FORUM

THE PROMISE AND CHALLENGE OF ARCHAEOLOGICAL DATA INTEGRATION

Edited by Keith Kintigh

This forum reports the results of a National Science Foundation–funded workshop that focused on the integration and preservation of digital databases and other structured data derived from archaeological contexts. The workshop concluded that for archaeology to achieve its potential to advance long-term, scientific understandings of human history, there is a pressing need for an archaeological information infrastructure that will allow us to archive, access, integrate, and mine disparate data sets. This report provides an assessment of the informatics needs of archaeology, articulates an ambitious vision for a distributed disciplinary information infrastructure (cyberinfrastructure), discusses the challenges posed by its development, and outlines initial steps toward its realization. Finally, it argues that such a cyberinfrastructure has enormous potential to contribute to anthropology and science more generally. Concept-oriented archaeological data integration will enable the use of existing data to answer compelling new questions and permit syntheses of archaeological data that rely not on other investigators' conclusions but on analyses of meaningfully integrated new and legacy data sets.

Este foro reporta los resultados de un taller auspiciado por la Fundación Nacional para las Ciencias (National Science Foundation), el cual se enfocó en la integración y conservación de las bases de datos digitales y de otros datos estructurados derivados de los contextos arqueológicos. Este taller llegó a la conclusión de que para que la arqueología alcance su potencial de avanzar en el entendimiento científico de la historia humana a largo plazo, hay una apremiante necesidad de que exista una infraestructura de información arqueológica que nos permita alcanzar, acceder, integrar, y extraer bases de datos diferentes. Este informe proporciona una evaluación de las necesidades informáticas de la arqueología, articula una visión ambiciosa para establecer una infraestructura de información disciplinaria distribuida (ciberinfraestructura), discute los retos presentados, y esboza los pasos iniciales hacia su realización. Finalmente, argumenta que dicha ciberinfraestructura tiene un enorme potencial de contribuir a la antropología y más generalmente a la ciencia. La integración de los datos arqueológicos orientados a los conceptos permitirá el uso de los datos existentes para resolver nuevas preguntas obligadas y conformar síntesis de los datos arqueológicos que se basan no en las conclusiones de otros investigadores sino en los análisis de bases de datos nuevas y heredadas integradas significativamente.

Executive Summary

Disciplinary Needs

For archaeology to achieve its potential to provide long-term, scientific understandings of human history, there is a pressing need for an archaeological information infrastructure that will allow us to archive, access, integrate, and mine disparate data sets. New technologies in information integration will enable archaeologists to (1) work at scales not currently possible to answer pressing questions that cannot now be addressed because of a lack of effective access to existing data; (2) foster the development of a new paradigm of integrative and synthetic research; (3) scale and integrate archaeological data so that they can be used to address compelling questions in other disciplines; and (4) sustain the scientific utility of existing digital data that are critically endangered by media degradation, software obsolescence, and inadequate data documentation (metadata). To meet pressing research needs and to help stem the loss of existing information, it is essential that we embark now on the task of creating an infrastructure that will allow us to archive and make available integrated databases of archaeological data.

Keith W. Kintigh ■ School of Human Evolution and Social Change, Arizona State University, Tempe, AZ 85287-2402 (kintigh@asu.edu)

American Antiquity, 71(3), 2006, pp. 567–578

Vision

The needed information infrastructure has five key components: (1) a Web-based interface designed for effective scholarly access; (2) sophisticated search capabilities using concept-oriented queries to locate and access relevant data sources while retaining the confidentiality and rights of data set creators; (3) data-integration tools that use concept ontologies and digital metadata to integrate data from multiple sources, yielding an output database of appropriately scaled observations with consistent variables; (4) reporting capabilities that include both the basic display of the integrated databases and their direct output in forms susceptible to further analysis; and (5) software tools that minimize the effort and expertise necessary to fully incorporate new and legacy data sets into the system. This cyberinfrastructure would encourage the research use of existing data, satisfy the data requirements of integrative and synthetic research, facilitate the entry of data into the infrastructure, sustain endangered and irreplaceable data, and increase the public accessibility of scientific findings.

Implications

Realizing this vision will entail both the development of innovative software tools that permit the cross-project integration of data and a sustained effort to document existing and newly created data sets. More specifically, we propose (1) establishing a national center for archaeological data integration that, along with a network of distributed data nodes, can address the emerging needs of a discipline asking increasingly wide-ranging questions that are impossible or impractical to answer with existing modes of research; (2) working with computer scientists to adapt existing informatics technologies from other disciplines (e.g., ecology and geology) and to develop general solutions for unsolved data integration problems posed by archaeology, thus contributing to a shared cyberinfrastructure for science; and (3) fostering a new group of professionals at the boundary of archaeology and computer science specializing in archaeological informatics. The national center would set priorities for the development of the cyberinfrastructure, coordinate the development of metadata standards, assemble the technical expertise and develop the necessary software, provide software maintenance, coordinate the efforts of regional nodes in a network of data providers, and maintain legacy databases lacking an institutional home.

Initial Steps

To begin, a steering group should be established (1) to develop a strategic plan that sets the direction for the long-term effort; (2) to work with the National Science Foundation and other funding agencies to fund cyberinfrastructure development; (3) to actively explore the utility of existing models for cyberinfrastructure development and to participate in collaborative efforts to develop shared (across disciplines) cyberinfrastructural tools; (4) to suggest compelling test cases that can demonstrate the potential of the system to address both focused and large-scale issues in both disciplinary and cross-disciplinary research; and (5) to encourage and support efforts by professional societies to advance the development of a cyberinfrastructure for archaeology.

Report

This report is the product of a workshop funded by a National Science Foundation grant to Arizona State University, "Enabling the Study of Long-Term Human and Social Dynamics: A Cyberinfrastructure for Archaeology" (SES 0433959). The workshop, hosted by the National Center for Ecological Analysis and Synthesis, was held December 5–6, 2004, at the Upham Hotel in Santa Barbara, California. In the workshop, the participants assessed the informatics needs of archaeology and the potential of a cyberinfrastructure for archaeology to benefit the discipline and science more generally. The workshop then formulated a vision for a cyberinfrastructure for archaeology, considered the challenges posed by its development, and outlined initial steps toward the realization of that vision.

The workshop included 31 distinguished participants: (1) 21 archaeologists representing different professional constituencies (academia, museums, government, and the private sector), time periods (prehistoric and historic), national settings (the United States, Mexico, England, and the Netherlands), and subjects of substantive expertise; (2) one physical anthropologist; and (3) nine

Table 1. Workshop Participants

Participant	Affiliation
Buchanan, Bruce	University of Pittsburgh, Computer Science
Candan, K. Selçuk	Arizona State University, Computer Science
Cowgill, George	Arizona State University, Archaeology
Davulcu, Hasan	Arizona State University, Computer Science
Doelle, William	Desert Archaeology Inc./Foundation, Archaeology
Giles, C. Lee	Pennsylvania State University, Computer Science
Goldstein, Lynne	Michigan State University, Archaeology
Hegmon, Michelle	Arizona State University, Archaeology
Kamermans, Hans	University of Leiden, Archaeology
King, Julia	Archaeological Conservation Lab/St. Mary's College of Maryland (Society for Historical Archaeology, Past President)
Kintigh, Keith	Arizona State University, Archaeology (Society for American Archaeology, Past President)
Kornfeld, Marcel	University of Wyoming, Archaeology
Larson, Mary Lou	University of Wyoming, Archaeology
Lin, Kai	San Diego Supercomputer Center, Geosciences Network Project, Computer Science/Informatics
Manzanilla, Linda	Universidad Nacional Autónoma de México, Archaeology
Martin, Worthy	University of Virginia, Computer Science
McCartney, Peter	Arizona State University, International Institute for Sustainability, Ecology/Archaeology/Informatics
McManamon, Francis	Department of the Interior, National Park Service, Archaeology
Nelson, Ben	Arizona State University, Archaeology
Nelson, Margaret	Arizona State University, Archaeology
Plog, Steven[a]	University of Virginia, Archaeology
Reichman, James	University of California, Santa Barbara/National Center for Ecological Analysis and Synthesis, Ecology
Richards, Julian	Archaeology Data Service, University of York, U.K.
Robertson, Ian	Stanford University, Archaeology
Schildhauer, Mark	University of California, Santa Barbara/National Center for Ecological Analysis and Synthesis, Ecology/Informatics
Schloen, David	University of Chicago, Oriental Institute, Archaeology
Simon, Arleyn	Arizona State University Archaeology
Snow, Dean	Pennsylvania State University, Archaeology (Society for American Archaeology, Secretary)
Spielmann, Katherine	Arizona State University Archaeology
Steponaitis, Vincas	University of North Carolina, Archaeology (Society for American Archaeology, Past President)
Walker, Phillip	University of California at Santa Barbara, Physical Anthropology (American Association of Physical Anthropologists, President)

[a]Steven Plog agreed that an infrastructure for preserving and accessing archaeological data is critical to the future of the discipline and agreed on the desirability of structuring a repository in a way that would enhance synthetic and comparative studies. However, he did not agree to the more particular route to such an infrastructure outlined in this report.

individuals from computer science and ecology with expertise in data integration and informatics. This report represents a consensus of all save one of the participants (Table 1).

For economy of presentation, "archaeology" is extended to include that large portion of physical anthropology that is concerned with human skeletal collections derived from archaeological contexts and related human osteological collections. These human skeletal collections constitute a vital component of the archaeological record of the human past and should be treated within the framework proposed here.

This report focuses on a key component of a cyberinfrastructure for archaeology, the integration of data represented in digital databases and other structured data sources. However, much important archaeological information also resides in unstruc-

tured documents. The discipline also needs ways to gain better access to these sources and to extract knowledge from them.

The Need

Intellectual Merit

Archaeology has the potential to provide the real-world, long-term data needed to illuminate such critical aspects of human history as demography, migration, health, economy, social stability, human environmental impacts, and environmental change from local to global scales. To date, efforts to recognize phenomena operating on large spatial and temporal scales have been crippled by the inherent complexities of archaeological data, the lack of data comparability across projects, and limited access to primary data. Indeed, scholars engaged in synthetic research are rarely able to compare their data-driven interpretations with data recovered by other archaeological projects but, rather, work with conclusions drawn by other researchers. Because the premises and data on which they are based may not be subjected to direct examination, erroneous conclusions may become entrenched in the literature as "facts" that serve as faulty premises of subsequent scientific arguments. For example, it was long thought that changes in sociopolitical complexity resulted from population pressure, and it was not until comparative demographic histories were assembled from combinations of archaeological and historical data that the subtleties of this problem came into focus (Cowgill 1975). Furthermore, important insights that could be observed when integrating multiple primary data sets are missed in working with summaries that have already "smoothed" the detail-rich data.

The potential for archaeological insights to contribute to the study of long-term social dynamics is enormous. The fundamental challenge is to enable scientifically meaningful use of the rapidly expanding corpus of data. Researchers have a pressing need for an information infrastructure that will allow them to extract a sensibly integrated and appropriately scaled database of analytically comparable observations from multiple data sets employing different recording protocols. The development of such an archaeological information infrastructure will propel synthetic and comparative research to a new level and enable researchers across scientific disciplines to address large-scale and long-term questions with a level of empirical support that heretofore has been unthinkable. With the development of appropriate data-integration tools, the Internet makes possible collaborative data sharing in ways not even imaginable 25 years ago. Archaeology must join the other sciences that have begun to tap this new reservoir of scientific potential.

The ability to integrate data will allow us to see patterns that are only visible with samples larger and more distributed than can be collected by any single project (Steckel et al. 2002). The ready availability of large-scale time-space distributional data will allow rigorous syntheses at macroregional scales (on such topics as long-distance exchange and interregional migration) that have been frustrated for decades. Because demography is a key variable underlying so many archaeological interpretations, the ability to derive consistent demographic estimates on large spatial and temporal scales is essential.

Over the last several years important issues of cultural variability have increasingly been conceptualized in terms of dimensions, rather than categories. For example, instead of classifying societies as more or less complex (i.e., chiefdoms or tribes), Nelson (1995) considers dimensions of complexity such as scale and hierarchy. Costin (1991) has discussed dimensions of craft specialization. Cowgill (1988) has similarly explored aspects of the collapse of ancient states. This focus on dimensions is a significant theoretical development that greatly enhances the meaningfulness of comparative studies. We need to encourage new work focused on other dimensions of variability instead of relying on rigid, simplistic typologies; for example, we need to move past "rural" versus "urban." The further development of theory in these arenas would be greatly enhanced by an ability to continually reevaluate arguments using refined conceptualizations of the dimensions. This task cannot be accomplished without access to *primary* data that can be provided through a cyberinfrastructure for archaeology. This information infrastructure would further allow us to find and correct interpretive errors that are embedded in contemporary interpretations, as a consequence of outdated or inappropriate inferential procedures in the past.

Because of the Native American Graves Protection and Repatriation Act and similar legislation, archaeologists now hold a huge amount of data that exist *only* in recorded form—the actual objects are now gone. Uncounted databases of those objects have been created in diverse forms, but there is no practical way for scholars or other interested individuals to access them. A cyberinfrastructure may be the only sensible approach to making that information available. Similarly, a cyberinfrastructure could help capture the vast research potential of the enormous body of—frequently underused—data from cultural resource management excavations associated with the development-related destruction of archaeological contexts.

Broader Effects

The proposed cyberinfrastructure will have disciplinary effects extending far beyond the traditional boundaries of academia. Private contractors, museums, and tribal enterprises would be active partners in building the data archives and major research users of the data-integration system. Federal, state, and tribal government entities all have key interests and responsibilities with respect to the collection and maintenance of archaeological data and would play an important role in the development and operation of the cyberinfrastructure. The shared, disciplinary cyberinfrastructure will provide a means to better sustain the utility and accessibility of irreplaceable primary data from all these sources in the face of inadequate original metadata and rapidly changing technology.

The proposed cyberinfrastructure will also have far-reaching effects on the infrastructure of social and natural science. An accurate rendering of the past is important for answering questions posed by researchers in the social, earth, and life sciences. Because of the complexity of archaeological data, it is virtually impossible for allied scientists to rely on primary data; they must depend on syntheses of archaeologists' conclusions that are themselves several steps removed from primary data. As things currently stand, outside use of the archaeological record is thus highly susceptible to faulty conclusions. A well-conceived cyberinfrastructure for archaeology, however, will allow specialists in other fields (as well as archaeologists) to gain meaningful access to an invaluable archive of primary archaeological data. They will be able to use systematized archaeological concepts (ontologies) in framing queries that will result in integrated data sets scaled and reconciled to match the scope of their research questions.

Multidisciplinary, landscape-based research will especially benefit from this information infrastructure. A geomorphologist might use archaeological ceramics to date human influences on landforms; an ecologist could use archaeological floral or faunal remains to assess the legacy effects of human use on a landscape; or a historical geographer might compare archaeological reconstructions of the demographic patterns of a region to those found historically. As archaeological data become more accessible to other scientists, we expect the interdisciplinary cooperation to increase, especially with respect to the study of coupled social and natural systems. By providing scholars in diverse fields with meaningful access to long-term data on society, population, and environment, archaeology can help explain the long-term social dynamics that have constituted today's social world and shaped the contemporary environment.

We expect both to draw on and to contribute to the body of informatics software used in the sciences. While building on current efforts in other fields, software development undertaken for a cyberinfrastructure of archaeology will lead to shareable data-representation and -integration techniques applicable to other scientific domains in which, as in archaeology, primary data are highly contextual and inconsistently collected, many inferential steps separate scientific understandings of major phenomena from observational data, and competing ontologies need to be maintained and used. We anticipate that key components of these developments will be generally applicable in a range of historical sciences. Finally, the access to real, large-scale archaeological data provided by a cyberinfrastructure will allow educators and outreach specialists to create analytical exercises and to provide public access to the results of scientific work in ways that, today, we are only beginning to imagine.

Archaeological Ethics and the Preservation and Sharing of Data

Archaeologists are ethically (and often legally) bound to deposit all records and photographs doc-

umenting fieldwork in a repository that can ensure the long-term preservation of irreplaceable archaeological data and to ensure that existing data are accessible and maintained. There is general agreement within the discipline that archaeologists are stewards of the archaeological record. Because the archaeological record is irreplaceable, archaeologists have a strong ethical responsibility to ensure the long-term preservation of archaeological materials, records, and photographs and to make them available to the scientific community so that they can inform current and future research. These issues are encoded in three of the "Society for American Archaeology Principles of Archaeological Ethics":

- *Principle No. 5—Intellectual Property*: Intellectual property, as contained in the knowledge and documents created through the study of archaeological resources, is part of the archaeological record. As such it should be treated in accord with the principles of stewardship rather than as a matter of personal possession. If there is a compelling reason, and there are no legal restrictions or strong countervailing interests, a researcher may have primary access to original materials and documents for a limited and reasonable time, after which these materials and documents must be made available to others.
- *Principle No. 6—Public Reporting and Publication*: Within a reasonable time, the knowledge archaeologists gain from investigation of the archaeological record must be presented in accessible form (through publication or other means) to as wide a range of interested publics as possible. The documents and materials on which publication and other forms of public reporting are based should be deposited in a suitable place for permanent safekeeping.
- *Principle No. 7—Records and Preservation*: Archaeologists should work actively for the preservation of, and long-term access to, archaeological collections, records, and reports (Society for American Archaeology 1996).

Although there is little professional disagreement about these principles, practice often falls short. Particularly in the case of digital data, we believe this is because of the lack of an established infrastructure and practical technologies to ensure preservation and access to databases.

Data Preservation, Access, and Integration

Data Preservation

Many, if not most, repositories that provide care and access for archaeological objects and paper records lack the expertise and facilities necessary to provide for the sustainable curation of and access to digital data. The reasons for this include the degradation or obsolescence of digital media, the obsolescence of software necessary to read the data, and a lack of sustained funding. It appears that few archaeological repositories are positioned to gain this capability in the foreseeable future. The problem of sustainable archiving is acute because, increasingly, key information is recorded exclusively in digital databases; it cannot be reconstructed from artifacts or paper records. Furthermore, newly created data sets continue to enter the same trajectory of degradation and loss that has been and is now experienced by legacy data sets. While archivists are still grappling with the problems of the long-term preservation of digital data, the incorporation of data in a cyberinfrastructure would add considerable life to critically endangered data and enormously facilitate future preservation efforts.

Data Access

Even when legacy data sets (existing data sets no longer in active use) are effectively sustained, there is no practical way to gain access to the overwhelming majority of them. In the United States, no effort to provide Internet access to existing digital data has approached the success of England's Archaeology Data Service (http://ads.ahds.ac.uk/), although there are notable initiatives. One effort to archive archaeological databases was Harrison Eiteljorg II's Archaeological Data Archive Project (centered at Bryn Mawr College), which became inactive in 2002. The National Archaeological Database (www.cast.uark.edu/other/nps/) is a comprehensive database of bibliographic references to archaeological project reports maintained by the National Park Service. Finally, online, geographic information system–based (but restricted) access to site survey data is available for a number of state and federal cultural resource inventories, primarily for cultural resource management–related purposes.

Data Integration

The simple maintenance of digital databases with their associated basic documentation (e.g., codebooks) and the provision of access to them are useful but wholly inadequate to meet the emerging demands of scientific research. What is most needed are means to integrate archaeological data across projects. Concept-oriented archaeological data integration will enable the use of existing data to answer compelling new questions and permit syntheses of archaeological data that rely not on other investigators' conclusions (that even their authors may now consider outdated) but on analyses of meaningfully integrated new and legacy data sets. In particular, because of archaeology's unique ability to provide centennial- and millennial-scale comparative data and comparative data from geographically dispersed areas, such a knowledge-based data-integration system would allow archaeology to contribute substantially to scientific understandings of long-term social dynamics.

Metadata and Ontologies

Metadata

The scientific utility of digital data is *absolutely* contingent on the availability of adequate metadata that document the data sets. Metadata include (1) syntactic information, having to do with how databases are formatted, how observations are scaled, and how the data fields are related, and (2) semantic information that documents the measured quantities; units; sampling procedures; temporal, spatial, and cultural contexts; recording procedures; and classification systems. It is only through these metadata that we can assess the comparability of observations in different data sets and determine the kinds of operations that can be meaningfully performed on them (e.g., the means of aggregating data). *Our ability to reconstruct the necessary metadata decays rapidly with time and often catastrophically with the death of the investigator.*

Ontologies

An ontology is a systematic representation of the relationships among concepts (e.g., deer *is-an* artiodactyl) and of procedural knowledge (to convert from feet to meters divide by 3.28). Embedding archaeological knowledge, definitions, and integration procedures in explicit ontologies will allow the cyberinfrastructure at once to operate using higher-level concepts (e.g., one could request counts of artiodactyls without individually listing each species) and to report fully and precisely the information it has used in processing a data request. (A highly readable summary that illuminates how ontologies can be productively employed appears in Berners-Lee et al. 2001.) The ontologies will permit, to the extent possible, the resolution of syntactic differences between data sets and the accommodation of differing semantics. It is expected that the cyberinfrastructure would maintain both widely shared ontologies for some categories of knowledge and alternative ontologies that could be selected where competing specifications of the same concepts are current (e.g., different ways of operationalizing faunal element utilities or stone tool types).

Vision: A Cyberinfrastructure for Archaeology

An effective information infrastructure for archaeology would encourage the research use of existing data, satisfy the data requirements of integrative and synthetic research, facilitate entry of data into the infrastructure, and provide for the preservation of irreplaceable data. The information infrastructure we envision will have five key components:

- a Web-based interface designed for effective scholarly access;
- sophisticated search capabilities that use concept-oriented queries to locate and access relevant data sets, images, and other systematized digital data sources while retaining the confidentiality and rights of their creators;
- data-integration tools that use syntactic and semantic digital metadata and ontologies to integrate disparate data sources, yielding a database of appropriately scaled observations with consistent variables;
- reporting capabilities that include both the basic display of integrated databases and their direct output in forms susceptible to further analysis; and

- software tools that minimize the human effort and expertise necessary to fully incorporate new data sets into the system (though both significant archaeological expertise and nontrivial effort will, of necessity, be involved).

Clearly these *demands* on a cyberinfrastructure are widely shared across scientific disciplines. Our review of the capabilities provided in or under development in geology's Geosciences Network (GEON; www.geongrid.org) and ecology's Science Environment for Ecological Knowledge (SEEK; http://seek.ecoinformatics.org) projects suggests that these initiatives can serve as reasonable models for the general functionality needed in an archaeological system.

Centralized and Distributed Functions

An archaeological cyberinfrastructure will involve both centralized and distributed maintenance of persistent data sources. Centralization will likely be effected at a national level but will also involve regional and fully distributed components. Centrally coordinated functions include

- the cyberinfrastructure technology development and maintenance,
- the establishment of metadata standards (standards and protocols that specify how data sets are documented, not how archaeological data are recorded),
- the maintenance of data discovery tools, global ontologies, and central access portals to data sources,
- systems for controlling access to sensitive information (e.g., in some cases, precise site location) and protecting intellectual property rights, and
- the acquisition and maintenance of "orphan" databases that lack an institutional home able to provide access.

Distributed functions include

- the maintenance and provision of access to institutional databases,
- maintenance of ties to central access portals,
- metadata acquisition and the incorporation of data sets into the system, and
- the research and educational use of the system worldwide.

Centralized functions are restricted to those dictated by efficiency (e.g., software development) or by compelling needs for consistency (e.g., agreement on metadata standards) or that appear necessary to ensure the longevity of the program. Regional efforts (including international nodes) will be required to deal with important metadata issues associated with regional systematics (such as typologies for ceramics, temporal periods, or archaeological cultures). Distributed functions are essential to foster professional inclusivity and to minimize long-term central costs that will be difficult to sustain. For distributed functions, it is noted that regional and local university repositories may not be best maintained by anthropology departments or museums. Libraries, information technology units, or professional data warehouse facilities may more reliably offer access to the necessary technology and personnel. Following the model of the National Center for Ecological Analysis and Synthesis, integrative research discipline-wide would be improved if the center were also to sponsor face-to-face, focused meetings of researchers collaborating on a synthetic research goal and to support these meetings with computing expertise for data integration.

Metadata Standards and Ontology Development

Data integration depends critically on a thorough encoding of both syntactic and semantic metadata and the development of useful ontologies. Although the infrastructure will not require investigators to standardize data-recording categories or procedures, metadata standards for the system will be necessary. Metadata standards specify a uniform language for the documentation of data sets, including project-, context-, and variable-specific information.

The development of metadata standards and ontologies should be undertaken by broad-based, expert work groups with the support of informatics professionals and user-friendly software tools. This development requires some centralized coordination. For some categories of information (e.g., fauna and chipped stone technologies) this may be accomplished by national and international working groups. Regional systematics (e.g., for ceramics or chipped stone typologies and regional chronologies) must be encoded by regionally based

groups. Focused, synthetic research projects that require the data integration of multiple sources may also foster the development of metadata standards.

Metadata Acquisition Tools. The infrastructure will require the development of tools for data documentation that are readily usable by an archaeologist. Data documentation means acquiring information on data set syntax and semantics, including the information needed to link more abstract concepts (used by the ontologies) to the data categories and archaeological contexts represented in the data set. These tools should minimize the human effort involved in developing complete metadata and incorporating data sets into the infrastructure's network of shared data.

Access Restrictions, Version Control, Badging, and Audit Trail. The system should allow users to place access restrictions, temporarily or permanently, to all or parts of the data they contribute (e.g., the locations of sensitive sites). The ability to impose access restrictions is intended to encourage the incorporation of new data sets into the infrastructure at or near the time of data set creation, not well after the fact. Of course, open access would be encouraged, and the contractual or permit requirements associated with the generation of the data may stipulate open access after some defined period.

Long-term research use of the system demands full version control (as is provided by GEON), such that any data sets may be revised by its authors but each earlier public version of the data set is preserved in an accessible form (thus maintaining the integrity of analyses that depend on a noncurrent version). In order that data use can be properly credited, the system must effectively "badge" data in a way that their sources are evident to users. Finally, the results of each query must be accompanied by full and readable documentation of which data archives are included, how each element of the query was transformed into primitive terms, and what data-integration procedures were utilized, so that the user can independently ensure that the results correspond to the intent of the query.

Technology Development

Sharing Cyberinfrastructure Components

A cyberinfrastructure for archaeology will involve addressing some unsolved information-integration challenges that are posed by archaeology as well as the adaptation of informatics technologies that have been newly developed in other scientific fields. As indicated above, our review of the capabilities provided in or under development in geology's GEON and ecology's SEEK projects suggests that these initiatives can serve as reasonable, initial models for the *functionality* needed in an archaeological system. This review further suggests that parts of these and related informatics efforts (including some metadata standards and limited aspects of the ontologies) could be incorporated or adapted to serve in an archaeological cyberinfrastructure. However, it is equally clear that considerable development is also needed to address the novel challenges posed by archaeology and to address as-yet-unsolved problems in semantic mediation. Open-source solutions to these archaeological and semantic mediation challenges will contribute to informatics efforts in other scientific domains. Such sharing of the software components of the cyberinfrastructure among disciplines will both promote interoperability among disciplines and reduce discipline-specific development costs.

Computer Science Challenges

Existing systems, such as SEEK and GEON, provide a base infrastructure for metadata acquisition and significant data-integration and -reporting capabilities. However, at their current stage of development, they do not address key data-integration problems posed in the domain of archaeology. As described above, a functioning integration system requires metadata that describe the resources provided by the data sets—so that relevant data sets can be identified—and describe the components of the data set in sufficient detail that superficially incompatible data sets can be integrated. Ontologies encode real-world scientific knowledge about the logical relationships among analytical concepts, permit necessary intermediate-level inferences, and supply the procedures (predicated on the metadata descriptions) needed to accomplish the reconciliation of different data sets from diverse archaeological contexts. These ontologies also allow the system to respond to concept-oriented (rather than data set–specific) queries. Obviously, some ontologies referenced by this information structure must be discipline specific.

It appears that some demands of the archaeo-

logical case pose significant computer science challenges. Primary archaeological data are extraordinarily contextual, are inconsistently collected, and are plagued by missing data that cannot be recovered through data cleaning or interpolation. In archaeology, many inferential steps typically separate scientific understandings of major phenomena from observational data. Furthermore, competing ontologies need to be maintained and used. To the extent that the meanings of some concepts are not fully shared among archaeologists, components of the system must be responsive to the needs of individuals or groups of users. Thus, for results to be meaningful to a given user, we need methodologies and semiautomated tools for mapping related (but inconsistent) ontologies for use in data resource discovery and in the on-the-fly integration of independently created data sets. Furthermore, as researchers use the system, they will formulate new concepts or alternative ways of operationalizing existing concepts. The system should "learn" by allowing the incorporation of these new conceptualizations into the ontology so that they can be reused in processing future queries.

Personnel Implications

The development of a cyberinfrastructure for archaeology will pose novel informatics challenges of interest to computer scientists. Computer scientists will have an incentive to become involved in attacking these problems, in part because solutions will translate to other scientific domains. However, the development of production versions of key tools is unlikely to come from computer science departments as they are currently constituted because, in part, of an incompatible reward structure in academic computer science departments.

Moving innovative solutions into production will require significant investment in software development by professionals at the boundaries of computer science and the knowledge domain of archaeology who are subject to reward structures compatible with achieving domain-specific accomplishments. At present, fulfilling this need is difficult and relies on finding individuals with quite rare combinations of talents. It is a priority to foster the development of an employment niche for informatics specialists who possess knowledge of the relevant domain science (in this case, archaeology and physical anthropology). This is a serious problem, already faced by other disciplines. Biology, for example, has formalized interdisciplinary training programs (in computational biology) to grow its own specialists. In the longer term, greater sharing of informatics technology across disciplines and the development of academic programs in science informatics may contribute to a solution.

Sociological Challenges: Disciplinary Participation

In the workshop, and in extensive communication with archaeological colleagues, we have encountered broad interest in, and very little resistance to, the development of a system for data sharing and a cyberinfrastructure, if it can be made practical. The authors believe that archaeologists' wide acceptance of ethical responsibilities for data sharing will translate reasonably smoothly into contributions of their data to and use of a cyberinfrastructure for archaeology. We are convinced that participation would be sufficiently widespread and the research payoff would be sufficiently large to more than justify the investment in development (by funding agencies) and incorporation of data sets in the system (by users). Furthermore, over time, the level of acceptance will rise with the demonstration of its benefits and as a failure to contribute becomes negatively valued professionally.

As in other sciences, the more useful the collaborative system is to those generating the data, the more likely they will be to participate. A productive strategy to encourage use would be to fast-track developments that have the potential to yield high returns for relatively low investments. With this, users would see the direct payoff of contributing to the system.

Were funding agencies (e.g., the National Science Foundation, National Endowment for the Humanities, National Geographic Society), permitting agencies (State Historic Preservation Offices, Tribal Historic Preservation Offices, the Advisory Council on Historic Preservation, and other federal, state, and tribal entities), or publication outlets to require incorporation of the data in the infrastructure as a condition of the grant, contract, or publication, participation could be greatly enhanced. Contracts and grants could make subsequent funding or permitting dependent on docu-

menting the data sets and making data available through the system proposed here. As has happened in other disciplines, journals could require certification that the data supporting published papers are available within the system.

Process Recommendations

It is imperative that the development of a cyberinfrastructure for archaeology begin immediately. Irreplaceable information is continually being lost, and the project of incorporating legacy data grows larger with delay.

Steering Committee

To begin, a formal steering group should be established that has strong linkages to the Society for American Archaeology, the American Association of Physical Anthropologists, and the Society for Historical Archaeology in order to

- develop a strategic plan that will set the direction for the long-term effort;
- acquire, in cooperation with the National Science Foundation and other funding sources, funding for cyberinfrastructure development and the execution of test cases that will demonstrate the potential of this approach;
- explore the utility of existing models for cyberinfrastructure development in other fields and participate in collaborative efforts to develop shared cyberinfrastructural tools;
- develop mechanisms for dialogue at regional, national, and international scales that will be needed for effective metadata standards; and
- work with professional societies to secure agreements needed in developing a cyberinfrastructure for archaeology and to foster the practice of data sharing.

National Center

Following the development of a strategic plan by the steering committee, a national center for archaeological data integration should be established with both startup funding for development and long-term, continuing funding. Continuing funding would also allow the national center to coordinate software maintenance and development.

The overall goal of the center would be to promote integrative and synthetic research using archaeological data. It will thus provide essential ongoing benefits to archaeology and to long-term studies in other scientific disciplines. It will administer the development of software tools for metadata acquisition and integrated data access. It will collaborate with informatics efforts in other scientific disciplines. It will coordinate the international, national, and regional development of standards for metadata. It will serve as a permanent node for data resource discovery and access and as a repository for documented data sets that cannot be maintained at a distributed node. Nonetheless, it must be recognized that the ultimate success of the cyberinfrastructure rests with the whole archaeological community.

The center should sponsor a small number of low-cost, high-reward "proof of concept" projects that would demonstrate the potential for data integration to the broader archaeological community. As suggested above, this goal would be advanced if the center (like the National Center for Ecological Analysis and Synthesis) were able to provide competitive funding and technical support for face-to-face collaboration on targeted integrative research topics.

Compelling test cases must be developed on both focused research issues and larger-scale research topics. Test cases should illustrate the relevance for different segments of the profession (including cultural resource management, physical anthropology, and historic archaeology), and some should have cross-disciplinary components (e.g., with ecology and earth science). For example, the workshop concluded that an excellent candidate for one of these test cases would be the archaeology of central Mexico. There, a lengthy sequence of prehistoric occupation has been well documented archaeologically—with excavation and survey—by numerous archaeological projects pursued by a number of investigators, both Mexican and international. The incorporation of a number of key databases into the system will demonstrate its ability to deal with real-world diversity in data-recording schemes. More important, the Web availability of integrated data derived from these projects will be an extraordinarily valuable resource for research into demographic change, settlement organization, and political complexity that would be of

international interest and utility. (Also, several of the key data sets would be immediately available for this purpose.)

Professional Societies

Professional societies, including the Society for American Archaeology, the American Association of Physical Anthropologists, and the Society for Historical Archaeology should make it a high priority to assist the profession in understanding the importance of metadata in sustaining the research utility of data. Professional societies can further contribute to this effort by encouraging agencies that permit or fund archaeological research to require the incorporation of data sets into the cyberinfrastructure, by fostering reward structures for systematic data sharing, and by ingraining the ethical obligations to share data as expected professional practice. They can promote the training of students to pursue synthetic research topics and the use of legacy data. Through their journals, they can develop standards for the citation of others' data (rather than just acknowledgment of their use). Similarly they could establish as a professional standard a general public license for data that states that the data may be used as long as it is properly credited and so long as any products of use of the data are also openly available. Finally, professional societies can promote funding for the development and maintenance of a sorely needed cyberinfrastructure for archaeology.

Acknowledgments. This report represents a collective effort of the project participants. I am grateful to them for both their contributions to the workshop and their efforts and perseverance in working through several drafts to achieve a near-consensus report. We gratefully acknowledge the National Science Foundation, whose Human Social Dynamics program funded this workshop through grant number SES 0433959.

References Cited

Berners-Lee, Tim, James Hendler, and Ora Lassila
2001 The Semantic Web. *Scientific American* 284(5):34–43.

Costin, Catherine L.
1991 Craft Specialization: Issues in Defining, Documenting, and Explaining the Organization of Production. In *Archaeological Method and Theory*, edited by M. Schiffer, 3:1–56. University of Arizona Press, Tucson.

Cowgill, George L.
1975 Population Pressure as a Non-Explanation. In *Population Studies in Archaeological and Biological Anthropology*, edited by A. C. Swedlund. Society for American Archaeology Memoir 30. *American Antiquity* 40:127–131.
1988 Onward and Upward with Collapse. In *The Collapse of Ancient States and Civilizations*, edited by Norman Yoffee and George L. Cowgill, pp. 244–276. University of Arizona Press, Tucson.

Nelson, Ben A.
1995 Complexity, Hierarchy, and Scale: A Controlled Comparison between Chaco Canyon, New Mexico, and La Quemada, Zacatecas. *American Antiquity* 60(4):597–618.

Society for American Archaeology
1996 Society for American Archeology Principles of Archaeological Ethics. *American Antiquity* 61:451–452.

Steckel, Richard H., Jerome C. Rose, Clark Spencer Larson, and Phillip L. Walker
2002 Skeletal Health in the Western Hemisphere from 4000 B.C to the Present. *Evolutionary Anthropology* 11:142–155.

Received July 25, 2003; Accepted July 28, 2005.