).

Django is a framework for building dynamic web applications written in the Python programming language (http://python.org). Django can connect to a variety of database back ends including PostgreSQL and SQLite (https://www.sqlite.org/), and it includes an administrator interface for viewing, querying, and managing data stored in the database. The administrator interface is customizable and allows the rapid development of online data-management interfaces. With the addition of a few extra software libraries (bundled as geoDjango, http://geodjango.org), the Django administration interface accommodates the presentation and editing of spatial data, which are presented through an interactive map widget.

Within the PaleoCore repository, each hosted research project is maintained as an independent Django software module, and each project has its own database tables, associated users, and user permissions. Each project also has its own data-management interface that can be tailored to the needs of the project. Furthermore, any number of interfaces can be built to manage the data, with separate interfaces designed for specific roles and activities. For example, one interface can be designed for basic specimen management, another for taxonomic identification, and another for collecting measurements.

When analyzing data, the PaleoCore FOSS4G software stack allows connections for multiple client applications. One can connect with desktop GIS clients such as QGIS or ArcGIS and immediately manipulate data natively without having to export or import anything. PaleoCore includes an application programming interface (API) for the statistical programming language R (PaleoCoreR), which allows R clients to download spatial data into an R dataframe with just a few lines of code. R itself has several libraries for visualizing and analyzing spatial data such as "sp," "rgdal," and "rgeos." The web interface in Django facilitates collaborative data management and editing, which are vital for distributed research teams. Finally, more complicated database manipulations can be implemented in SQL (a universal database programming language) with the PostgreSQL database client software pgAdmin (http://www.pgadmin.org/) or any of a number of generic database clients, including DB Visualizer (https://www.dbvis.com/).

## MOBILE DATA COLLECTION

Digital data-collection workflows for mobile devices form an integral part of the PaleoCore system; this system include routines for collecting digital data in the field and migrating those data into the spatial database. Denné Reed and colleagues (2015) provide a detailed account of a data-collection system for paleoanthropology that allows researchers to capture specimen data on mobile devices such as smartphones and tablets, as the items are collected in the field, including spatial location, time of collection, stratigraphic position, and other vital details. In this system the data are born digital, they do not have to be manually entered later, and an updated digital catalog is readily available for analysis.

This system employs mobile GIS apps for data collection and uses Keyhole Markup Language (KML) to transfer data from the mobile devices into the spatial database. When the system was developed, there were no mature FOSS4G mobile solutions available, and a proprietary solution was adopted (GIS Pro, http://garafa.com/wordpress/all-apps/gis-pro). Several new mobile initiatives, including QField (http://www.opengis.ch/android-gis/qfield/) and geoODK (http://geoodk.com/), are now filling the mobile-collection niche in the FOSS4G ecosystem.

Mobile data-collection tools provide an important incentive for researchers to contribute data into the PaleoCore system. These applications facilitate data collection and make it easy to move that data directly into the PaleoCore repository. When combined with the benefits of collaborative online data management through the PaleoCore web interface, this integrated approach supports researchers and encourages their participation in the system.

## PALEOCORE (META)DATA STANDARDS

A widespread challenge when integrating data from multiple sources is mapping the content of one data set to another. This process can be conducted at many levels of refinement depending upon how the combined data set will be used. Imperfect mappings are suitable for data search and discovery, such as knowing generally what data are available where. Biodiversity data standards such as Darwin Core (Wieczorek et al. 2012) and the Access to Biological Collections Data (ABCD; Holetschek et al. 2012) are well suited to this task and are used, for example, in the Global Biodiversity Information Facility (GBIF, http://gbif.org) to integrate biodiversity data across a federation of data providers. PaleoCore has implemented a composite set of standard terms (http://paleocore.org/standard/) compiled from Darwin Core (http://rs.tdwg.org/dwc/terms/) and the library metadata standard Dublin Core (http://dublincore.org/) to begin reconciling paleoanthropological data sets stored in its repository. This reconciliation provides the basis for the future development of global searches and queries of items in PaleoCore originating from many different contributing projects.

However, aggregating data in this general way is unsuitable for synthetic analyses — a widely acknowledged problem in information science (Zimmerman 2008; Veen et al. 2012; Wallis, Rolando, and Borgman 2013). Efforts to integrate data for joint analyses require higher levels of collaboration between different data providers, for example, in developing and implementing standardized procedures and best practices. This higher-level metadata and provenance information must also accompany the data themselves. Bechhofer and colleagues (2013) argue that standard metadata are necessary but not sufficient for conducting fully synthetic meta-analyses using data sets combined from diverse sources. Richer context is required, and these authors advance the idea of research objects, encapsulations of data, metadata, publications, provenance information, and other vital details needed to use the data appropriately. Such encapsulations would embody a more open, reproducible, and reusable scientific product.

## OPEN-SOURCE COLLABORATIVE SCIENCE

Reduced cost is one advantage of deploying FOSS4G technology, but for research perhaps a greater benefit comes from the communities that develop around FOSS software. The development and maintenance of FOSS software require a community with the attending infrastructure, such as collaborative code repositories like GitHub (http://github.com), along with the many blogs and wikis for discussions and assistance in using the software. Thus, one of the biggest advantages to using FOSS is access to the community of people using the same software and the rich exchange of ideas and information that stems from them

A second advantage, especially for a field like anthropology, is the way that FOSS software allows researchers to evaluate and modify the software's source code. This option makes it possible to customize the software for boutique applications that may be ignored by proprietary software developers. An example from PaleoCore has been the desire to export mobile data in different formats such as GeoJSON. Some proprietary mobile app developers have not implemented this feature in their software, and users must continue to request the feature from the developer, but with FOSS software this feature could be added by the user.

The prospect of adding or modifying source code may seem daunting to some anthropologists, but it is more feasible now that programming languages have become increasingly easier to use. For example, Python is now the most widely taught language in introductory computer-science courses (Guo 2014). Its friendly syntax and broad academic and nonacademic user base are draws. It is also the primary geoprocessing language adopted by ESRI and is the language used to develop many FOSS software applications, including Quantum GIS (QGIS, http://www.qgis.org/en/site/). Python, like many programming languages, can import libraries written by others, and the libraries bring a functionality that is easy to call upon. Thus, it is possible to write simple software that has powerful capabilities without engaging in much "low-level" programming. Similarly, editing existing software creates a point of entry for inexperienced programmers and can add the critical missing piece to an existing software package that makes it ideal for boutique applications.

The communities fostered by FOSS provide a resource and model for a more open and collaborative science. By employing the tools available for FOSS, we can encourage the rapid exchange of ideas between research groups, fostering a fast-paced, discovery-based science rather than a divided and contentious one.

## PALEOCORE IN PALEOANTHROPOLOGY

An online collaborative spatial data infrastructure has several potential roles in anthropological research: (1) assisting researchers in collaborative data management; (2) allowing researchers, educators, and the public to search and browse museum and research collections; and (3) providing a framework by which researchers may begin synthesizing data collected by different teams in order to address broader questions.

Paleoanthropology is the scientific study of paleontological and archaeological remains relevant to human origins and evolution. PaleoCore provides a fully digital workflow for the collection of fossils at several paleoanthropological sites in Africa. In these use cases, researchers log fossils and artifacts on mobile devices in real time—as they are discovered and collected or alternatively observed but left in place. The trick is to develop streamlined routines that allow piece-proveniencing of many items without interfering with or reducing the pace of fieldwork (Reed et al. 2015). Data collected on mobile devices are then transferred in the field to a spatial database, where they can be analyzed with all the capabilities of desktop GIS systems.

At the project level, mapping fossil occurrences helps researchers estimate fossil densities across a study area, establish collection priorities, and better plan field activities. Mapping individual fossils also helps reveal biogeographic patterning on the landscape and can be used to estimate the age of different collection areas based on biochronology. Element distributions can also reveal taphonomic processes such as size or preservation biases across different areas.

Paleoanthropological research teams comprise several full-time scientific specialists, all of whom must be able to access, edit, and manage a shared set of data. PaleoCore provides a central repository for research teams and allows anyone with the appropriate permissions to collaboratively view and edit data at the same time over the Internet. The PostgreSQL RDBMS is designed to serve multiple users simultaneously, and the Django user interface can be customized to suit the specific needs of each project.

The ability to manage media associated with artifacts is also key to any system for use in archaeology and paleontology. Digital images of fossils and artifacts are now a standard part of documentation and are readily incorporated, as are videos, three-dimensional surface scans, and other related data files. One advantage of an online system over conventional desktop GIS is that these resources — the data about the specimens and the associated images and media — can be readily shared through a variety of Internet mechanisms. Web pages allow users to search and download data, while other mechanisms, such as APIs, allow other computers to automatically locate and access data.

Beyond the scope of a single project, PaleoCore presents researchers with the opportunity to begin integrating data across projects. Global searches and queries across fossil-occurrence data sets allow researchers, educators, students, and the public to learn what fossils occur where and when. They also allow the paleoanthropology community to begin answering fundamental questions: How many fossils make up the human fossil record? What is the complete catalog of *Australopithecus afarensis* from Ethiopia? What is the geospatial extent of all Neanderthal fossils? Our inability to address these basic questions illustrates the limitations of our current geospatial digital infrastructure and the need to address the problem.

But where do we begin? Each project deploys its own system for data collection, and each project has its own field techniques and its own data structures. Some projects record spatial provenience for every find, while others record provenience using localities or collecting areas. Some use databases and some record information in spreadsheets, while others still rely on pen and paper field notes. PaleoCore begins by offering a flexible and affordable digital workflow.

When a project is added to the PaleoCore database, the terms in the project's data schema are mapped to the standard data schema (fig. 11.2). Some project terms do not map onto any standard terms and remain unique to the project, or they may be added to the standard. The mapping process gives projects the freedom to use whatever terms they prefer without being forced to conform

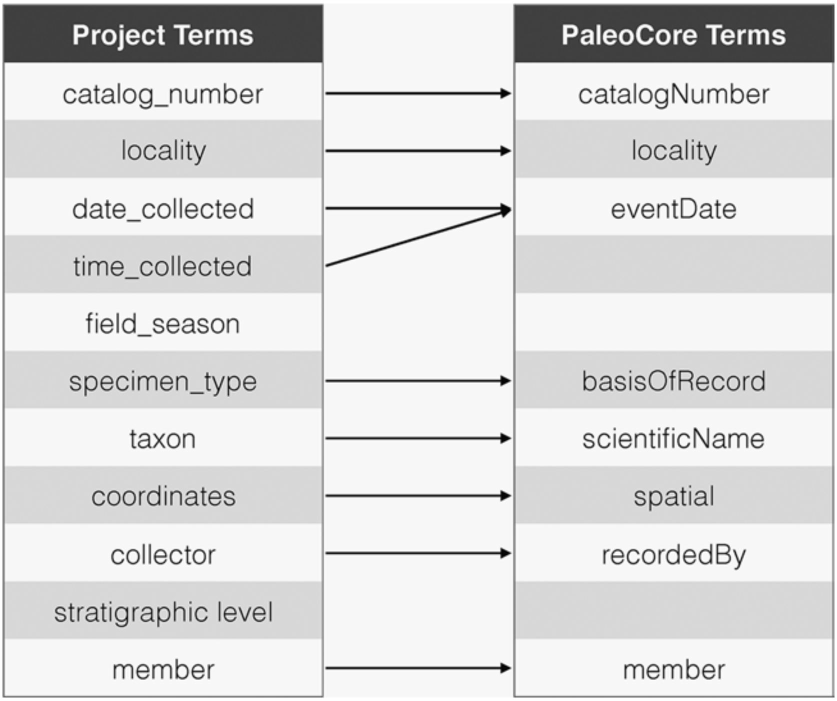

Figure 11.2. The mapping of project data terms with PaleoCore standard terms. The terms used in a project-data model are mapped to terms in the PaleoCore data standard, which is derived from the Darwin Core and Dublin Core data standards. Some terms may need to be combined, and some do not match any standard terms.

to a standard. At the same time, PaleoCore provides the mapping that allows translation of data from one project to the next. The standards are defined in a way that is flexible enough for sufficient ease of mapping. The downside is that the mapping across projects may sometimes be inexact. Generally, this level of mapping is sufficient for search and discovery but insufficient for data synthesis and analysis; these latter steps require a much greater degree of coordination between projects in order to establish best practices, common data collection protocols, and procedures.

## BROADER APPLICATIONS AND FUTURE DIRECTIONS

The benefits of online spatial-data infrastructure are not restricted to paleoanthropology. Any domain that needs to manage spatial data collaboratively would benefit from an online spatial-data infrastructure, and these resources are being developed. For example, the Ethoinformatics project (http://ethoinformatics.org) is a data-management and cyberinfrastructure initiative for primatology and behavioral data similar to PaleoCore but focusing more on data standards (for behavioral data) and data-collection platforms or field observations of behavior). The Digital Archaeological Record (tDAR) project provides data-archiving services for archaeological sites (https://www.tdar.org/about/). This system facilitates long-term data storage and archiving, but it does not seek to integrate the data into a shared data structure. The site provides searching and indexing of data sets but not the data contained in them, which can be stored in a wide variety of formats, including spreadsheet files, text documents, or project database files. It is quite distinct from PaleoCore and from Open Context (http://opencontext.org), which both focus on implementing data standards to present original archaeological data in a more structured format. A similar approach is used by the OCHRE Data Service (https://ochre.uchicago.edu/), which focuses on Near Eastern archaeology. OCHRE does not advocate data standards but does help researchers to collect and store data in a structured and consistent way.

These efforts reflect a new approach to spatial analysis, one focusing on the entire spectrum of the spatial data management lifecycle from design and documentation to data collection and storage, collaborative maintenance and management, analysis, publication and sharing, and reuse and synthesis. One of the key advances in geospatial analysis is our ability to wholistically manage large amounts of spatial data and to leverage the data more effectively.

## SUMMARY

Over the past twenty years spatial data analysis has become a standard part of paleoanthropology, evolving along the way to incorporate advances in GIS and remote-sensing software. The ability to maintain spatial data inside an RDBMS opens new frontiers for geospatial science, especially one immersed in the collaborative framework of free open-source software and networked software components that comprise spatial-data infrastructure. The technologies for implementing these systems have matured to the point where they offer excellent tools for collaborative research that allow scientific teams to begin integrating data in new ways so that they can more easily investigate the patterns and processes of human evolution and prehistory.

## ACKNOWLEDGMENTS

This work was funded by a grant from the National Science Foundation under award 1244735. PaleoCore is maintained by a core group of software developers including Tomislav Urban, W. Andrew Barr, Shannon McPherron, Andrew Cook, Nick Pereira, and Jeanette Tang. PaleoCore is supported in part by software donations from JetBrains Research Inc. and DbVisualizer software. The PaleoCore database and website are maintained at the Texas Advanced Computing Center, whose scientists have provided guidance and support throughout the project.