INFORMATION SCIENCE

Cybertools and Archaeology

Dean R. Snow,[1]* Mark Gahegan,[2] C. Lee Giles,[3] Kenneth G. Hirth,[1] George R. Milner,[1] Prasenjit Mitra,[3] James Z. Wang[3]

In archeology and other historical sciences, diverse, widely distributed data include artifacts, notes, field logs, and other records. Future research requires that these archives be electronically accessible and user-friendly.

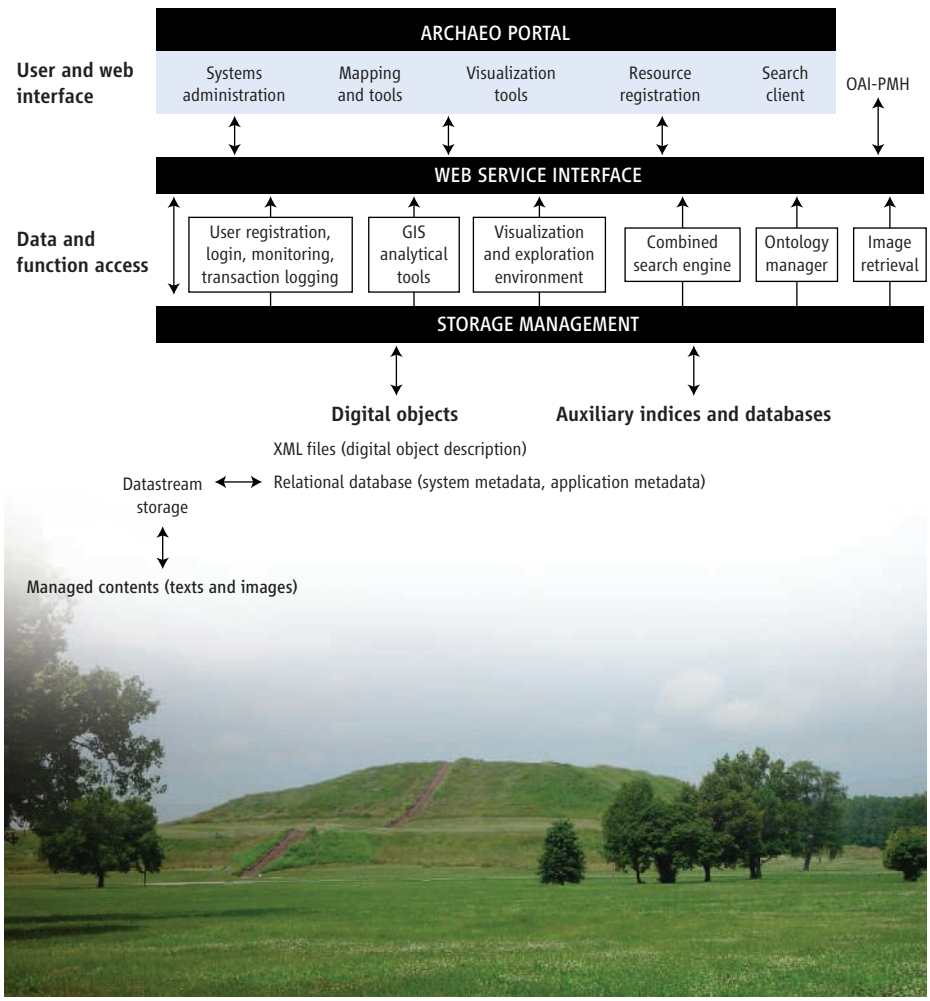
The need for service-oriented cyberinfrastructure (CI) has been reported (*1*–*4*). Further development of archiving and search tools to accommodate the explosive growth in many fields, particularly in biomedical research, has been emphasized. Archaeology often depends on archived data acquired by other researchers for other purposes, often long ago. Differences in recording protocols, terms, measurement units, and language are commonplace. Data are often obscurely archived and difficult to access, and policies regarding confidentiality vary considerably. Even when databases are accessible, they often differ in size, format, structure, and semantics and seem to defy fusion. In archaeology, research on the most important issues in today's society—the evolution of culture, the growth in population, and the long-term interaction of cultures with their physical and biological environments—will remain impoverished in the absence of a new generation of cybertools.

Enhanced online at www.sciencemag.org/cgi/content/full/311/5763/958

Modern archaeological science depends on large collections of diverse, mundane objects (such as potsherds, stone tools and debris, and animal and plant remains), rather than small collections of treasures. Sites are unique, nonrenewable resources easily destroyed by erosion or modern land use. Thus, old collections, original field notes, and reports of prior work have enduring research value.

At present, there are three types of data that are impossible to access simultaneously because of the highly individualized nature of traditional archaeological field and laboratory research. First, there are separately compiled databases held by museums, governmental agencies, and individuals that reside on different computer platforms. Data classifications and terminology vary, are regionally and temporally specific, and are inconsistently applied. Increasingly, these are Geographic Information System (GIS) databases based on years of accumulated paper records. Second, there is a voluminous unpublished "gray literature" consisting of limited distribution reports (produced mainly by cultural resource management firms and government agencies). Third, there are images, maps, and photographs embedded in museum catalogs and archaeological reports (published and unpublished). Difficulties in accessing data have been aggravated by the boom in cultural resource management (CRM) research in the United States. Government agencies, museums, universities, and private companies have acquired, and now care for, tens of millions of artifacts plus associated field notes and metadata.

The dimensions of the problem for the United States were outlined in a recent white paper commissioned by the Society for American Archaeology (SAA), in which it was estimated that six federal agencies alone require



Cyberinfrastructure architecture for archaeology. This proposed system integrates digital library middleware, document and image search, GIS analytical kits, visualization tools, and content management. The OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) provides an application-independent interoperability framework for metadata harvesting by other repositories and similar systems. The architecture could be built completely from existing systems, some of which are open source. For example, Fedora is an open-source middleware for managing and serving digital objects and repositories. GeoTools is an open-source Java toolkit for producing interactive maps on the Web and GeoVISTA Studio is an open software development environment designed for geospatial data that allows users to quickly build applications for geocomputation and geographic visualization. SIMPLicity is a content-based image search and automatic learning-based linguistic indexing system. These are meant to be examples of software possibilities; no specific product endorsement is intended. Monk's Mound, Cahokia Mounds, Collinsville, IL, is pictured.

PHOTO CREDIT: D. R. SNOW

[1]Department of Anthropology, [2]Department of Geography, [3]School of Information Sciences and Technology, The Pennsylvania State University, University Park, PA 16802, USA.

*Author for correspondence. E-mail: drs17@psu.edu

Published by AAAS

~64 million cubic feet of collection storage space in addition to over 40,000 linear feet of documentation (*5*). These numbers do not include collections, images, and catalog documentation maintained by universities and museums (including the Smithsonian Institution) or state agencies across the country (*6*).

The problem is not only one of access. The various autonomous user communities currently hosting these resources understand and describe data in different ways. There is a need to classify and search for numerical, textual, and visual data simultaneously. Standards and protocols, where they exist, are typically defined at state or local levels and have changed over time. Diverse research interests and reliance on old collections require archaeologists first to understand each other's concepts and procedures in order to comprehend others' data, methods, and results. Translation protocols are needed. Regional data have often been collected and organized according to modern political boundaries that have no meaning whatsoever in prehistoric, early historic, or environmental contexts. Furthermore, some data, particularly precise site locations, must be kept confidential at levels that vary according to state and local policies. Consequently, archaeological research remains a mosaic of parochial efforts. Even the spatial and temporal dimensions of regional cultures must be extrapolated from selected sites and radiocarbon dates. Research on large geographical areas is particularly difficult at present.

Recent developments in computer and information science provide the computational tools, protocols, and standards that can help devise an integrated infrastructure. Federated databases and ontology-based database integration (*7*–*9*) provide the means of coordinated data management. Portal technology and Web services provide customizable access points to methods and data (*9*), and grid technology, coupled with high-speed data connections, provides distributed but high-powered computational resources (*10*). In several sciences, it is already apparent that a coordinated approach to describing, archiving, and disseminating research products and services can provide significant gains in productivity and quality. For example, the National Science Foundation (NSF) is currently funding the development of CI in support of the human-environment interaction and geoscience communities via the Human-Environment Regional Observatory Network (HERO) (*11*) and Geosciences Network (GEON) (*12*) projects. Within government, the National Map, the National Spatial Data Infrastructure (*13*), and Geospatial OneStop are federally mandated initiatives to improve sharing and collaboration among data collection agencies and users. A digital library and search engine for computer and information science, CiteSeer (*14*, *15*), is very popular with the computer science community. The SIMPLIcity (Semantics-sensitive Integrated Matching for Picture Libraries) content-based image retrieval engine has been used to manage large-scale databases for art and cultural, remote-sensing, biomedical, and Web applications (*16*). Readers can try it out online. Rather than try to impose a single formal data model and associated semantics on the community of researchers, new tools should instead take an approach that encompasses many different perspectives by developing database mediation services based on successful current approaches such as those used by GEON (*17*), the Fedora Digital Library (*18*), and ONION [Ontology Composition System (*19*)]. Thus, researchers should be able to query electronic data using search terms that have meaning to them, and these terms should be mapped to semantic equivalents when used to query remote data collections, then translated back. The CI architecture shown (see figure, page 958) is an example of an approach to facilitate use of archaeological resources via Web/grid services.

At the same time, any new system should facilitate future efforts within the archaeological community to establish common, minimal standards for metadata descriptions of artifacts, sites, maps, and other academic resources (*20*). Thus, interoperability is not simply a technical end-goal. It is instead a design strategy that also promotes effective cooperation between human and electronic components of the research process. Such efforts have begun at universities such as The Pennsylvania State University, Arizona State University, and the University of Arkansas and at the National Park Service. We require an e-science that marries the interconnectedness of digital research tools with the introspection enabled by traditional record-keeping (*21*, *22*).

Sustainability can be assured in two ways. First, data collections should be distributed and sharable. Host institutions should retain the freedom to manage their own databases for their own purposes, thereby spreading costs and maintaining institutional autonomy. Second, digital libraries and associated services should be made available to researchers and organizations to store their own data and mirror data of others. Because databases can remain in place, once the infrastructure is completed, the running costs would be restricted to maintenance and refinement of the metadata collections and associated services. Such a distributed system avoids trying to manage an ever-increasing centralized digital archive into the foreseeable future, with significant recurrent annual costs, but does bring up problems associated with federated search and management. Harvesting and indexing services can continue with minimum support and can be replicated by other organizations, so functionality can become truly distributed over time. The best system will be one that has commitments from government, academic, and commercial organizations.

Emerging cybertools can transform the way in which researchers collaborate to solve long-standing problems by providing: (i) a stable set of catalogs to preserve what is known about important data resources; (ii) tools to help researchers locate, access, and contribute data resources; and (iii) shared virtual workspaces in which researchers can collaborate virtually on larger tasks (see figure, page 958). Any attempt to impose a single complex (and expensive) system will fail. CI will be successful if it is allowed to evolve as it is adopted, used, and contributed to by a community. Encouraging archaeologists to do so also involves solving problems of confidentiality and trust, and securing long-term commitment from agencies.

References and Notes

1. I. Foster, *Science* **308**, 814 (2005).
2. T. Hey, A. E. Trefethen, *Science* **308**, 817 (2005).
3. K. H. Buetow, *Science* **308**, 821 (2005).
4. S. M. Maurer, R. B. Firestone, C. R. Scriver, *Nature* **405**, 117 (2000).
5. S. T. Childs, K. Kinsey, "Costs of curating archeological collections: A study of repository fees in 2002 and 1997–98" (Tech. Rep. 1, National Park Service, Washington, DC, 2003).
6. S. T. Childs, K. Kinsey, "A survey of SHPO archeological report bibliographic systems, 2002" (Tech. Rep. 5, National Park Service, Washington, DC, 2004).
7. A. Elmagarmid, M. Rusinkiewicz, A. Sheth, Eds., *Management of Heterogeneous and Autonomous Database Systems* (Morgan Kaufmann, San Francisco, 1999).
8. M. A. Rodriguez, M. J. Egenhofer, *IEEE (Inst. Electr. Electron. Eng.) Trans. Knowl. Data Eng.* **15**, 442 (2003).
9. M. Chau *et al.*, paper presented at the Second Association for Computing Machinery and IEEE Computer Society, Joint Conference on Digital Libraries (JCDL '02), Portland, OR, 14 to 18 July 2002.
10. The Globus Alliance (www.globus.org).
11. HERO (http://hero.geog.psu.edu/index.jsp).
12. GEON (www.geongrid.org).
13. NSDI (www.nsdipa.gr.jp/english/index.html).
14. C. L. Giles, K. Bollacker, S. Lawrence, in *Proceedings of the Third ACM Conference on Digital Libraries* (ACM Press, New York, 1998), pp. 89–98.
15. S. Lawrence, C. L. Giles, K. Bollacker, *IEEE Comput.* **32** (6), 67 (1999).
16. J. Z. Wang, J. Li, G. Wiederhold, *IEEE Trans. Pattern Anal. Machine Intell.* **23**, 947 (2001); (www-db.stanford.edu/~wangz/project/imsearch/SIMPLIcity/TPAMI/).
17. P. Tooby, "Speeding scientific workflows: The open-source Kepler Project," 3 May 2005 (San Diego Supercomputer Center, San Diego, CA, 2005); (www.sdsc.edu/Press/2005/05/050305_kepler.html).
18. C. Lagoze, S. Payette, E. Shin, C. Wilper, *Int. J. Digital Libr.*, in press; (draft available at http://arxiv.org/ftp/cs/papers/0501/0501012.pdf).
19. P. Mitra, G. Wiederhold, in Workshop on Ontologies and Semantic Interoperability, *Proceedings of ECAI 2002, the 15th European Conference on Artificial Intelligence,* Lyon, France, 21 to 26 July 2002 (IOS Press, Amsterdam, 2002).
20. W. A. Pike, O. Ahlqvist, M. Gahegan, S. Oswal, Workshop on Semantic Web Technologies for Searching and Retrieving Scientific Data, at the Second International Semantic Web Conference, Sanibel Island, FL, 20 to 23 October 2003; (http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-83/sia_1.pdf).
21. M. S. Carroll, Ed., *Delivering Archaeological Information Electronically* (Society for American Archaeology, Washington, DC, 2002).
22. R. D. Drennan, S. Mora, Eds., *Archaeological Research and Heritage Preservation in the Americas* (Society for American Archaeology, Washington, DC, 2001).

10.1126/science.1121556