

# The Challenge of Archaeological Data Integration<sup>1</sup>

Keith W. Kintigh  
School of Human Evolution & Social Change  
Global Institute of Sustainability  
Arizona State University  
Tempe, Arizona 85287-2402, USA  
[kintigh@asu.edu](mailto:kintigh@asu.edu)

## Abstract

Archaeological insights have enormous potential to contribute to the understanding of long-term social and socioecological dynamics. However, the complexities of archaeological datasets, the lack of data comparability across projects, and limited access to primary data have crippled our efforts to understand phenomena operating on large spatial and temporal scales. The fundamental challenge is to enable scientifically meaningful integration and use of the expanding corpus of systematically collected archaeological data. A two-year long investigation of the information-integration demands of archaeology has revealed fundamental technical challenges that cannot be met by simply adapting existing technologies. Our work suggests the need for a knowledge-based archaeological data integration system. It would include a distributed archive of digital data with rich semantic descriptions of dataset content that is accessed by a concept-oriented, query-driven integration system capable of mediating between the semantics of a research question and the semantic content of archive's data sources. Its output would be an informed database of sensibly integrated and appropriately scaled observations and consistent variables. This paper discusses the key design considerations and the pragmatic challenges presented by the development of such a system.

Key words: data integration, data archiving

## Le défi de l'intégration des données archéologiques

### Résumé

Les efforts de l'archéologie pour saisir les phénomènes qui interviennent à des niveaux spatiaux et temporels larges se heurtent à la complexités des données archéologiques, à l'absence de données de comparaison et à l'accès limité aux données primaires. Nos travaux indiquent qu'une intégration scientifiquement significative des données archéologiques systématiquement recueillies requiert un knowledge-based archaeological data integration system (système informatique d'intégration des données archéologique basé sur la connaissance). Un corpus des données numériques en réseau, avec de riches descriptions sémantiques, pourrait être accessible via un système — conceptuellement orienté et de questionnements intégrés — capable d'intervenir entre le contenu sémantique d'une question et le contenu sémantique d'un corpus de référence. Le résultat serait une base de données de variables cohérentes et d'observations scalaires 'raisonnablement' intégrées. La question est ici celle des concepts fondamentaux et des défis pragmatiques que posent ce type de système.

Key Words: système informatique d'intégration des données, corpus des données numériques

---

<sup>1</sup> Paper presented in the Commission 4 session, Technology and Methodology for Archaeological Practice: Practical Applications for the Reconstruction of the Past, organized by Alexandra Velho and Hans Kamermans, at the meeting of the Union Internationale des Sciences Préhistoriques et Protohistoriques, Lisbon, 4-9 September 2006.

## **Introduction**

In this paper, I discuss some recent research at Arizona State University (ASU) directed to the development of one component of an advanced information infrastructure for archaeology. First, I will summarize a shared vision for a Web-accessible, distributed information system for systematically collected archaeological data—data of the sort that we ordinarily store in databases. Then, I will present some of our conclusions about the shape that such an infrastructure must take and outline some of our ongoing and prospective research on a proof-of-concept system focused on faunal data.

However, at the outset I also want to acknowledge that some of the fundamental problems have recognized for many years (notably, by Harrison Eiteljorg) and that related efforts are ongoing, notably in England with the Archaeological Data Service at the University of York (<http://ads.ahds.ac.uk>). This is an enormous problem and will require international engagement and collaboration in a collective effort to improve the research access, research utility, and sustainability of irreplaceable archaeological data.

### **Developing a Vision for an Information Infrastructure for Archaeology**

#### **Archaeological Informatics**

Archaeology desperately needs a sophisticated, computer-based, information infrastructure that will both enable the effective use of new and legacy data to advance archaeological knowledge and that will stem the loss of irreplaceable data, much of which exists only in digital form (Snow et al. 2006). This will require that archaeology develop active initiatives in archaeological informatics much as other scientific disciplines are doing. Of course, bioinformatics is now a well-developed field, but other scientific disciplines (e.g., geology and ecology) are also building information infrastructures.

Within the scope of what we might conceive of as archaeological informatics there are at least three domains that might be considered separately. The first, and the one on which the ASU team has chosen to focus, is systematically collected primary data, information that is typically recorded on forms or electronically and that today ordinarily ends up in computer databases. Secondly, a great deal of archaeological information is available only in reports, notes, and other sorts of unstructured text. Some of this information is published in articles, books and other relatively accessible forms, but much of it has very limited distribution (the so-called gray literature) or exists in only one or a few copies in scholars' files or museum archives. Third, we need to consider graphical data including maps, drawings, and photographs as well as GIS

#### **Initial Steps**

In 1999, we began to build a team, develop an approach, and seek funding for an information infrastructure in archaeology, but it was not until 2004 that we received funding from the US National Science Foundation (NSF). While we had sought funding to begin implementation, NSF provided a smaller grant and asked that we work to develop a disciplinary consensus on the desirability and feasibility of the ambitious initiative of the sort that we had proposed.

As a consequence, in 2004 we organized a workshop hosted by the National Center for Ecological Analysis and Synthesis in Santa Barbara, California. The workshop's goals were: 1) to develop a shared disciplinary vision for information infrastructure for archaeology; 2) to assess the sociological and technical challenges to such an infrastructure; and 3) to outline an implementation strategy. Thirty-one scholars participated in the workshop including archaeologists and physical anthropologists from diverse areas and with varied interests, computer scientists concerned with information integration and informatics, and practicing scientists working on informatics projects in other disciplines.

## The Importance and Difficulty of Archaeological Synthesis

In the workshop, there was broad agreement that archaeology's potential to contribute to scientific understandings of socio-ecological dynamics depends upon our ability to synthesize our long-term and spatially extensive data on society, economy, and environment. It was also abundantly clear that synthetic analyses of primary archaeological data are, at the very least, extremely time-consuming and exceedingly difficult. It appears that syntheses generally rely on the author's personal experience and reading of other excavators' conclusions, *not their primary data*.<sup>2</sup> Thus, faulty inferences easily become entrenched in the literature as "facts," and are often difficult to detect and correct.

Anyone who has tried to integrating primary data across projects will need no convincing of the difficulty of this task. The problems are numerous. Most obviously, it is often difficult or impossible to locate the primary data in any form—digital or otherwise. Given the data, it is often difficult or impossible to gain adequate *metadata*, documentation that describe what the data *mean*. If an investigator reports a certain ceramic type, how was that type distinguished from a similar one? What size animals fit in the category "unidentified small mammal"—is a dog large or small? Without systematic metadata, the research value of the data are severely limited. More broadly, there is a lack of data comparability across projects. How are inconsistent typologies to be used together? Finally, while archaeological databases tend not to be large—compared with many scientific domains (climatology, for example)—they almost always have exceedingly complex hierarchical structures, reflecting both the complicated relationships among observational contexts, the different scales at which we record data, and the complexity of our recording strategies for the many kinds of observations we make.

## Archaeological Data Preservation

In addition to the goal of improving the accessibility and overall utility of archaeological data, the field faces extraordinary problems of data loss. Archaeological data obtained at great expense are being lost, often irretrievably, at an alarming rate. In some of our recent synthetic efforts, we have not been able to locate primary data, even for recent projects with the full cooperation of the investigator. Digital data are being lost through degradation of electronic media, and software obsolescence. Increasingly, primary archaeological data are initially acquired in digital form. Furthermore, today the management of primary data (including corrections to original data and linkages to other data observations) is almost always through digital databases. This increases the urgency of providing a means to preserve digital data in a ways that are sustainable in the long term. However, probably the most difficult problem that the discipline faces in maintaining digital data is the loss of the metadata necessary to give meaning to the observations contained in the databases. As databases evolve as a project proceeds, coding keys (if they ever existed) become outdated. Often essential metadata is irreversibly lost with the death or incapacity of the excavator or specialist responsible.

We frequently—and rightly—maintain that archaeological data are irreplaceable—once we have excavated a context, it is gone. As a consequence, I don't think we can overstate the urgency of action to preserve the records of our observations.

---

<sup>2</sup> By "primary data" I mean observations that are relatively direct and contain little interpretation. Thus, I would say that a radiocarbon assay of 1200BP±60 is a primary observation, whereas the dating of an occupation of a structure is inferential. Similarly I would say that reporting 27 potsherds of a particular type from a certain context is also a primary observation, though there might be reasonable debate about the reliability of that observation.

## Vision for an Information Infrastructure for Systematically Recorded Data

The Santa Barbara workshop included presentations of ongoing (well-funded) informatics efforts in geology (GEON; <http://www.geongrid.org/>) and ecology (SEEK; <http://seek.ecoinformatics.org>). These demonstrations and subsequent discussion helped inform a vision for systematically collected archaeological data. There was a consensus that what is needed is an archaeological information infrastructure that will sustain the utility of new and legacy data and that will provide tools to integrate data so that we can address compelling research questions at large spatial and temporal scales. Such tools can be expected to foster a new paradigm of synthetic and integrative research.

As we now see it, this information infrastructure would depend on open-source software and would have several key components:

- A network of distributed data sources that would allow scholars to control their own datasets (to the extent desired) and would provide for the maintenance in regional or national repositories of “orphan” datasets for which no one wishes to take ongoing responsibility.
- A “registration” process through which datasets would join this information network. Registration of a data source on the network would include specification of its location on the Internet, access restrictions, and metadata describing the content of the dataset.
- Software tools that facilitate the registration of new data sources while fostering the collection and maintenance of adequate metadata.
- A Web-based *concept-oriented* interface designed for scholarly inquiry. Through use of structured representations of archaeological concepts and their relationships to one another (encoded in ontologies), scholars would be able to use the system without knowledge of the details of the structure or content of individual data sources.
- Sophisticated search capabilities that use metadata associated with the registered projects to locate potentially relevant datasets based on the scholar’s query.
- Data integration tools that, based on a query, would use these ontologies, digital metadata that describe the datasets, and user guidance to integrate data from multiple sources.
- Reporting capabilities for basic data display and, more importantly, output of digital databases of appropriately scaled and comparable observations that the scholar could further analyze.

## Workshop Report

The workshop’s full report was published in *American Antiquity* (Kintigh 2006) and is available on-line at <http://cadi.asu.edu>. A summary appeared in the American Anthropological Association’s *Anthropology News* (Kintigh 2005). This report provides a much more complete specification of this vision and a thorough discussion of the ethical mandates and the implications of such a network for archaeological scholarship—referred to above as the sociological challenges. It also provides some pragmatic discussion of how such a system could get up and running and how scholars could be induced to contribute and register their datasets.<sup>3</sup>

---

<sup>3</sup> The Society for American Archaeology has developed a Digital Data Interest Group. Its purpose is “To promote the preservation and sharing of archaeological data maintained in digital form. The long-term conservation and protection of the archaeological record demands that we preserve digital documents, images, and databases, and make them available to other scholars in order to advance archaeological understandings of the past. The interest group will foster the development of shared digital archives of archaeological data. It will promote data sharing and

To gain credibility both in the scholarly community and by US funding agencies, we sought and received endorsement of this report and its conclusions from three major scholarly organizations in the US, the Society for American Archaeology, the American Association of Physical Anthropologists, and the Society for Historical Archaeology.

### **Implementation Research**

In addition to the Santa Barbara Workshop, we've been working over the last two years to refine this vision, to frame a development strategy in more technical terms and to seek implementation funding. This effort has involved intensive interaction of the ASU team of computer scientists and archaeologists in the context of work on a more focused issue of archaeological informatics—integrating datasets of archaeological fauna. The effort was greatly improved through the generous assistance of a working group of expert faunal analysts.

The vision described above was fundamentally formulated in archaeological terms within an archaeological frame of reference. A key step was to articulate this vision in a language that communicates these objectives within a Computer Science perspective. One of our most challenging tasks has been for the computer scientists and archaeologists to develop a common understanding of the problem and common language to discuss it. Well into the project, at a time that we thought we were pretty fully comprehending each others' perceptions of key issues, the discussion would occasionally reveal fundamental misunderstandings between the two groups.

Quite understandably, the computer scientists, initially, tried to locate our problem specification within the realm our information sharing and integration problems that had already been successfully addressed in computer science. If that could be done, then what was needed was perhaps a great deal of work, but ultimately amounted to an application of existing knowledge to our particular problem. *If* there were really no computer science challenges the project was of marginal interest to the computer scientists.

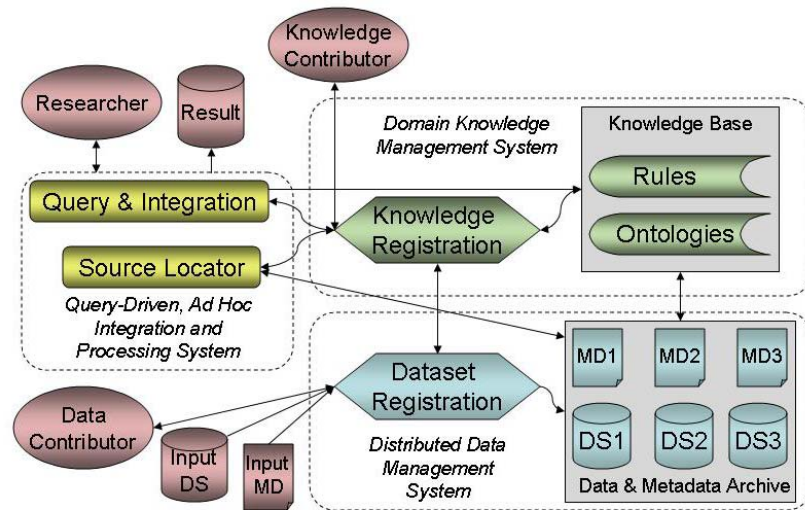
However, it has turned that out as we worked through what needed to be done to integrate relatively straightforward faunal datasets, that there are indeed unsolved computer science problems, and with the archaeologists, the computer scientists began to grapple with a reformulated view of the problem and to start to discern, in more detail, where the interesting problems lay and to begin to sketch a system design (Figure 1).

### **Technical Demands and Conclusions**

As we refined our statement of the problem it became clear that the system must: integrate data collected at different scales, at different times, by different investigators using variable data recovery strategies and inconsistent typologies; adequately encode complex typologies, data recording schemes, archaeological contexts, and recovery techniques; and, most importantly that it is neither necessary nor advisable to reduce data to a single standard at registration time. *Instead, the semantics of new and legacy data must be preserved.*

---

preservation to the broader archaeological community and enhance communication and collaboration among data sharing initiatives.” Contact: [ekansa@alexandriaarchive.org](mailto:ekansa@alexandriaarchive.org).



**Figure 1. Schematic Design for a Knowledge-Based Archaeological Data Integration System (KADIS).**

Standardizing data at the time it is registered is a common (e.g., in economics), seemingly obvious, and often useful solution that makes data integration easy but, for archaeology, introduces unacceptable compromises. If data were standardized at registration time and the data standards were high, much legacy data would be discarded. If the standards represented a lowest common denominator of recovery techniques and data recording over the last century, we would lose all of our high precision data. Any sort of intermediate compromise would entail loss of access to data that did not meet whatever standard was set, regardless of the fact that *for some questions* these low resolution datasets may be perfectly adequate. Furthermore, the history of our field shows that data standards do, and should, continue to evolve.<sup>4</sup>

The key conclusion that we reached is that we must reconcile the semantic requirements of a user query with the semantic content of the available data sources. Using this strategy of ad-hoc data integration, the integration is accomplished in the context of a query, not at the time of dataset registration. Thus, the system must: 1) determine the semantic data requirements entailed by the query; 2) identify potentially relevant data sources among the registered data sets; 3) must perform on-demand metadata matching to align key portions of the selected datasets (which entails reasoning with potentially incomplete and inconsistent information and using multiple ontologies and user assistance in deciding on irresolvable mismatches and data quality/quantity compromises); and 4) output appropriately scaled, analytically comparable observations along with a readable log of the decision-making and transformations that went into the integration.

Our final conclusion was that direct adaptation of existing information integration technologies is unworkable, though there are important components of existing applications that can be used or adapted. As is discussed more below, we believe that the tools that we develop will be useful in some of these other efforts.

<sup>4</sup> This is not at all to say that I am opposed to data standards of any sort. There is much to be gained if investigators can agree, at given time for a given class of contexts, to standardize their observations. My opposition is to imposing a one-time standard, both because I think it is philosophically unwise over the long run, and pragmatically, because it would cause the collapse of the proposed enterprise before it ever got started.

## **Prospective Research**

As our next step, we have proposed a major effort to develop a proof-of-concept application that will implement the key components of this information infrastructure, for a sub-domain of archaeology, faunal analysis. Even with this limited scope the project is ambitious because many of the key software components must be built and a number of the key technical problems must be solved. However, we will be working with related science informatics efforts and it will be possible to adapt some of their work to our needs.

The choice of faunal data as a testbed perhaps warrants some explanation. First, in our initial grant we had begun with this topic and it had worked well for developing joint understandings with our computer science colleagues. Second, we felt that it was of intermediate difficulty, because it combined both natural taxonomies (taxon and element; with variant elaborations for various groupings of taxa and elements that could not be fully identified) and alternative classification schemes for such things as fragmentation and burning. Third, we believe that we can get something up and running that will have international utility much more quickly than with a material such as ceramics with their innumerable regional taxonomies and conventions. In this effort we have enlisted the support of the International Council for Archaeozoology (ICAZ) which had, in 2004, established a Data Archiving Task Force whose goals aligned well with ours.

The goal for our prospective research is then to establish an international, Web-accessible, distributed network of archaeological (fauna) data sources. This network will be populated using open-source software tools that will facilitate entry of new datasets along with the metadata necessary to interpret them. As described above, the interface will be concept-oriented so that knowledge of specific datasets will not be necessary. We will first develop the software to identify and download relevant datasets and their associated metadata and then tackle the more challenging task of using archaeological knowledge to integrate information within diverse databases to provide users with truly integrated data across multiple data sources.

Expanding our faunal system to other classes of material will provide some additional challenges. However we are optimistic that the proof-of-concept system will provide a model that is sufficiently persuasive to attract both the user support and the funding to make an expansive information infrastructure for systematically collected archaeological data a reality.

### **Promise of Data Integration**

The vision articulated in this article would do much to provide for the long term preservation of archaeological data and its semantic content and to stem the ongoing loss of irreplaceable data. It would expand international access to primary archaeological data and the ability of archaeologists to do integrative and synthetic work, allowing us to address questions on temporal and spatial scales that have heretofore been unthinkable, and to reevaluate key premises of our discipline's accepted wisdom. Such a system will massively improve our ability to model long-term stability and change in coupled social and environmental systems. Finally, it will make archaeological and environmental data meaningful to scholars in other fields, e.g., biologists looking at long-term species decline.

We believe that that our conceptual design of an ad-hoc, query-driven data integration tools will be useful in historical and other sciences that share key characteristics with archaeology: heterogeneous data sources; highly contextual data; competing taxonomies and definitions; inconsistent methods of data collection; frequent missing values; and data in which many inferential steps separate observations from the variables of interest.

This paper has not addressed many key problems to which we have devoted considerable attention. However, I hope that it will stimulate interest and collaborative efforts to contribute to a

workable, international information infrastructure for archaeology. There will be an enormous payoff if, together, we can succeed; the cost in lost data of not moving forward is incalculable.

### **Acknowledgments**

First, I must thank the US National Science Foundation for its support of this research through a grant entitled “Cyberinfrastructure for Archaeological Data Integration Enabling the Study of Long-Term Human & Social Dynamics (SBE 0433959).” For all their hard work, I am extraordinarily grateful to my co-principal investigators on that project, John M. Anderies, Chitta R. Baral, K. Selçuk Candan, Hasan Davulcu, Michelle Hegmon, Subbarao Kambhampati, Ann Kinzig, Huan Liu, Peter H. McCartney, Ben Nelson, Margaret C. Nelson, Charles L. Redman, Arleyn W. Simon, Katherine A. Spielmann, and Sander van der Leeuw. Of this group, K. Selçuk Candan, Hasan Davulcu, Margaret C. Nelson, Subbarao Kambhampati, and Katherine A. Spielmann have been instrumental in our most recent implementation efforts as have our graduate assistants, Tiffany Clark, Yan Qi, and Toufeeq Ahmed. The members of our Faunal Working Group: Jonathan Driver, Donald Grayson, Elizabeth Reitz, and Christine Szuter generously provided us with invaluable insight and guidance. The participants in our Santa Barbara Workshop donated their time and their idea to help shape this important vision for the discipline (their names are listed in Kintigh 2006). Finally, ASU’s Global Institute of Sustainability has provided marvellous logistical support for the project.

### **References Cited**

KINTIGH, K. W. (2005) — The Promise and Challenge of Archaeological Data Integration. *Anthropology News* 46:7, p. 16.

KINTIGH, K.W. (Ed.) (2006) — The Promise and Challenge of Archaeological Data Integration. *American Antiquity* 71:3 567-578

SNOW, D.R.; GAHEGAN, M.; GILES, C.L.; HIRTH, K.G.; MILNER, G.R.; MITRA, P; WANG, J.Z. (2006) — Cybertools and Archaeology. *Science* 311, p. 958-959.