# A community approach to data integration: Authorship and building meaningful links across diverse archaeological data sets

**Eric Kansa**
*The Alexandria Archive Institute, 125 El Verano Way, San Francisco, California 94127, USA*

## ABSTRACT

**The ability to link and compare diverse archaeological data sets will catalyze innovative research of great scope and analytic rigor. However, information heterogeneity and limited budgets and information technology skills challenge data dissemination initiatives. This paper argues for new methods of community-based data integration pioneered by the University of Chicago's Extensible Markup Language (XML) System for Textual and Archaeological Research project (XSTAR). With XSTAR, data integration takes place in two steps: (1) syntactic-schematic integration: Legacy data sets are migrated for representation in the data structures described by the Archaeological Markup Language (ArchaeoML), and (2) Semantic integration: Mappings must be established between related terms and classes in each source database. Because the nuances of meaning are often very subtle, human experts must classify related items in each data set. Initial syntactic-schematic mapping of data into XSTAR is simple and fast but occurs at a relatively abstract level of meaning. Nevertheless, this initial step can accommodate diverse archaeological (and other) data sets, and will facilitate community-led development of more semantically specific data integration. XSTAR hopes to enable multiple semantic data integration schemas to develop and keep pace with changing research agendas. though rooted in archaeology, this paper discusses challenges faced by many disciplines in encouraging more powerful diachronic and regional syntheses. ArchaeoML's highly generalized data model has applicability outside archaeology, especially with subdisciplines of the earth sciences yet to develop formal ontologies. In addition, because this is a community driven approach, incentives for community participation must be explored. Intellectual property and professional rewards are key factors in determining the success of online dissemination systems across many disciplines.**

**Keywords: archaeology, databases, digital archives, data models, intellectual property.**

## INTRODUCTION

The construction of integrated digital collections of world historical and archaeological information stands as an important goal with wide-ranging research, instructional, and cultural heritage stewardship applications. Thus far, the Internet and related technologies have played an important role in informal scholarly communication, project planning, and in some limited promotional efforts. However, the Internet has yet to play a significant role in the archaeological publication process. In this regard, archaeological data dissemination lags behind other disciplines (such as the geological sciences).

However, data dissemination in archaeology faces many unique challenges that require the development of special technical and institutional frameworks. Archaeology differs from many other field sciences in that many archaeologists do not see their research as science (Trigger, 1989). The discipline developed and continues to evolve at the intersection of the natural sciences (geology, zoology, botany, and other environmental sciences), humanities, and social sciences (anthropology, history, architecture, classics, and other area studies, etc.). This array of disciplinary perspectives encourages a diversity of research methods, agendas, and theoretical approaches to the study of human material culture. Data, evidence, interpretations, and syntheses all have very different roles across this widely varying community. Such heterogeneity (not to mention uneven levels of technical expertise and funding) makes the construction of a digital dissemination infrastructure for this discipline a great challenge. The goal of data integration must take into account the highly divergent nature of archaeological data.

To accommodate archaeology's diversity, this paper advocates a community driven iterative approach to data integration. Because archaeology has so few standards, creating mappings across data sets is often not straightforward or easily automated. Such mappings should be regarded as contingent and open for revision. Encouraging data sharing and efforts at data integration requires that individual researcher participation incentives be recognized as critical aspects of the system's architecture. Therefore, intellectual property and professional concerns must be addressed. Finally, the relevance of this discussion is not limited to archaeology. Many new and emerging areas of study see wide experimentation with research methods, theoretical approaches, and terminological systems. As is the case with archaeology, sharing information where there is little or no standardization represents an important challenge. In addition, many areas of investigation lay at the margins of funding priorities and infrastructural support. Cost-effective and technically feasible solutions are a must, since archaeology, like many other areas of the field and environmental sciences, lacks the funding and organizational infrastructures needed to support the kinds of solutions so successfully demonstrated by the geosciences, genetics, neurosciences, etc. Therefore, it is hoped that this discussion of archaeology's struggles with digital dissemination can provide some useful experience for other communities faced with similar challenges.

## INFORMATION ACCESS AND PRESERVATION

The problem of data integration is part of larger concerns in archaeological communi-

*Geosphere*; October 2005; v. 1; no. 2; p. 97–109; doi: 10.1130/GES00013.1; 6 figures.

cation. It is widely recognized that we face tremendous problems in the publication, dissemination, and preservation of archaeological research (Denning, 2003; Gaffney and Exon, 1999; Richards, 2001). There has been a general trend in the history of our discipline toward more detailed and comprehensive levels of documentation. Great efforts are now used to recover and record animal bones, seeds, micromorphological samples, lithic debitage, and other types of evidence that earlier generations of archaeologists ignored and did not value. Stratigraphic recording, intensive intrasite and regional survey techniques, photography (and video), geographic information systems (GIS), and computer aided design (CAD) all add to the richness and comprehensiveness of contemporary archaeological records. While this information is never wholly complete or objective, it presumably has some value (Richards, 2003). The Council for British Archaeology's Publication of Archaeological Projects: A User Needs Survey (PUNS) further explores these issues with a survey of professional archaeologists in the United Kingdom (Jones et al., 2003; Jones et al., 2001). The level of documentation now demanded by our discipline impedes the function of the publication process as a tool to preserve the archaeological record. Information management, coordination of specialists, editing, compiling, and filtering the masses of documentation, are often expensive and tedious tasks. In practice, the complete corpus of observations and documentation rarely sees full publication. Jones et al. note (Jones et al., 2003, http://www.britarch.ac.uk/pubs/puns/punsrep6.html): ''There is widespread suspicion that decisions on what should or should not go into print are being shaped less by scholarly principle than by financial expediency. . . Many specialists now take it for granted that much of their material will not be published.'' Because high-level abstractions are easier to share and win more professional rewards, researchers have more incentive to produce such distillations instead of their primary documentation. Where the detailed observations have relevance to a specific research question, argument, or point of interpretation, they are likely to be published. Through publication and dissemination such observations have some hope of long-term preservation. The rest of the observational corpus remains in the hands of individual researchers, either slowly prepared for more complete publication, or deteriorating as the researcher's interests and priorities change. Without active preservation efforts, this digital documentation can quickly become unusable and forever lost (Condron et al., 1999; Eiteljorg, 1997; McCartney et al., 2000; Richards, 1997).

## DISSEMINATION AND INTERPRETATION

Scholarship is better served if claims about the past can be evaluated in terms of appropriate use of evidence to support arguments and interpretations. Without dissemination of that large information resource, it is impossible to support challenging interpretations that have authoritative support from primary documentation (Gaffney and Exon, 1999). Frustration with the current state of publication is highest among specialists and researchers in the archaeological sciences, where the links between primary observations and interpretive claims are especially clear (see Jones et al., 2003, their Fig. 14). Without effective dissemination, claims and counter claims become much less meaningful, since all refer to essentially hidden, inaccessible, and controlled evidence.

Incomplete publication limits the value of context as an interpretive resource. We see meaning and value in the vast and complex web of spatial and stratigraphic associations that intertwine the various components of the archaeological record. Yet, meaning suffers when observations and primary interpretations that provide such context are missing. Essentially, the incomplete dissemination of primary evidence prunes away (often drastically) this complex array of contextual associations. Context, however, extends beyond the purview of a single archaeological study. The meaning of archaeological observations is also constructed by how these observations relate to the larger body of scholarship. Yet, here too, the limitations of current publication practice impede contextual understanding. Data sources from multiple studies are scattered, fragmented, inaccessible, and thereby extremely difficult to integrate (Grayson and Cannon, 1999; McCartney et al., 2000; Nelson et al., 1994; Richards, 1997; Robinson, 2000). PUNS shows that this is a general concern in archaeology, and 79% of respondents desired measures to encourage integration of research from different field projects (Jones et al., 2003).

Syntheses and data integration are also difficult because access is only one facet of the usability of information. ''Gray literature'' earns its dubious distinction not only because these reports (cultural resource management studies, dissertations, and unpublished drawings, maps, and field notes) have limited circulation, but also because their sheer quantity presents formidable needle-in-a-stack-of-needles problems. For instance, professional researchers generate over 790 archaeological reports a year just in the state of Oregon (Gilsen, 2001). Few can keep pace with so many reports, most of which are never archived in libraries and therefore see little of the indexing and cataloging needed for retrieval. The PUNS report clearly demonstrates that this is a great concern for many archaeologists, since over 71% of respondents expressed worry that new and relevant information was being produced of which they had no knowledge (Jones et al., 2003). Such gray literature represents an underutilized and vulnerable resource for scholarship. Forty percent of archaeologists responding to PUNS indicated that they rarely (if ever) use gray literature (Jones et al., 2003). Jones et al. (2003) note that university researchers tend to draw on openly published field research, while underutilizing the gray literature that circulates primarily among government officials and archaeological contractors. Thus, the research community that attempts to develop and expand general understanding of the past does so with a limited sample of archaeological literature. Since the overall corpus of gray literature is poorly known by anyone, assessing the archival status and preservation risks of this corpus is difficult. However, by virtue of its limited distribution, gray literature is at risk for loss and destruction—undermining the very intent behind cultural and historical preservation legislation.

## DATABASES AND COMMUNICATING ARCHAEOLOGY

At the outset, archaeologists must recognize that much archaeological knowledge extends beyond written narratives and is encoded in databases of various types. For instance, as GIS becomes an increasingly important aspect of archaeological research, it is receiving much more critical theoretical attention and theoretically informed application (Aldenderfer and Maschner, 1996; Lock and Harris, 2000). Because of their importance as recording and organization tools, understanding databases and strategies to communicate databases has great theoretical significance. Theoretical understanding of the role of these media, including databases, must keep pace with technology if we are to maximize their interpretive potential as we use and communicate these new types of documentation (Denning, 2003; Hodder, 1999). By recognizing the diversity of archaeological meaning, we can limit false starts that apply new technology in flashy, yet shallow, ways.

A database is a model or representation of observed and interpreted reality. Such models help to organize and guide interpretations. Constraints imposed by the data schema can constrain the observational and interpretive process. Researchers will tend to avoid making certain types of observations or interpretations if such observations cannot be accommodated within an existing database design. A poor database design can exclude unexpected observations and limit interpretive choices. In other words, poor database design may distort an observed picture of a very heterogeneous reality by locking recording choices into rigid and standardized entries in data tables. Most archaeological databases use familiar tables of rows and columns of data in class-based data schema (Schloen, 2001). Each record represents a single instance of items belonging to a given class of observation. Typically excavation deposits (a class of observations) are recorded on a table. Other tables may be used to record observations on pottery, others on lithics, and so on. Each class has a table with a fixed number of columns of predetermined variables (e.g., soil type, pottery fabric, taxon, etc.) used to describe items in that class. A class-based approach is not necessarily always unadvised. It is easy to develop data entry forms that help guide excavators (some with little experience) to make standard observations. A data schema built around such classes is typically adequate for individual researchers and their own observation recording and analysis needs, although difficulties can be encountered with unexpected finds that do not adhere well to the classes defined in the schema.

Database design is therefore an integral aspect of archaeological methodology, since design choices both reflect and help determine the interpretive process. The theoretical choices and assumptions that guide database design have profound interpretive implications for one research project (Ryan, 2004). Such design choices are also critical factors in the interpretive possibilities of online systems of data sharing. Terminologies, recording strategies, and artifact typologies are all subject to revision and are predicated on diverse research interests and theoretical perspectives. This heterogeneity is common to databases generated in the social sciences and humanities. The authors, research methods, uses, and audiences of social science data sets are all highly diverse. Such heterogeneity makes database integration a theoretically informed interpretive process (Paterson, 2003). That is why we, as a discipline, should carefully develop and apply standards while working toward online data dissemination. Simply put, inappropriate application of standards can limit interpretive possibilities and overly constrain research design. When developing standards for archaeological (and other) information architectures, we should aim to maximize interpretive possibilities and limit constraints on methodological, terminological, and theoretical choices.

## EXPERIMENTAL METHODS IN ARCHAEOLOGICAL DISSEMINATION

A key challenge in digital dissemination for archaeology and other disciplines is the delivery of highly diverse data sets (with no standard taxonomies, vocabularies, or structures) while enabling analytic queries and specific searches. Such delivery mechanisms need to be scalable to accommodate new data sets and new institutional participants. Scalable Web databases not only enable powerful searches and retrieval of specific items of content, but they can also encourage the integration and synthesis of data generated by disparate archaeological projects. To achieve integration, scalable Web databases require the development of global schemas where content can be mapped from the local schemas of data sets generated by different researchers and different projects. Such global schemas can vary in terms of their specificity. They may include constrained common vocabularies (such as the common typologies and recording systems required by some government heritage organizations). Alternatively, global schemas may be highly abstract and generalized to represent widely varying and diverse collections of archaeological data, but support less specific semantic mappings.

Data integration through use of global schemas can be achieved with two basic strategies: (1) data mediation, and (2) data warehousing. Data mediation strategies join several individual databases together through software translation tools called ‘‘wrappers’’. These wrappers essentially translate between the local vocabularies and structures of each individual database and a more generic global schema. Users can query a system running the generic global schema, and this query gets passed on and automatically translated to retrieve information from various individual databases. With the data warehousing approach, the translation from local to global schemas happens by directly importing individual project data sets into one larger database system. Both of these approaches can be combined, and data mediation and data warehousing systems can be integrated.

The most significant hurdle for use of scalable Web databases as publication tools comes from the highly diverse nature of archaeological knowledge. Archaeology uses a vast and ever-changing array of recording systems, all based on diverse theoretical perspectives and methods. Typologies and nomenclatures are frequently contested and vary among specialist communities and individual researchers. This makes the development of global schema to represent archaeological information critically important (Zhang et al., 2002). One temptation would be to create rigid class-based standards of recording, data organization, vocabularies, typologies, etc. and to compel researchers to follow such standards. This approach would indeed facilitate data integration (Moon, 1993). The technical aspects of defining such standards would be relatively easy. Agreeing to the standards, however, would be much more difficult (Kilbride, 2005). And enforcing the standards would probably be even more difficult, akin to herding cats who each have their own interests, research questions, and agendas. Given that new questions and perspectives continually emerge in this evolving discipline, attempting to lock in rigid standards is probably unworkable and unadvised. Moreover, how would one integrate legacy information generated before such standards became enforced? Would this information be adapted into the new standards framework? Would there be universal agreement on such adaptations?

## GLOBAL SCHEMAS FOR ARCHAEOLOGY?

There are several digital dissemination efforts in archaeology, including research driven initiatives and cultural heritage preservation programs. We are now witnessing the first attempts to build Web databases that explicitly embrace the diversity of archaeology and its divergent research agendas. My purpose here is not to discuss each individual project, but to present a few to highlight significant points.

## DATA MEDIATION INITIATIVES

There are several archaeological data mediation efforts currently in development and in use. The Digital Archive Network for Anthropology and World Heritage (DANA-WH) (www.dana-wh.net), a project led by Jeffrey Clarke, from the University of North Dakota, has pioneered these data mediation techniques in the United States. It has attracted some participation, but because of funding limitations, limited organizational support, and constraining data schema requirements, DANA-WH has not grown into a widely used resource. Despite these limitations, DANA-WH is online and serving data that would otherwise likely see little dissemination or preservation. It has also provided a valuable basis of experience that benefits other endeavors.

Following DANA-WH, another leading data mediation project is the Electronic Tools and Ancient Near Eastern Archive Digital Library (ETANA-DL) (http://feathers.dlib.vt.edu/). Sponsored by the National Science Foundation and the ETANA consortium, ETANA-DL benefits from the guidance of James Flanagan, an archaeologist specializing in the Southern Levant, and Edward Fox, a computer scientist and digital library theoretician. The ETANA-DL project attempts to apply Fox's conceptually abstract and generalized 5S framework to develop user services, data structures, and management tools that link together a series of archaeological projects from the Southern Levant (Fox et al., 2001). Because this project is in its early stages, the global schema used by ETANA to create mappings that link different project databases together is still in development (Flanagan et al., 2004). In this approach, individual researchers have the flexibility to design their own databases with information structured according to their own individual research agendas. The database wrappers essentially eliminate the need for users to query each project's idiosyncratic database separately. The database wrappers enable users to benefit from integrated Web access to several individual databases mapped into ETANA's global schema.

The ETANA-DL project is significant and may prove to be an important avenue for building scalable Web databases. One concern over this approach regards the global schema developed by ETANA. The project thus far attempts to integrate several databases of projects exploring broadly similar research questions (Southern Levantine Bronze and Iron Age settlements). All of these projects stem from a common disciplinary history, and all use similar methodologies and recording practices. These individual project data sets therefore represent only a small portion of the range of archaeological research questions, methods, and recording systems developed across the whole discipline. Individual local data schemas developed by archaeologists will vary much more widely than the data sets currently available to ETANA. ETANA researchers acknowledge that their global schema may need revision if an individual project data set cannot be accommodated. However, it is unclear how decisions about global schema revisions would be determined.

As with changes to the global schema, the development and modification of wrappers around each individual ETANA data set remain the domain of technical specialists (ideally working in conjunction with archaeologists) (Ravindranathan, 2004). Information technology specialists must still intervene, and for this reason, debate and experimentation with data mappings will probably be limited. Domain experts need the ability to author, debate, and revise data mappings since some mappings will be contested. For instance, many archaeological databases have fields to record the length, width, and mass of finds, and such fields can be easily related without much controversy. However, many archaeological databases will have fields with project-specific meanings, relating to individual research approaches and agendas. For instance, some database fields may relate to very abstract, and highly contestable concepts. One can classify certain excavation contexts as ritual spaces, domestic spaces, public spaces, private spaces, activity areas, and so on. It is unlikely that mapping meaning between these abstract and highly theoretical concepts can ever be definitive.

The International Committee for Documentation of the International Council of Museums (CIDOC), a European led initiative, uses a similar data mediation approach to data integration. CIDOC's global schema (called a conceptual ontology or conceptual reference model) is based on approaches stemming from cognitive science and semantic web research (Gruber, 1993; Hendler, 2001). A conceptual ontology is often developed within a specific domain (or discipline) to help support the integration of diverse and distributed databases within that discipline. Therefore, a conceptual ontology must be sufficiently generalized to represent the meaning of data sets produced within a specific domain. Here many archaeologists would argue that there is not one but many ''archaeologies,'' each with their own domains of meaning. Nevertheless, CIDOC developers claim that their contextual reference model is generalized enough to be a global schema for archaeology (Doerr, 2003).

CIDOC, whose goal is to make museum data semantically interoperable, has gradually emerged through several years of discussion among museum staff, some archaeologists, and computer scientists. It has great institutional support, especially in the European Union, and is slated to become an ISO standard (Cripps et al., 2004). Terminologies and data structures local to each museum database system are mapped to the general CIDOC scheme. Because of CIDOC's complexity and the limited availability of CIDOC authoring tools, these mappings are conducted by information technology specialists with specific CIDOC expertise. Thus far, CIDOC sees most of its application in large institutional settings (especially museums and digital archives), and has seen little application as a solution for individual researchers.

The CIDOC framework provides a common and explicitly defined way of expressing historical knowledge. It acts as a conceptual bridge that integrates each of the local museum databases with one another. Whether or not meaning is lost or distorted in translation to CIDOC's scheme is open to theoretical discussion. Can one conceptual scheme, such as CIDOC, adequately express historical and archaeological meanings and interpretations stemming from diverse research traditions and agendas? Can CIDOC really express contested theoretic constructs such as place, rank, and culture, which may see multiple forms of expression in different databases? Mapping such meaning into CIDOC's framework may be contestable and difficult to establish. Some concepts may have no possibility of CIDOC representation at all, because they stem from theoretical perspectives far removed from the perspectives of CIDOC's creators. In any event, CIDOC is new and because it has primarily seen implementation in institutional settings, it has yet to receive critical review from practicing field archaeologists. Nevertheless, as CIDOC sees wider application it will see more and more thoughtful criticism. Schloen (D. Schloen, 2004, personal commun.) has voiced some basic concerns on the CIDOC approach, which uses a global schema that is on one hand too abstract for easy comprehension among the research community, and on the other hand has some 120+ relationship types that may be too over-specified and hard-coded.

ETANA and CIDOC can and do play an important and powerful role in integrating archaeological and cultural heritage knowledge. One of CIDOC's creators, Martin Doerr (Doerr, 2003) explicitly calls for much more theoretical examination of CIDOC and its concepts. This discussion merely raises issues that should be more fully debated by the community. Any implementation of a given data architecture often involves a significant financial investment. Asking theoretically informed questions about such architectures is essential to maximizing their ability to support innovative scholarship. In addition, there are financial and organizational issues to consider.

It remains to be seen if ETANA will see any support beyond its initial NSF funding, or if it will continue to grow and gain an active user community. On the other hand, CIDOC has seen a great deal of institutional commitment from museums and heritage organizations in the European Union, but the technical and conceptual difficulties in using CIDOC have so far hampered its growth as a data dissemination tool for individual field researchers.

## THE XSTAR DATA WAREHOUSING APPROACH

The funding, organizational, and technical constraints of archaeology require approaches for data dissemination and integration that are technically feasible and cost-effective. In many cases, a data warehousing approach may be more appropriate than data mediation strategies since archaeologists often lack the technical skills and facilities needed to maintain their own database servers or build data wrappers required for mediation. A data warehouse acts as a somewhat more centralized repository system to deliver data to users and accept new data submissions. With this approach, individual researchers and small research groups no longer need to support their own database servers, but mapping to a global schema is still required. Similarly, implementation of a global schema appropriate for archaeology must reflect the realties of limited funding and technical resources. In this case, mapping to such a global schema must be simple enough for individual researchers to apply, since there are few financial resources to support dedicated technical experts. As demonstrated by the CIDOC example, individual researchers create many local data sets but typically lack the time, inclination, or skills to perform data mapping themselves. Though powerful, the elaborateness and specificity of the CIDOC ontology makes it too difficult for individuals to perform their own mappings. In some ways, the sophistication of the CIDOC ontology may be too much of a good thing for widespread implementation among the community.

An alternative approach attempts to circumvent these constraints. David Schloen of the University of Chicago leads a data warehousing initiative called the Extensible Markup Language (XML) System for Textual and Archaeological Research project (XSTAR). Schloen (2001) implemented XML to describe the Archaeological Markup Language (ArchaeoML), which is intended to be a global schema for archaeology and philology (see www.oi.uchicago.edu/OI/PROJ/XSTAR/ArchaeoML.html). ArchaeoML provides a common framework for expressing archaeological observations, their descriptive properties, and their contextual relationships. However, its inherently flexible item-based structure insures that the content of that information is not predetermined (Schloen, 2001). ArchaeoML's key features include:

1. **Flexibility in Scale:** An ArchaeoML item can be any type of archaeological observation at any scale, ranging from a region, to a site, to a specific deposit, to an artifact, ecofact, or even microscopic observation. Each item has its own unique label (site name, context ID, bone ID, etc.) created at the discretion of the researcher.
2. **Flexibility in Description:** Similarly, the names, terminologies, and values of the descriptive properties of each item are also created at the discretion of the individual researcher. For instance, one is free to describe the composition of pottery with a property like fabric, ware type, or any other set of variables. In other words, descriptive variables and terminologies are left to the researcher's discretion, and are not hard-coded into the data structure. Multiple media, including video, images, GIS, etc., also can be used in addition to alphanumeric text to describe specific items. Figure 1 represents how ArchaeoML enables such flexible description of spatial items (locations and objects).[1]
3. **Accommodates Heterogeneity:** New descriptive variables can be tailor-made for a specific unit without changing the descriptive framework for a whole class of finds. Researchers can create new observational criteria and descriptive properties very easily if they encounter unexpected or unique items.
4. **Multiple Observations and Observers:** ArchaeoML easily represents multiple observations (even contradictory observations) made on a single item. Each observation can be authored individually, thus explicating much of the process of knowledge construction.
5. **Expresses Contextual Relationships:** Extrinsic contextual relationships organize the mass of individual items into archaeologically meaningful structures. These relationships include spatial hierarchies (some items contain smaller items, which contain even smaller items), stratigraphic relationships of sequences of deposition (shown graphically in Harris matrices), and relationships of spatial adjacency. These archaeologically meaningful structures (many of which are recursive) provide the framework that guides searches and analytically powerful queries (see Figures 2 and 3).[2] Users will also have the option to declare their own customized types of relationships.

Schloen claims that ArchaeoML is sufficiently generalized to represent the diversity of archaeological knowledge. ArchaeoML is conceptualized as a very generalized, item-based model, where individual atomic units of observation are related to each other and their descriptive attributes. Each item does not belong to a predetermined observational class (pottery, bone, deposit, grave good, etc.). Users can group items into multiple and overlapping classifications defined according to their changing interests and assumptions.

### Syntactic and Schematic Integration

Because ArchaeoML data structures are highly abstracted and generalized, mapping local schema into the ArchaeoML global schema is simplified. Mapping an individual data set involves classifying local schemas into a few highly abstracted schemas that make up ArchaeoML. These include:[3]

1. Locations or objects (a site, an excavation deposit, an artifact, a sample of an artifact, etc.);
2. People or organizations (authors, editors, observers, museums, sponsoring organizations);
3. Properties (variables and values that describe other items);
4. Resources (digital media objects, such as images, video, etc.); and
5. Relationships (links between items, such as stratigraphic relationships).

The majority of archaeological data sets will map very readily to this simple set of schemas. Because of this relative simplicity, we hope to encourage more widespread adoption throughout the individual research community. Easy to use tools are essential for widespread adoption. Our development of

[1] This figure is a simplified representation of spatial units, for its complete formal schema, please refer to http://oi.uchicago.edu/OI/PROJ/XSTAR/spatialunit/SpatialUnit.html.

[2] The relationship-document type describes how XSTAR represents contextual relationships between different items (especially spatial items). Graphs (including networks and Harris matrices) are represented as trees of related Relationship documents. The complete formal schema is at http://oi.uchicago.edu/OI/PROJ/XSTAR/relationship/Relationship.html.

[3] ArchaeoML has several other schema, some with more specialized application for philology, defined in various XML document types. Most archaeological data sets can be mapped into these five listed schema. Full schema definitions can be found at http://oi.uchicago.edu/OI/PROJ/XSTAR/DocumentTypes.html.

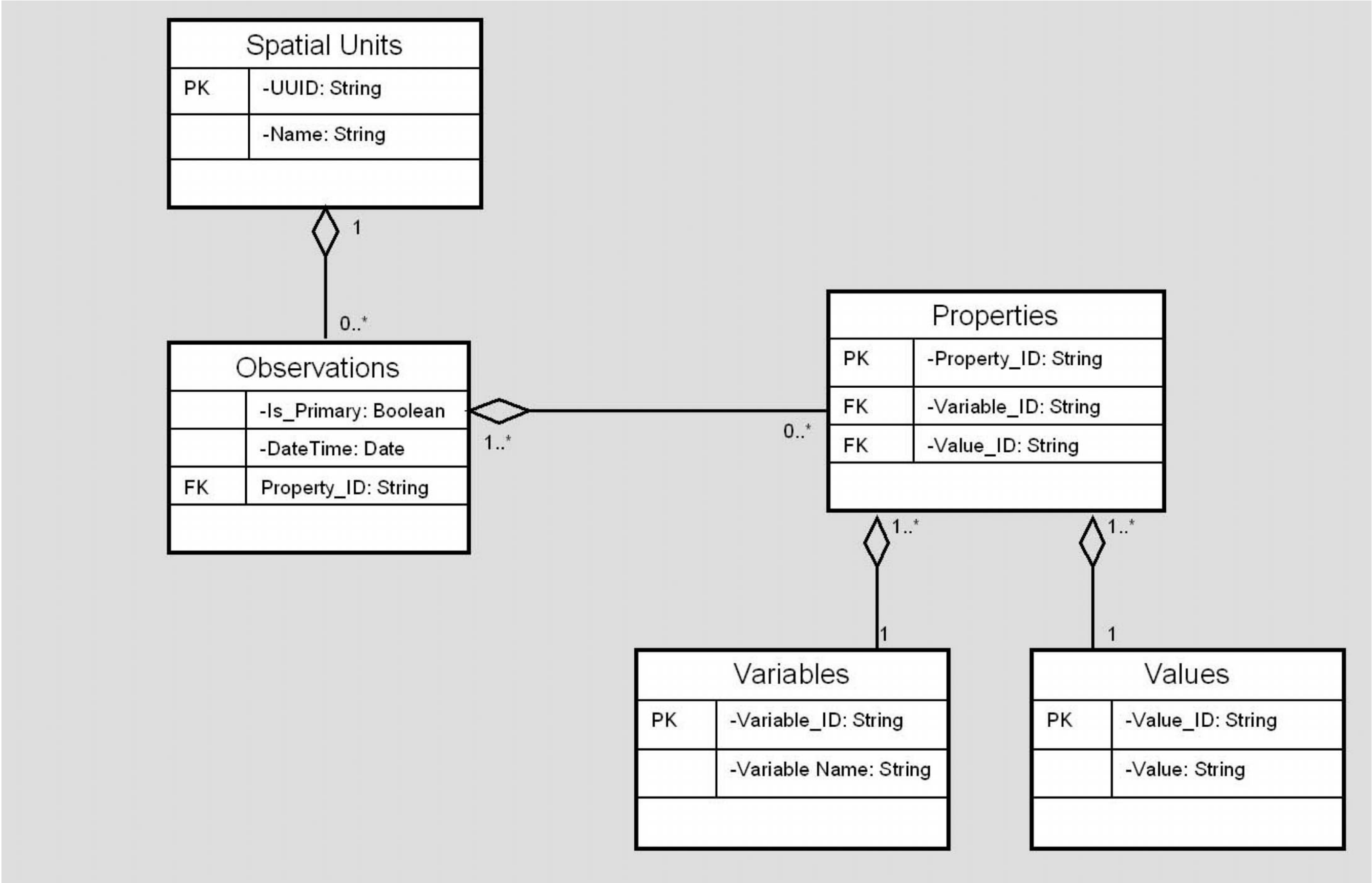


**Figure 1. Unified-modeling-language diagram of Archaeological Markup Language (ArchaeoML) spatial units, observations, and descriptive properties.**

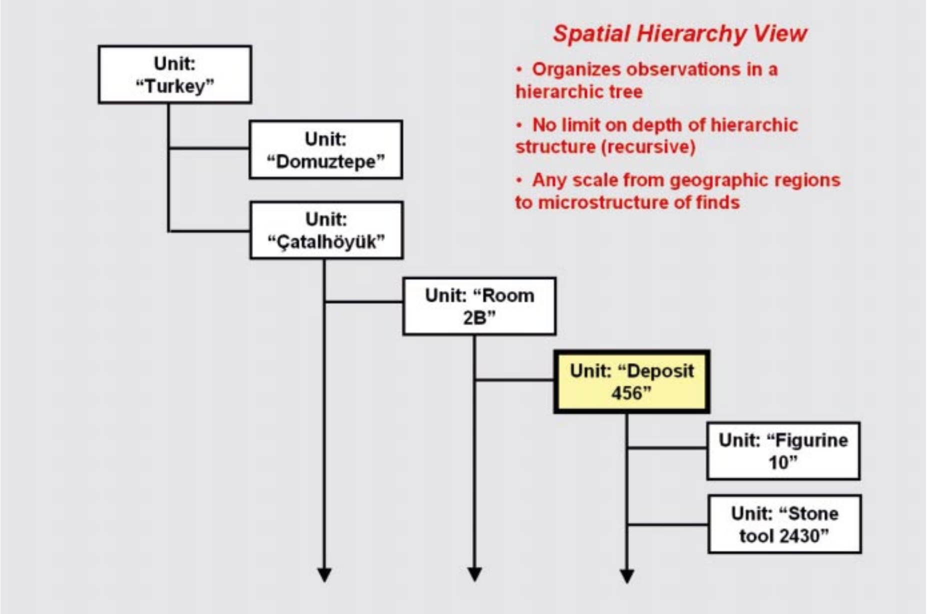

Spatial Hierarchy View
• Organizes observations in a hierarchic tree
• No limit on depth of hierarchic structure (recursive)
• Any scale from geographic regions to microstructure of finds
Unit: "Turkey"
Unit: "Domuztepe"
Unit: "Çatalhöyük"
Unit: "Room 2B"
Unit: "Deposit 456"
Unit: "Figurine 10"
Unit: "Stone tool 2430"

**Figure 2. Illustration of recursive Archaeological Markup Language (ArchaeoML) spatial hierarchies.**

data mapping tools assumes that most individual researchers will have developed their own idiosyncratic data schemas implemented in tables within relational databases. We are currently developing a java application that guides individual data contributors through a step-by-step process to classify each field in their legacy data table according to the above schema (see Fig. 4 and Fig. 5). They also describe how the various fields of their data table relate with one another and previously imported data. For instance, the application asks users to note which field contains records of items described by fields classified as properties. In another example, users can use the application to define relationships among items classified as locations and objects. Spatial containment relationships (e.g., ‘‘this field containing records of excavation deposits *contains* this field containing records of artifacts’’), stratigraphic relationships, and links between spatial items, image files, and people (as authors or observers) can all be defined with the Java application.

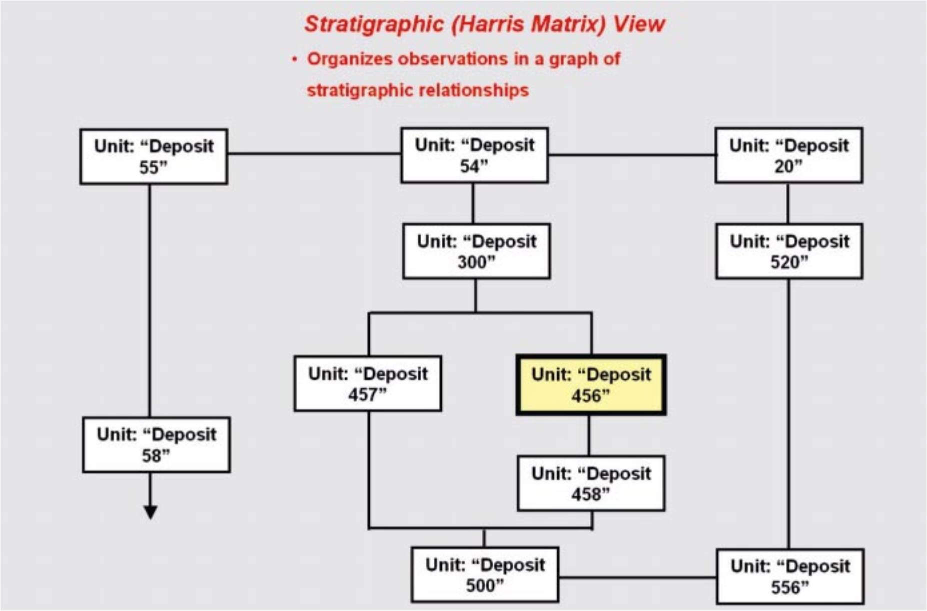
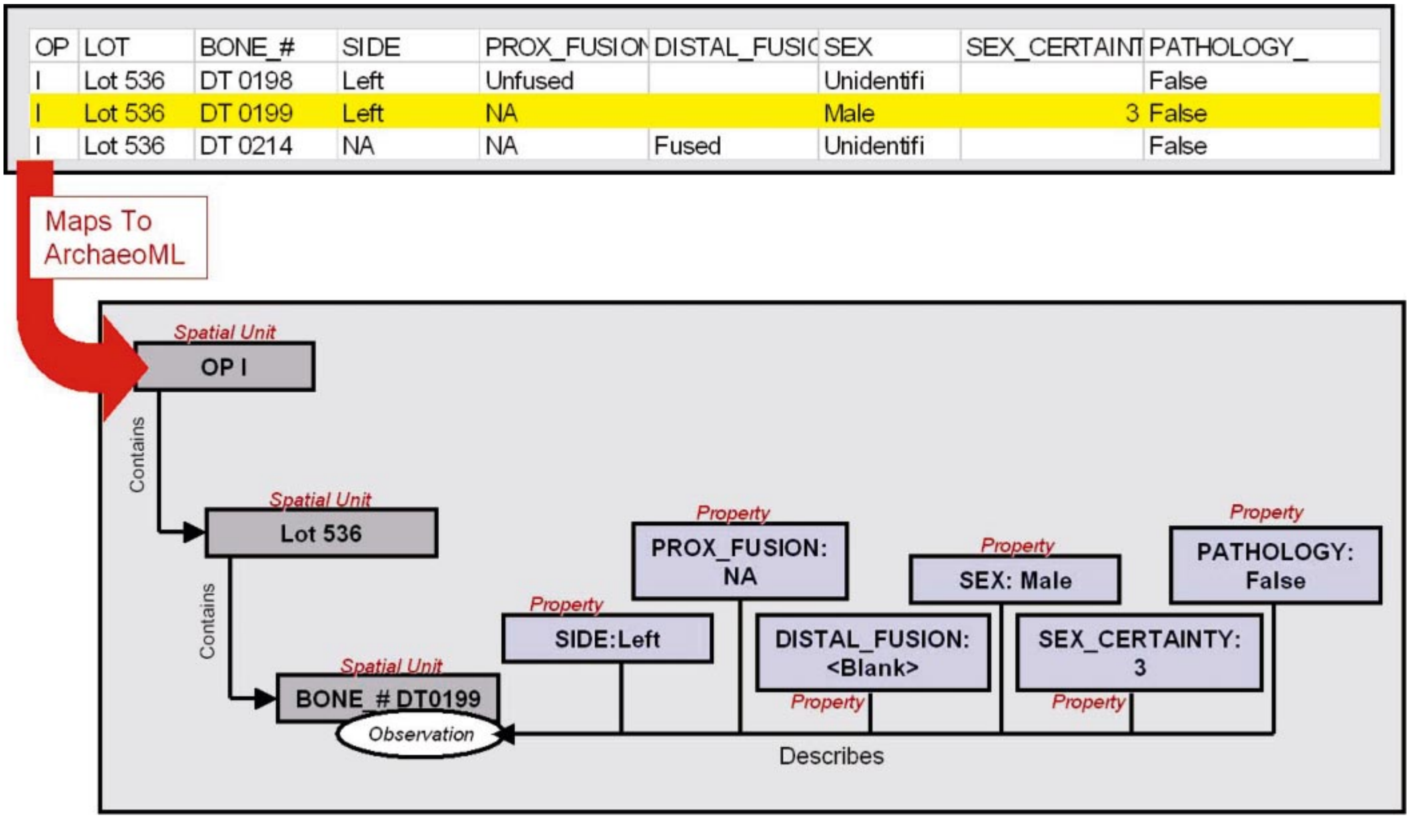


**Figure 3. Illustration of Archaeological Markup Language (ArchaeoML) stratigraphic relationships.**

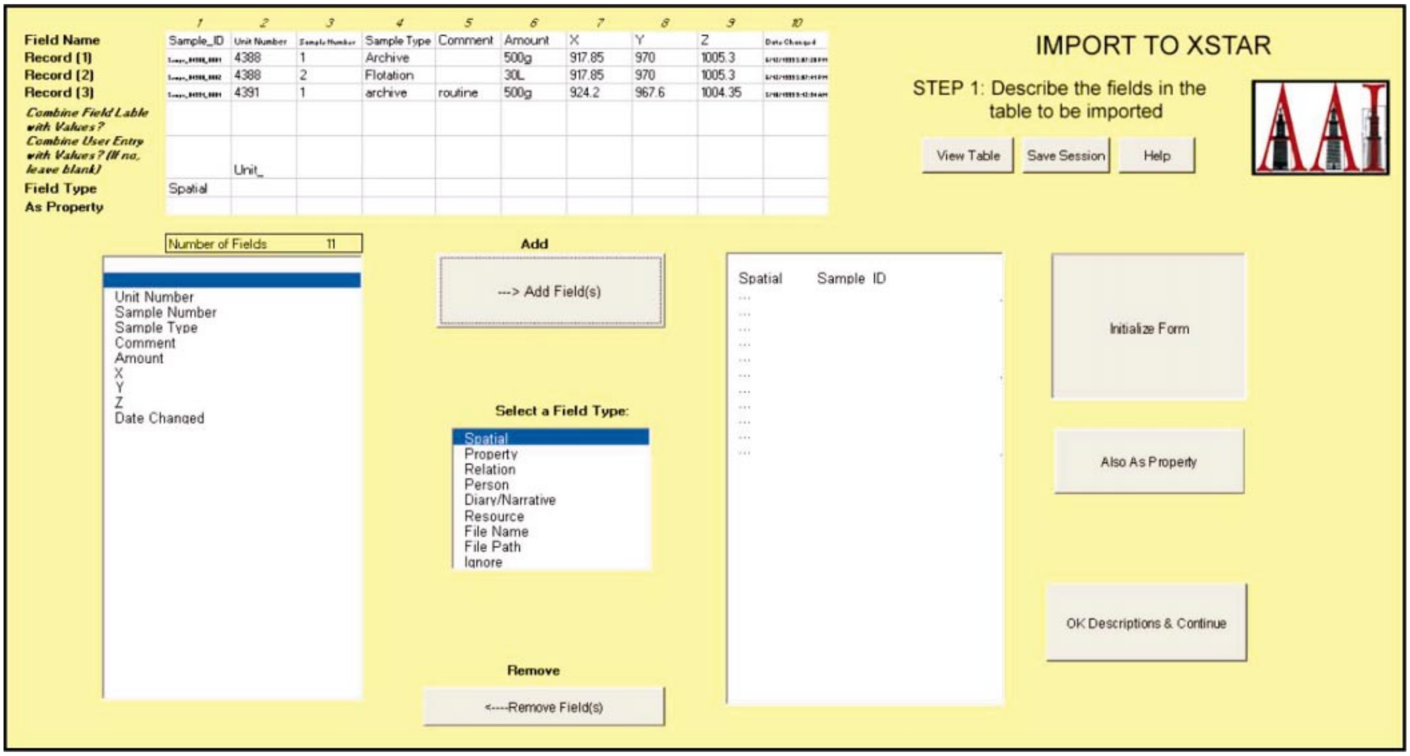
Since XSTAR is a data warehousing project, this java application is also an import device used to migrate legacy data sets into the XSTAR database. After the user maps a local data schema to ArchaeoML, the import tool then sends the imported data set to the XSTAR data warehouse in the form of a properly structured XML document. After importation, a researcher's original idiosyncratic vocabularies, observational variables, and values are all retained. Thus, while ArchaeoML is a standard, researchers retain full flexibility to develop and continually refine their methods, vocabularies, and recording systems. Because of this, the community would be more likely to accept the ArchaeoML standard than other standards (which prescribe certain terminological and recording systems). Finally, data can make a round-trip into and out of expression in ArchaeoML. In other words, a data table can be imported to ArchaeoML and exported back into a file identical to the original imported table. The ability for data to make this round-trip demonstrates that local meanings are retained in ArchaeoML expression.

The mapping tool we are creating works primarily on a syntactic and schematic basis. Data originating from diverse research proj-



| OP | LOT | BONE_# | SIDE | PROX_FUSIOI | DISTAL_FUSIC | SEX | SEX_CERTAINT | PATHOLOGY_ |
|---|---|---|---|---|---|---|---|---|
| I | Lot 536 | DT 0198 | Left | Unfused | | Unidentifi | | False |
| I | Lot 536 | DT 0199 | Left | NA | | Male | 3 | False |
| I | Lot 536 | DT 0214 | NA | NA | Fused | Unidentifi | | False |

**Figure 4. Example of mapping into Archaeological Markup Language (ArchaeoML).**



|  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Field Name | Sample_ID | Unit Number | Sample Number | Sample Type | Comment | Amount | X | Y | Z | Date Changed |
| Record (1) | [illegible] | 4388 | 1 | Archive |  | 500g | 917.85 | 970 | 1005.3 | [illegible] |
| Record (2) | [illegible] | 4388 | 2 | Flotation |  | 30L | 917.85 | 970 | 1005.3 | [illegible] |
| Record (3) | [illegible] | 4391 | 1 | archive | routine | 500g | 924.2 | 967.6 | 1004.35 | [illegible] |
| *Combine Field Lable with Values?* |  |  |  |  |  |  |  |  |  |  |
| *Combine User Entry with Values? (If no, leave blank)* |  | Unit_ |  |  |  |  |  |  |  |  |
| Field Type | Spatial |  |  |  |  |  |  |  |  |  |
| As Property |  |  |  |  |  |  |  |  |  |  |

Number of Fields 11

**Figure 5. Screenshot of data mapping–import tool.**

ects are migrated into XSTAR and are represented in a common schema (ArchaeoML). However, since ArchaeoML is highly abstracted, little data integration at the semantic level takes place. Individual terms that may exist in a legacy data set are not mapped into an elaborate ontology (that defines relationships of meaning such as: *jar* is an instance of *pottery*). Though migrating data sets into XSTAR does not immediately build highly specific semantic mappings, it does include mapping contextual relationships into the ArchaeoML global schema. Common representation of contextual relationships is sufficient to provide users with powerful browse, search, and some query capabilities across multiple data sets in the XSTAR system. A clear example is spatial containment, which describes recursive spatial hierarchies[4] between locations and objects. Each data set imported into XSTAR will have spatial containment relationships defined in the same way. This provides a powerful tool for navigation among observational units of varying scales as well as understanding contextual relationships of assemblages across multiple data sets. Finally, the XSTAR schemas can be used with other high-level conceptual mapping frameworks such as TimeMap (www.timemap.net) to provide enhanced spatial-temporal data integration (Johnson, 2004).

## Semantic Integration

Limiting data mapping to a high level of conceptual abstraction has some practical advantages. An experienced user requires ~15 min to migrate a table with 20 fields and 10,000 records into the ArchaeoML global schema, using the current version of the mapping tool (not yet released). Since these tables are often the product of several months of data collection, this is not an overly onerous process. In contrast, more semantically specific data mappings will require the investment of greater levels of time and effort. By lowering the bar to make data mappings relatively fast and easy and not requiring time-consuming semantic mappings, we hope to encourage more community participation in XSTAR. Since data sharing and data longevity are fundamental problems for archaeology, data dissemination methods must be as painless as possible.

Lowering the bar and only requiring high-level mappings for users to participate in the XSTAR system does not preclude more specific semantic mappings to develop over time. Because observational class structures are not hard-coded in ArchaeoML, data integration efforts can focus on drawing appropriate (as defined by individual researchers) links between individual items in different data sets. Items are described by specific variables and values, and such variables and values can be related across different data sets to create equivalences. Essentially, individual researchers must make decisions (with the option of sharing their decisions) about which individual items to include into classes that they develop. They must base their decisions on assessments of the meaning of various idiosyncratic variables and values that describe items from different data sets. They assign membership to classes based on criteria specific to each individual data set. The Relationship, Set, and Query document types support such semantic mappings, as do thesaurus functions.[5] Building such equivalences requires expert contextual knowledge, to know, for instance, that *fabric* is equivalent to *ware type* in a particular case, but not neces-

[4] See the Tree document type definition at http://oi.uchicago.edu/OI/PROJ/XSTAR/DocumentTypes.html.

[5] Full schema definitions can be found at http://oi.uchicago.edu/OI/PROJ/XSTAR/DocumentTypes.html. For an additional discussion of semantic mapping issues, follow the thesaurus subheading at http://oi.uchicago.edu/OI/PROJ/XSTAR/Integration.html.

sarily in other instances. In some cases, researchers may judge that data sets may be too incompatible to establish reasonable mappings. Any set of equivalences created to relate multiple data sets also would, and should, be highly contested. Enabling multiple semantic mapping schemes to be authored, evaluated, and revised avoids locking the community into a single interpretive structure, even if this framework was developed through consensus. Thus, by working toward personalized data mapping, the XSTAR project has similar goals as other digital research infrastructure projects, such as BIRN and GEON (Gupta et al., 2002; Lin and Ludaescher, 2004).

Thus, data integration is not always a problem that can see a final definitive answer, but can itself be a long-term area of research. Authoring links across individual data sets will take time, effort, and deep understanding of the data sets to be integrated. However, even a little data integration (at the syntactic and schematic level) and data pooling will help jump-start the process and facilitate the development of semantically richer data mappings. It is likely that such integration efforts will take place in the context of problem-oriented research. Instead of attempting to map together the entirety of multiple large data sets, many researchers will likely focus their efforts on mapping specific parts of these data sets to explore specific research questions. Exploring where highly specific semantic mappings will yield the highest returns in terms of science will be an important area of exploration. Initial semantic data integration efforts may focus on areas with relatively more community consensus in recording practices and methods, such as zooarchaeology, paleobotany, and human osteology. Potentially valuable scientific insights about regional trends in subsistence, health, and demography may be gained through such data integration efforts. Data integration issues may also present important research opportunities that will encourage more rigorous methods and recording practices to develop. The XSTAR approach does not exclude the possibility of standards or global classes of observations. Recording standards, shared vocabularies, and more semantically elaborated community ontologies (such as CIDOC) can still be developed, shared, and applied, but each standard need not be regarded as definitive. In the XSTAR system, multiple evolving standards can coexist. Thus, the same information architecture can allow and encourage standards to develop and keep pace with changing research agendas.

## Applications of XSTAR

Thus far, the XSTAR project has largely concerned itself with schema development and implementation (in a Tamino, native-XML database) and development of a custom user interface to browse, search, and query the database (implemented in Java, see screenshot, Fig. 6). Public demonstration of the XSTAR system is scheduled for October of 2005. This public demonstration will include a host of diverse pilot data sets either developed directly in the XSTAR environment or migrated into XSTAR. These projects range from philological initiatives to large-scale archaeological excavations of major Near Eastern sites. The Alexandria Archive Institute (AAI) is actively building a corpus of ArchaeoML data collected from projects outside of the University of Chicago. Building this corpus involves migrating diverse legacy data from research projects that come from Paleolithic archaeology, archaeological survey, classical archaeology, and Anatolian archaeology. Some of these data sets represent smaller specialist studies, and others represent large projects with contributions from several participants. These projects include Domuztepe, Çatalhöyük, and Monte Polizzo, all projects that employ different methods, theoretical frameworks, and recording standards. The application of ArchaeoML to the complete record of a variety of major excavations stands as an important way to evaluate the general applicability of the ArchaeoML global schema.

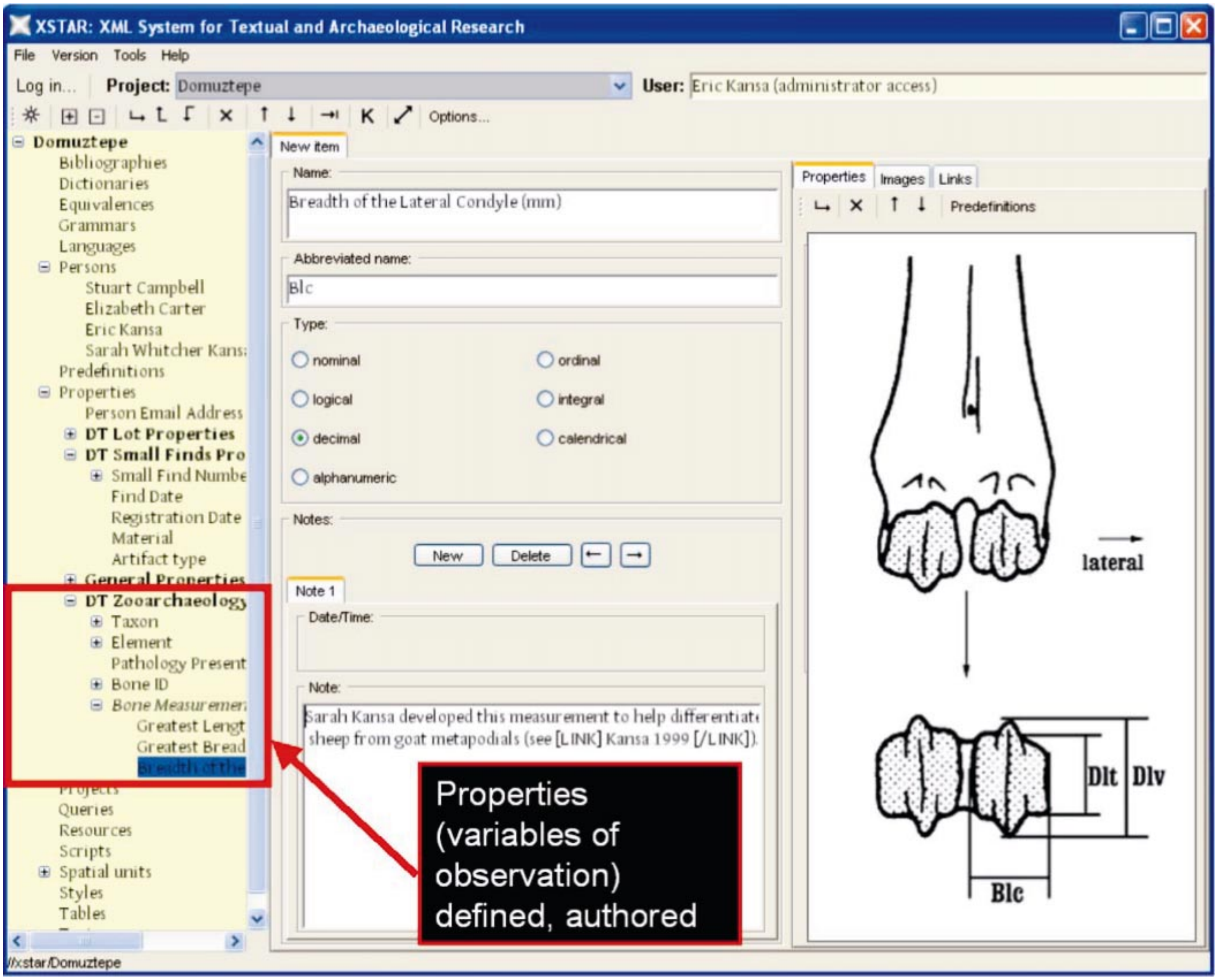
Finally, since ArchaeoML data structures are so highly abstracted, we are experimenting with migrating other related types of research content into the XSTAR system. Regional environmental data and environmental proxy data sources (data records from field ecology stations, some pollen analyses from lake-cores, zoological and some paleontological studies) often have little cyberinfrastructure support and few efficient dissemination channels. Even a very general level of data-mapping can add value to these data sets and facilitate multidisciplinary studies. A set of micromorphology databases generously provided by Paul Goldberg (Boston University) will soon be publicly available, offering a valuable demonstration of mapping geology-related data sets into the XSTAR system.

## Authorship and Intellectual Property: Creating Incentives for Online Dissemination

As discussed, because of the diverse nature of archaeological data sets, data integration should be seen as a contestable interpretive process. This requires the active participation of individual researchers, which in turn requires that these systems must work within archaeology's social infrastructure. Any solution must acknowledge these social realities or strong incentives will continue to inhibit the dissemination of large data sets. The tools required for online publishing and analyses must be intuitive and easy to learn. Many archaeologists recognize the failure of current dissemination methods and incentives. Estimates place the rate of primary data publication (of any level of comprehensiveness) at an abysmal 10%–30% (Ottaway, 2001).

Raw data often have rich and under-realized interpretive potential, and are often collected at great expense (see examples in genetics and the environmental sciences [Campbell et al., 2002; Helly et al., 1995] and examples in archaeology and museums [Gaffney and Exon, 1999; Jones et al., 2003]). An example from the environmental sciences helps to illustrate this point. In 1898, Hermon Bumpus published a landmark study on the evolutionary process of stabilizing selection by investigating mortality of house sparrows (Bumpus, 1898). Along with his conclusions, he comprehensively published his primary observations along with his theoretical interpretations. This set of raw data has proven to be tremendously valuable to later researchers, and has helped inspire the publication of at least eight (sometimes highly influential) peer-reviewed papers (Calhoun, 1947; Crespi and Bookstein, 1989; Grant, 1972; Harris, 1911; Johnston et al., 1972; Manly, 1976, 1985; O'Donald, 1973). If one measures the value of raw data by the number of publications it spawns, then sharing this set of raw data makes it at least ten times more valuable than it would have been without dissemination. This data set has even more value if we consider how useful it has proven for student instruction and exploration of real world data (Price, 1996). Without dissemination, data are very vulnerable to loss through overly restrictive intellectual property policies or simple neglect (Condron et al., 1999; Richards, 1997). An open-knowledge commons with freedom to build on, recombine, and re-evaluate research data will promote scholarship in a fundamental way.

In the United States, data, or facts, are not protected by copyright (though some other international jurisdictions, including the European Union, extend copyright over some data). Telephone white pages stand as an almost quintessential example of unprotected facts (Feist Publications, Inc., C. Rural Tel.



Project: Domuztepe

User: Eric Kansa (administrator access)

- Domuztepe
  - Bibliographies
  - Dictionaries
  - Equivalences
  - Grammars
  - Languages
  - Persons
    - Stuart Campbell
    - Elizabeth Carter
    - Eric Kansa
    - Sarah Whitcher Kans
  - Predefinitions
  - Properties
    - Person Email Address
    - DT Lot Properties
    - DT Small Finds Pro
      - Small Find Numbe
      - Find Date
      - Registration Date
      - Material
      - Artifact type
    - General Properties
    - DT Zooarchaeology
      - Taxon
      - Element
      - Pathology Present
      - Bone ID
      - Bone Measuremen
        - Greatest Lengt
        - Greatest Bread
        - Breadth of the
  - Projects
  - Queries
  - Resources
  - Scripts
  - Spatial units
  - Styles
  - Tables

Name: Breadth of the Lateral Condyle (mm)

Abbreviated name: Blc

Type:

○ nominal ○ ordinal

○ logical ○ integral

● decimal ○ calendrical

○ alphanumeric

Notes:

Date/Time:

Note: Sarah Kansa developed this measurement to help differentiate sheep from goat metapodials (see [LINK] Kansa 1999 [/LINK]).

Properties (variables of observation) defined, authored

**Figure 6. Screenshot of Extensible Markup Language (XML) System for Textual and Archaeological Research project (XSTAR) user interface.**

Service Co., 499 U.S. 340 [1991]) (cited by Onsrud et al., 2004). They contain only lists of numbers (facts) and no original or creative information in their organization or expression. Similarly, lists, tables, and databases of objectively determined measurements collected during field research will probably not be protected in current copyright law. Such data represent an important component of much field research. On the other hand, researcher field notes may take the form of written or recorded narratives. Copyright law would inhibit reproduction and distribution of such documentation since the facts in such documentation are embedded within their expression. Similarly, photographs, drawings, and other types of recording all mix fact and expression. Thus, the copyright status of much field documentation (at least in archaeology) is likely to be mixed (depending on the specifics of the records involved) and open to interpretation.

In any event, standard legal copyright protections are in themselves too blunt an instrument to best serve data dissemination among scientific communities. Copyright law provides either no protection (in the case of factual data) or over-protection (most other works) that inhibits the sharing and reuse of content. The lack of protection for factual data can act as a disincentive for open sharing, since data producers may lack assurances that they will be credited for their contributions. At the other extreme, all-rights-reserved copyright greatly reduces the efficiency of sharing because it requires content users to ask for permission for each instance of reusing or even copying of an item of content. Encouraging information sharing requires understanding the incentives and needs of researchers, since they are the most important sources of content. A some-rights-reserved approach that balances the interests of data authors with data users should therefore be explored for scientific cyber-infrastructure projects. This approach, pioneered by Creative Commons, uses licenses to legally secure certain protections (especially attribution) for content developers while enabling efficient communication by granting certain permissions (including permission to copy and redistribute works) for content consumers. Content developers benefit from less restriction on their material since their contributions (along with their attribution) can enjoy wider circulation. Attribution, exposure, and influence in the academic discourse are all essential career-building motivators, and potentially can be measured by tenure committees and university administrators. By granting dissemination permissions with a Creative Commons license, text, images, drawings, graphs, GIS files, and any other media can be freely used and disseminated without specific permission from the author so long as the source is attributed. Use of these licenses removes the legal ambiguities, expenses, and inefficiencies of standard copyright and frees users from locating rights holders, negotiating licensing terms, and paying royalty fees (Samuelson, 2004). Several important applications of these licenses, most notably the PloS biomedical journals (http://www.plos.org/), demonstrate their success in facilitating scientific communication (Association of College Research Libraries, 2003; Brown, 2003; Eisen et al., 2004).

As they stand now, these licenses work for copyrighted content and would not be applicable with factual data. Nevertheless, the overall some-rights-reserved concept can be extended and used as a model for future legal and licensing mechanisms more appropriate for factual data. While the legal status of data is important for dissemination initiatives, the professional value of primary scientific data is perhaps of even more critical importance. Published articles, the currency of the academic market place, are typically narrative syntheses of primary data. Raw data, in itself, is not highly valued unless it is synthesized into publishable paper. Even in relatively open disciplines with traditions of sharing, such as genetics, researchers often fear opening access to raw data, since this would provide resources for competing rivals to publish their own syntheses (Campbell et al., 2002). Since raw data can be exploited by rivals, and their dissemination is not directly rewarded, new licensing terms may be required to build sharing incentives. For example, a noncompete–do-not-republish term might allow the public free access and use of primary research so long as use does not include unauthorized publication in a peer-reviewed journal or similar outlet. This parallels Creative Commons's noncommercial term, a tool that encourages sharing without weakening one's position in the commercial marketplace. A do-not-republish term similarly counters disincentives to sharing in the academic market.

Moreover, a noncompete–do-not-republish term may actually encourage greater collaboration between researchers (Kansa et al., 2005). For example, an interested researcher may develop significant interpretations while exploring large open data sets available under a license from a previous researcher. With a do-not-republish term in effect, however, the interested researcher would then have to negotiate terms, including co-author arrangements, with the raw data's original creator in order to publish his findings professionally (this term should expire after a few years to enable more open use of the data set in question). Such negotiations would not only ensure proper recognition and attribution in the publication process, but also provide exactly the type of interaction where synergies and early peer-review could emerge. Researchers who openly disseminate raw data can thereby attract more co-authoring partners, enhancing their own publication record as well as enhancing the quality of overall research and analysis based on their work.

The Society for American Archaeology's ethical code suggests a great deal of interest in promoting openness and the sharing of information. However, the actual practice of archaeology suggests that the great mass of primary excavation observations and interpretations belong as proprietary knowledge, hoarded by individual researchers. Many researchers have deep concern that they will not be properly acknowledged for their contributions unless their research is disseminated through the protected realm of peer-review journals. However, new intellectual property frameworks, such as the Creative Commons–type licenses can provide scholars with the protections they need to share their primary interpretations. These author protections and permissions are essential tools in building information systems that will see contributions and use. Finally, because archaeological information is so often the subject of political disputes, intellectual property frameworks must also consider the interests of other groups of stakeholders (especially indigenous communities) and the vulnerable and sensitive nature of many cultural sites (Nicholas and Bannister, 2004). Addressing these intellectual property concerns is also an active area of licensing research (Kansa et al., 2005).

### Paying for Free Content

How the scholarly community can sustain an open-knowledge commons presents challenges and requires new funding models and approaches. There are profound challenges to upholding these projects even after they begin to operate (Solla, 2002; Zorich, 2003). Problems in maintaining projects for the long term have already resulted in the dissolution of some pioneering endeavors, including the Archaeological Data Archive Project (www.csanet.org/archive/adap/), which closed in 2002 after operating for nearly a decade (Eiteljorg II, 1997). Collecting data sets from authors, maintaining overhead costs, and continuing to keep abreast of technological change requires sustained funding mechanisms.

Perhaps the viable financially sustainable model is one of community support (Krowne, 2003). This is also the most appropriate model, since the community benefits from an open-knowledge commons. While this may represent a wonderful ideal, how do we prevent the tragedy of the commons, where everyone uses the resource but nobody bothers to sustain it? The question goes back to larger issues of why certain granting institutions and universities support archaeology in the first place. Most such research yields very little in the way of direct profits. The theoretical, philosophical, political, and prestige motivations that support the humanities and social sciences generally outweigh any hope for financial returns. It is no great extension of these motivations to include support for a common digital resource that expresses our understanding of the diversity of human experience. Recently, the journal *Nature* has organized a fascinating online discussion on open access methods, including much discussion on the role of granting institutions in the creation and dissemination of knowledge. The issue has recently come to a head in medicine, where the director of the National Institutes of Health (NIH), Elias Zerhouni, recently announced that the NIH would require open access to all NIH-funded research (Park, 2004). Other granting agencies, including those more directly relevant to archaeology, may opt for similar open dissemination requirements. This approach, where a fraction of grant money is dedicated to knowledge preservation, would potentially generate the required funds for updating and maintaining such a commons. An open-knowledge commons for archaeology may always depend on the same charitable donations, philanthropic granting institutions, and government resources that typically support much of the arts and humanities today. In that sense, a digital commons is only as sustainable as the discipline itself.

## CONCLUSIONS

Julian Richards, a pioneer in digital archives for archaeology, noted that digital archiving should be much more than dumping large data sets onto the Internet (Richards, 2003). Making such information meaningful and well integrated with our narratives, visions, and debates about the past can help create the value that sustains this information. Building tools that can support the meaningful application and integration of this content stands as a great challenge. Archaeology represents an interesting case where field research often varies tremendously because of different trajectories of historical development, different levels of time and financial constraints, and complex legal and social factors that vary regionally and internationally. In addition, archaeologists themselves generally lack the technical and financial resources needed to support elaborate cyberinfrastructures. Many of the same factors impact other environmental sciences.

Because of these realities, there is little immediate prospect for community consensus to converge around a common ontology that supports highly specific semantic integration. Thus, the strategies for successful data integration so prominently demonstrated by the Biomedical Informatics Research Network (BIRN)'s Knowledge Integration of Neuroscience Data (KIND) mediator, and by the ontology-mediation approach of the Geosciences Network (GEON) (Ludäscher et al., 2003; Martone et al., 2002) are not feasible for archaeology. Data integration efforts must therefore aim for what is more achievable, which is data mapping at a much more abstracted and generalized level. The XSTAR project discussed in this paper aims to first support such highly generalized data integration in the expectation that this will ultimately facilitate the development of more semantically rich data mappings. Mapping into XSTAR's global schema is relatively straightforward and fast—key factors in lowering entry costs for individual researchers to participate in data dissemination initiatives. XSTAR's aim to enable data integration for archaeology means that it has potential application in a wide variety of disciplines. Because archaeology is an inherently multidisciplinary endeavor, involving inputs from specialists in the earth, biological, physical, social, and environmental sciences, any global schema for archaeology must accommodate mappings from these other disciplines. Thus, many of the highly abstracted and generalized data structures described by ArchaeoML may see wider application well beyond archaeology. XSTAR's data schema may be appropriate (with or without minor modification) for many other disciplines, including areas of the earth sciences, that have yet to develop formal ontologies of their own.

As discussed, for domains as diverse and divergent as archaeology, data integration should not just be thought of as a tool to facilitate research; data integration should be regarded as the product of research. Drawing a community of researchers to author, share, and integrate data requires new institutional frameworks and incentive structures. All of this must take place within the context of a relatively poorly funded community with few technical resources or institutional structures dedicated to data dissemination. Thus far, the social and institutional context of archaeological publication has impeded the free flow and application of data. Digital information must be used and valued if members of the research community will expend the effort required to share and preserve these resources even if adequate technological solutions were available. Great effort must be focused on how to shape the incentive structures of publication, especially intellectual property and recognition frameworks, so that they encourage, rather than discourage, open dissemination. Intellectual property issues will probably grow in significance as the power of Internet dissemination technologies makes research data more important for commercial, government, instructional, public, and research interests. All of these social, legal, and technological systems must be in place if we are to realize this vision of greater sharing, transparency, and value of the digital record of humanity's past.

## ACKNOWLEDGMENTS

I would like to thank David and Sandra Schloen for their long collaborative partnership, discussion of XSTAR and ArchaeoML, and for their support and encouragement in our work to develop services around ArchaeoML. Stuart Campbell, Elizabeth Carter, Ian Hodder, Ruth Trigham, Ian Morris, John Rick, Paul Goldberg and many others deserve thanks for providing invaluable pilot data sets to evaluate XSTAR and XSTAR related tools. Thanks also go to Creative Commons and Science Commons, especially John Wilbanks, Heather Ford, and Lawrence Lessig, for their continued collaboration in our scientific data and traditional knowledge licensing efforts. Special thanks go to Chris Hoffman (University of California–Berkeley) and the Presidio Trust–National Park Service Presidio Archaeology Lab for encouraging and assisting development of software and services around ArchaeoML. Sarah Kansa, Michael Ashley, Ruth Tringham, and David Schloen deserve special thanks for their invaluable comments and critiques, encouragement, and suggestions on earlier versions of this document. Needless to say, any omissions or errors in this work are solely the fault and responsibility of the author. Finally, I owe special thanks to the William and Flora Hewlett Foundation for their generous financial support that enables this and other endeavors.

## REFERENCES CITED

Aldenderfer, M., and Maschner, H., 1996, Anthropology, space, and geographic information systems: Oxford, Oxford University Press, p. 132–154.

Association of College Research Libraries, 2003, Principles and strategies for the reform of scholarly communication: Chicago, Illinois, Association of College and Research Libraries, Intellectual Freedom Committee, http://www.ala.org/ala/acrl/acrlpubs/whitepapers/principlesstrategies.htm (Accessed June 2004)

Brown, G.O., 2003, Out of the way: How the next copyright revolution can help the next scientific revolution: PLoS Biology, v. 1, no. 1, p. e9, doi: 10.1371/journal.pbio.0000009.

Bumpus, H.C., 1898, The elimination of the unfit as illustrated by the introduced sparrow, *Passer domesticus*: Biological Lectures, Woods Hole Marine Biology Station, v. 7, p. 209–226.

Calhoun, J.B., 1947, The role of temperature and natural selection in the variations in size of the English sparrow in the United States: American Naturalist, v. 81, p. 203–228, doi: 10.1086/281513.

Campbell, E.G., Clarridge, B.R., Gokhale, M., Birenbaum, L., Hilgartner, S., Holtzman, N.A., and Blumenthal, D., 2002, Data withholding in academic genetics: Journal of the American Medical Association, v. 287, no. 4, p. 473–480, doi: 10.1001/jama.287.4.473.

Condron, F., Richards, J., Robinson, D., and Wise, A., 1999, Strategies for digital data-findings and recommendations from digital data in archaeology: A survey of user needs: York, United Kingdom, Archaeology Data Service, University of York, Section 11.7, http://ads.ahds.ac.uk/project/strategies/11.html (Accessed May 2004).

Crespi, B.J., and Bookstein, F.L., 1989, A path-analytic model for the measurement of selection on morphology: Evolution; International Journal of Organic Evolution, v. 43, no. 1, p. 18–28.

Cripps, P., Greenhalgh, A., Fellows, D., May, K., and Robinson, D., 2004, Ontological modelling of the work of the Centre for Archaeology: English Heritage Centre for Archaeology, p. 4–7.

Denning, K., 2003, ''The Storm of Progress'' and archaeology for an online public: Internet Archaeology, v. 15, http://intarch.ac.uk/journal/issue15/denning_index.html (Accessed May 2004).

Doerr, M., 2003, The CIDOC conceptual reference module: An ontological approach to semantic interoperability of metadata: AI Magazine, v. 24, no. 3, p. 75–92.

Eisen, M.B., Brown, P.O., and Varmus, H.E., 2004, PLoS Medicine—A medical journal for the Internet age: PLoS Medicine, v. 1, no. 1, p. 31, doi: 10.1371/journal.pmed.0010031.

Eiteljorg, H., II, 1997, Electronic archives: Antiquity, v. 71, p. 1054–1057.

Flanagan, J.W., Fox, E.A., Clark, D.R., Fan, W., Ravindranathan, U., Shen, R., and Goncalves, M.A., 2004, Jordanian archaeology and ETANA: Developing a digital library for Near Eastern archaeology, *in* History and archaeology of Jordan: Petra, Jordan, ETANA Publications, p. 1–16, http://feathers.dlib.vt.edu/publications.html (Accessed October 2004).

Fox, E.A., Gonçalves, M.A., and Kipp, N.A., 2001, Digital libraries, *in* Adelsberger, H., Collis, B., and Pawlowski, J., eds., Handbook on information technologies for education and training: Springer, p. 19.

Gaffney, V., and Exon, S., 1999, From order to chaos: Publication, synthesis and the dissemination of data in a digital age: Internet Archaeology, v. 6, http://intarch.ac.uk/journal/issue6/gaffney_toc.html (Accessed May 2004).

Gilsen, L., 2001, Archaeological gray literature: The SAA Archaeological Record, p. 30–31.

Grant, P.R., 1972, Centripetal selection and the house sparrow: Systematic Zoology, v. 21, p. 23–30.

Grayson, D.K., and Cannon, M.D., 1999, Human paleoecology and foraging theory in the Great Basin, *in* Beck, C., ed., Models of the millennium: Salt Lake City, University of Utah Press, Great Basin Anthropology Today, p. 141–151.

Gruber, T.R., 1993, A translation approach to portable ontologies: Knowledge Acquisition, v. 5, no. 2, p. 199–220, doi: 10.1006/knac.1993.1008.

Gupta, A., Ludäscher, B., Martone, M.E., Qian, X., Ross, E., Tran, J., and Zaslavsky, I., 2002, A system for managing alternate models in model-based mediation: Sheffield, United Kingdom, British National Conference on Databases, p. 54–57.

Harris, J.A., 1911, A neglected paper on natural selection in the English sparrow: American Naturalist, v. 45, p. 314–318, doi: 10.1086/279213.

Helly, J., Case, T., Davis, F., Levin, S., and Michener, W., 1995, The state of computational ecology: Santa Barbara, California, National Center for Ecological Analysis and Synthesis, Research Paper No. 1, http://www.nceas.ucsb.edu/papers/compecol/ (Accessed January 2005).

Hendler, J., 2001, Agents and the semantic web: IEEE Intelligent Systems Journal, v. 16, no. 2, p. 30–37, doi: 10.1109/5254.920597.

Hodder, I., 1999, Archaeology and global information systems: Internet Archaeology, v.6, http://intarch.ac.uk/journal/issue6/hodder_toc.html (Accessed May 2004).

Johnson, I., 2004, Putting time on the map: Using TimeMap for map animation and Web delivery: GeoInformatics, v. 7, no. 4, p. 26–39.

Johnston, R.F., Niles, D.M., and Rohwer, S.A., 1972, Hermon Bumpus and natural selection in the house sparrow *Passer domesticus*: Evolution; International Journal of Organic Evolution, v. 26, no. 1, p. 20–31.

Jones, S., MacSween, A., Jeffrey, S., Morris, R., and Heyworth, M., 2001, From the ground up. The publication of archaeological projects: A user needs survey: York, United Kingdom, Council for British Archaeology, http://www.britarch.ac.uk/pubs/puns/ (Accessed November 2003)

Jones, S., MacSween, A., Jeffrey, S., Morris, R., and Heyworth, M., 2003, From the ground up. The publication of archaeological projects: A user needs survey. A summary: Internet Archaeology, v. 14, http://intarch.ac.uk/journal/issue14/puns_toc.html (Accessed May 2004).

Kansa, E.C., Schultz, J., and Bissell, A., 2005, Protecting traditional knowledge and expanding access to scientific data: Convergence of two intellectual-property agendas via a ''some rights reserved'' model: International Journal of Cultural Property (in press).

Kilbride, W., 2005, Past, present and future: XML, archaeology and digital preservation: The Center for the Study of Architecture/Archaeology Newsletter, v. 17, no. 3, http://www.csanet.org/newsletter/winter05/nlw0502.html (Accessed April 2005).

Krowne, A., 2003, Building a digital library the commons-based peer production way: D-Lib Magazine, v. 9, no. 10, http://www.dlib.org/dlib/october03/krowne/10krowne.html (Accessed November 2004).

Lin, K., and Ludaescher, B., 2004, GEON: Ontology-enabled map integration, *in* 2004 ESRI International User Conference: Redlands, California, ESRI Professional Papers, http://gis.esri.com/library/userconf/proc04/abstracts/a1796.html (Accessed June 2005).

Lock, G., and Harris, T., 2000, Beyond the map: Archaeology and spatial technologies, *in* Lock, G., ed., Beyond the map: Archaeology and spatial technologies: Amsterdam, IOS Press, 236 p.

Ludäscher, B., Lin, K., Brodaric, B., and Baru, C., 2003, GEON: Toward a cyberinfrastructure for the geosciences—A prototype for geologic map integration via domain ontologies, *in* Digital Mapping Techniques '03—Workshop Proceeding: U.S. Geological Survey Open-File Report 03-471.

Manly, B.F.J., 1976, Some examples of double exponential fitness functions: Heredity, v. 36, no. 2, p. 229–234.

Manly, B.F.J., 1985, Detecting and measuring stabilizing selection: Evolutionary Theory, v. 7, no. 4, p. 205–217.

Martone, M.E., Gupta, A., Ludäscher, B., Zaslavsky, I., and Ellisman, M.H., 2002, Federation of brain data through knowledge-guided mediation, *in* Otter, R.K., ed., Neuroscience databases: A practical guide: Boston, Kluwer Academic Publishers, p. 275–292.

McCartney, P., Robertson, I., and Cowgill, G.L., 2000, Using metadata to address problems of data preservation and delivery: Examples from the Teotihuacan Data Archiving Project, *in* Meeting of the Society for American Archaeology, Digital Data: Preservation and Re-Use Session, Philadelphia: Bryn Mawr, Pennsylvania, Center for the Study of Architecture/Archaeology, http://www.csanet.org/saa/mccartney.html#Info (Accessed September 2003).

Moon, H., 1993, Archaeological predictive modelling: An assessment: Victoria, British Colombia, Earth Sciences Task Force Resources Inventory Committee, p. 1–39.

Nelson, B.A., Kohler, T.D., and Kintigh, K.W., 1994, Demographic alternatives: Consequences for current models of southwestern prehistory, *in* Gummerman, G., and Gell-Mann, M., eds., Santa Fe Institute Studies in the Sciences of Complexity, Addison Wesley, p. 113–146.

Nicholas, G.P., and Bannister, K.P., 2004, Copyrighting the past? Emerging intellectual property rights issues in archaeology: Current Anthropology, v. 45, no. 3, p. 327–350, doi: 10.1086/382251.

O'Donald, P., 1973, A further analysis of Bumpus' data: The intensity of natural selection: Evolution; International Journal of Organic Evolution, v. 27, no. 3, p. 398–404.

Onsrud, H., Camara, G., Campbell, J., and Chakravarthy, N.S., 2004, Public commons of geographic data: Research and development challenges: Public Commons of Geographic Data, http://www.spatial.maine.edu/geodatacommons/ (Accessed November 2004).

Ottaway, J.H., 2001, Publish or be damned, *in* Aslan, R., Blum, S., Kastl, G., Schweizer, F., and Thumm, D., eds., Mauershau, Festschrift für Manfred Korfmann (Vol. 3): Remshalden-Grunbach, Germany, Verlag, Bernhard Albert Greiner, p. 1101–1111.

Park, P., 2004, NIH research to be open access: The Scientist: http://www.biomedcentral.com/news/20040729/04 (Accessed September 2004).

Paterson, A., 2003, The design and development of a social science data warehouse: A case study of the Human Resources Development Data Warehouse Project of the Human Sciences Research Council, South Africa: Data Science Journal, v. 2, p. 12–24.

Price, F., 1996, Bumpus' house sparrow data, *in* Biology in action: New approaches to teaching and learning science: Radford, Virginia, Radford University, http://www.radford.edu/~biol-web/ActionPoster/infoweb.html (Accessed January 2005)

Ravindranathan, U., 2004, Prototyping digital libraries handling heterogeneous data sources—An ETANA-DL case study [M.S. Thesis]: Blacksburg, Virginia, Virginia Polytechnic Institute and State University, 82 p.

Richards, J., 1997, Preservation and re-use of digital data: The role of the Archaeology Data Service: Antiquity, v. 71, no. 274, p. 1057–1059.

Richards, J., 2001, Anglian and Anglo-Scandinavian Cottam: Linking digital publication and archive: Internet Archaeology, v. 10, http://intarch.ac.uk/journal/issue10/richards_toc.html (Accessed May 2004).

Richards, J., 2003, Online archives: Internet Archaeology, v. 15, http://intarch.ac.uk/journal/issue15/richards_toc.html (Accessed May 2004).

Robinson, D., 2000, Digital archiving pilot project for excavation records (DAPPER), *in* Society of American Archaeology, Digital Data: Preservation and Re-Use Session, New Orleans: Bryn Mawr, Pennsylvania, Center for the Study of Architecture/Archaeology, http://www.csanet.org/saa/dapper.html (Accessed September 2003).

Ryan, N., 2004, Databases: Internet Archaeology, v. 15, http://intarch.ac.uk/journal/issue15/ryan_toc.html (Accessed December 2004).

Samuelson, P., 2004, Preserving the positive functions of the public domain in science: Data Science Journal, v. 2, no. 24, p. 192–197.

Schloen, D., 2001, Archaeological data models and Web publication using XML: Computers and the humanities, v. 35, p. 123–152, doi: 10.1023/A:1002471112790.

Solla, L., 2002, Building digital archives for scientific information: Issues in science and technology librarianship, v. 36, http://www.istl.org/02-fall/article2.html (Accessed November 2003)

Trigger, B.G., 1989, A history of archaeological thought: Cambridge, Cambridge University Press, 516 p.

Zhang, C., Cao, C., Gu, F., and Si, J., 2002, A domain-specific formal ontology for archaeological knowledge sharing and reusing: Lecture Notes in Computer Science, v. 2569, p. 213–225.

Zorich, D.M., 2003, A survey of digital cultural heritage initiatives and their sustainability concerns: Washington, D.C., Council on Library and Information Resources, p. 1–47.

MANUSCRIPT RECEIVED BY THE SOCIETY 9 MARCH 2005
REVISED MANUSCRIPT RECEIVED 19 JULY 2005
MANUSCRIPT ACCEPTED 2 AUGUST 2005

Printed in the USA