Future Generation Computer Systems 29 (2013) 2117–2127

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

ArchaeoSTOR: A data curation system for research on the archeological frontier

Aaron Gidding [a,b,*], Yuma Matsui [b], Thomas E. Levy [a,b], Tom DeFanti [b], Falko Kuester [b]

[a] University of California, San Diego, Department of Anthropology, 9500 Gilman Drive, La Jolla, CA 92093-0532, USA

[b] California Institute for Telecommunications and Information Technology (CALIT2), University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0436, USA

HIGHLIGHTS

- We review the needs in archeology for data management.
- We introduce a pipeline to characterize archeological legacy and field data.
- We introduce a spatial based application for calling up archeological data.
- We establish that a tool like ArchaeoSTOR is necessary for modern field research.

ARTICLE INFO

Article history:
Received 6 March 2012
Received in revised form
29 January 2013
Accepted 6 April 2013
Available online 19 April 2013

Keywords:
Archeology
Visualization
Data comprehension enhancement
Database technology
Web applications

ABSTRACT

The broad adoption of diagnostic and analytical techniques in the field of archeology, presents a unique opportunity for e-Science in the form of scientific explanation, drawing from methodologies aimed at recording, archiving, analyzing, and disseminating, rich data collections to create the needed infrastructure for both research and web-based curation and data management system. This paper presents a needed stepping stone towards synergy between information technology and archeology, by introducing a data acquisition, tagging and characterization pipeline along with a novel method for spatially querying archeological data.

1. Introduction

The primary goal of archeological research is to tell the story of past cultures using the material record that is left behind. To tell that story archaeologists have increasingly been turning to more efficient methods within the context of the modern research environment. This paper outlines the introduction of a new web based data management system that we call ArchaeoSTOR. ArchaeoSTOR provides the organization necessary for the purpose of speeding up the process of storing complex archeological data and delivering that data to novel tools for data visualization and analysis. Importantly, the design of the system allows for the user to generate data in the field, off the grid, using a remote server and copy data back to a central repository after field research is complete. This approach is essential to archeological research, which has begun to adopt a number of data intensive research techniques, but can take place in remote locations with unreliable or slow internet access preventing regular access to an off-field-site server. While ArchaeoSTOR has been developed for archeology, it provides a model for infrastructure that can be repurposed and used in other research fields that include long research agendas in more remote areas.

A number of new technologies have become increasingly available offering rapid, diagnostic analysis of artifacts. However, innovative technologies introduce new problems with increased data volume and complexity. This complexity becomes more challenging for archeology as it can come from novel instruments and techniques applied to many sites, consisting of dissimilar material types that date to diverse time periods. The multi-dimensionality represented by this variability in data requires the adoption of e-Science to help create a valid narrative of the past. In archeology most research is driven by the public consumption of shared cultural heritage through various media outlets (e.g. museums, books,

* Corresponding author at: University of California, San Diego, Department of Anthropology, 9500 Gilman Drive, La Jolla, CA 92093-0532, USA. Tel.: +1 858 822 1676.
E-mail address: agidding@ucsd.edu (A. Gidding).

0167-739X/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.future.2013.04.007

television) and now the Internet. The challenge for cultural heritage research is how to fully integrate the new technologies into a comprehensive research program for the purpose of communicating not just polished results to the general public but also the data used to generate effective, collaborative research.

Total digitization of research in archeology offers a way for archaeologists to meet the challenge of rapidly bringing excavated data to the public and professional communities efficiently. By digitizing the entire archeological research process it is possible to provide a means by which archeological data can be queried and analyzed easily leading to simplified dissemination of data for research and the public.

2. Archeological data management

Managing complex data sets has always been one of the most difficult aspects of archeology. Archeological data is nuanced both in its recording and subsequent presentation. This is inherent to the nature of archeological data, which is drawn entirely from artifacts that present a 'silent' record of human activity for which multiple interpretations are possible. Building such a narrative for archeology has traditionally meant spending countless hours analyzing and comparing material from a variety of sources, derived through different methodologies. Additionally, data is generated in remote and harsh environments, without the ability to connect to a central server. The best way to process this varied data is to link the methods of data collection through a manageable data system that uses a common denominator to relate findings. Not only does the data need to be well managed, but the system should also allow for on-the-fly recovery of the data as it is being excavated to facilitate different visualization environments to aid in the excavation and real-time analytical processes. Finally, data that is added in the field needs to be rectified with a central storage server that is used for high performance computing when not in the field.

Archeological data management takes many different forms from excavation to excavation. Every project has its own methods that are used to archive the excavation process. Nonetheless, there are congruencies between the kinds of data that are collected at all excavations. Various researchers have worked on ontological issues to describe these similarities in methodology and as a way to bridge the gap between the different methodologies between researchers [1,2]. For instance, excavators will always map both the horizontal components of a site along with the vertical stratigraphy. Additionally, the site is always divided up into sections in order to give better context to artifacts collected. Artifacts are collected within these spatial partitions in order to maintain contextual integrity and important finds are recorded as individual points. Similarities between archeological excavations allow for the identification of facets of archeological data collection strategies that offer lowest common denominator data types necessary to begin using digital data management.

As a whole, the discipline of archeology has been slow to adopt digital techniques that encompass every facet of their excavation methodology. This is especially true when compared to other fields that have been able to mobilize large-scale efforts to organize data management strategies as a field [3]. Typical still are the paper recording sheets to archive and manage excavated material. Maps are drawn manually as the excavation proceeds without digital recording.

Besides issues of funding, one reason that a complete digital methodology has not been adopted comes from a lack of standard ontology for aspects of research outside basic excavation methods. Although archaeologists have been using computers for archeological research since the first personal computers became available, the use of digital technologies is often only applied to specific facets of an excavation and not with data while it is collected in the field. There is no universal software or organizing principle for archeological data storage and a result communication between projects has been historically difficult. Volumes describing applications of digital technologies in archeology have been published through the years but have not attracted widespread appeal [4]. A few systems have very recently been developed commercially, such as the Archeological Recording Kit (ARK) [5], to help guide archeological research from excavation through publication and digital dissemination. Conferences like Computer Applications and Quantitative Methods in Archeology have discussed many of these issues over the years and presented different ways of handling these problems [6]. Unlike other data management projects, we are not looking to use ArchaeoSTOR as a data management system that can readily ingest outside data, as much as its ontology is readily accessible for other users and archeological data services.

The majority of examples from archeology concerning the adoption of digital technologies highlight test cases that use a dataset to investigate a narrow analytical objective [7], but do not account for the total archeological process. The independence of excavations and the resulting lack of standards in ontology have not helped the development of a single system. Nonetheless, digital recording methodologies are currently being employed by archaeologists in a variety of ways to enable rich description of artifacts through diagnostic analysis of material. Some projects such as VERA (A Virtual Environment for Research in Archeology) have started the process of experimenting with using digital tools in the field for direct digital data recording and later dissemination [8]. However, these applications do not fulfill the needs of all long-term excavation projects, especially those like ours that are in remote areas with unreliable and slow internet connectivity. We present a cohesive archeology system, in the form of ArchaeoSTOR, as an alternative here.

3. Archeological data recording—the Edom Lowlands Regional Archaeology Project (ELRAP), Jordan

Typically archeology seeks to create a narrative from three sets of data: the material record, its spatial context and the temporal setting that allows for comparison of records between sites. Over the years and inter-regionally, archaeologists have used a number of methodologies to record artifacts and their spatial context to answer different questions about the past. The application of different methods has meant that comparing archeological data is difficult and fraught with complexities of melding different research methodologies established for dissimilar research objectives. This variation can even exist within a single project making the integration of 'legacy' data with newly derived data difficult. In order to deal with this problem a digital research environment that is flexible enough to handle multiple interpretive frameworks and offers a framework that allows for easier comparison and integration of data with a well-defined ontology is necessary.

Within the UC San Diego ELRAP expedition we have considerable experience with various facets of collecting archeological data digitally. All of ELRAP's work has been focused in the area of the southern Jordan knows as Faynan—one of the largest southern Levantine copper ore resource zones exploited in antiquity. Our research has focused on a 'deep-time' study of social change as it relates to ancient mining and metallurgy from Neolithic to Islamic times (∼8700 BCE–∼1700 CE). To record the material record for this project our excavations have been using digital recording methods for over a decade. The digital recording methods that we have employed over the years have been outlined recently [9]. However, over that decade of digital data collection it has become increasingly clear that maintaining an organized flat file structure was simply not a long-term solution given the increasing complexity and size of data collected season to season.

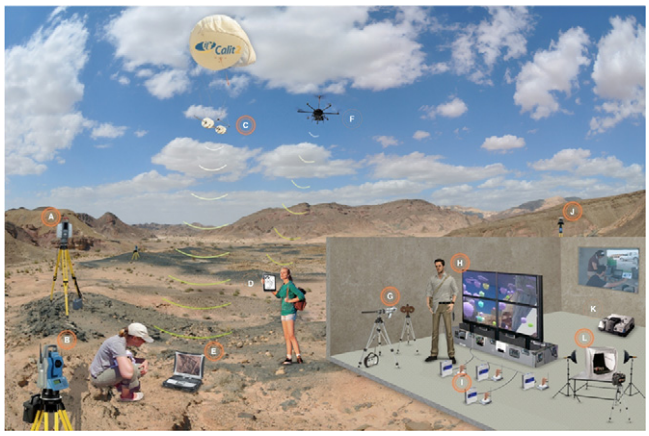


Fig. 1. A model of the different components that are a part of our field recording methodology. A. LiDAR Scanner B. Total Station C. Aerial photography platform D. Digital Note taking (Opendig) E. Control for Aerial photography platform F. Octocopter forensic imaging platform G. XRF H. Nexcave portable display I. Nextengine 3D scanners J. dGPS K. FTiR L. Digital object photography.

For ELRAP, digital recording was initially adopted to aid the laborious mapping of archeological surveys and excavations. Over time this basic goal has evolved using different technologies and methods as they became commonly available, feasible and affordable. The technologies we have adopted are varied in application and data output; to date we have published using high-resolution digital imaging, light detection and ranging (LiDAR) scanners, airborne imaging platforms fitted with high-resolution digital cameras, high-resolution desktop 3D artifact scanners, portable X-ray fluorescence (XRF), differential GPS (dGPS), Fourier transform infrared spectroscopy (FTiR) and different absolute dating techniques (high precision radiocarbon and paleomagnetic dating; see Fig. 1). These new sources of data offer a number of challenges over time: they need to be integrated into existing data acquisition channels; they all use different data types/formats that are often incompatible effectively partitioning their use to discrete aspects of analysis; they represent a digital 'data avalanche' with increasing density/complexity that has become difficult to manage; and new technologies that use these data-sources are constantly being developed and applied which use already collected data in their own novel ways [10].

A good example of how a single type of data has transformed in utility for archeology through the years is the photograph. Before digital photography, images were taken with great care of valuable finds and to document excavated site features of interest. The number of images in a season rarely exceeded 1000. These images were highly prized and primarily served a single purpose, publication. The early adoption of digital photography mirrored that process in archeological research. Now we photograph every artifact multiple times, with more comprehensive images of interesting features on site, and have added aerial and GigaPan photography. We generate images by the tens of thousands each season in a raw format that takes up over a terabyte of data. Not only are these images used for publication, but they are streamlined into many different analytical functions for scientific visualization including the creation of structure-from-motion point clouds, display in three-dimensional models of the site, and guiding the illustration of features of the site. Thus, the increased complexity of how we deal with photography represents a subset of the exponential growth of data in terms of size, management complexity, and processing power need for applications using the data that we collect each field season.

Two other data systems are used by the ELRAP to record data in the field and are complementary databases used by ArchaeoSTOR to describe the archeological excavation. Firstly there is data recording software called Archfield [11]. The advantage of Archfield is that all spatial data is stored in a PostgreSQL database, which makes connections between the databases simpler. This also facilitates rapid data entry through barcodes assigned to artifacts and contexts. OpenDIG is a piece of software that allows for digital site description and note taking [12]. The three systems work together in order to provide the necessary recording of field excavations.

Other data organization systems have been built within the archeological community over the years to try to deal with some of these problems. All of these focus on three different levels of analysis: macro-scale, micro-scale, or something in between. On the macro-scale the most common databases store information regarding individual site locations and some basic details as opposed to comprehensive data regarding the excavated materials examples of these types (see Digital Archeological Atlas of the Holy Land (DAAHL) [13] and Pleiades [14]). Other systems have been designed to help organize single-site level data focusing primarily on the basic spatial recording of artifacts and their contexts (see ARK [5], Online Cultural Heritage Research Environment (OCHRE) [15], Reconstruction and Exploratory Visualization: Engineering meets Archeology (REVEAL) [16], Integrated Archeological Database (IADB) [17]). More recently there has also been a push to create digital repositories that can accept any kind of data, with coding sheets provided, in order to provide a central place for broad access to research (see the Digital Archeological Record (tDAR) [18], Archeological Data Service (ADS) [19], or Open Context [20]). However, none of the systems mentioned above deal with the combination of intra- and inter-site analysis that includes a number of diagnostic tools for material analysis such as XRF and FTIR.

The most significant differences between ArchaeoSTOR and the aforementioned systems is the explicit focus on artifacts rather than contexts and independence from the internet for daily usage. As one might expect, the majority of the different archeological data management systems have some capability to input data regarding artifacts. What many lack, however, are tools that complement the digital record with the physical nature of artifacts. ArchaeoSTOR offers a system to manage the location of artifacts and to build inventories for shipping, museum loans, and other common occurrences. Our focus on tightly integrating microscopic and diagnostic methods from material science offers a method for artifact characterization within the database. Additionally, the data input mechanism relies on digital context records established during excavation in order to build the artifact records. Rather than relating context and artifact during data entry, relationships between the two are defined from the start. Importantly, all of this data is recorded in the field without a connection to a central server.

4. System map/conception

In order to develop our system we identified the most basic units of analysis and expanded workflows in order to maintain analysis at the lowest common denominator possible to establish a well-defined ontology. The primary data that all archaeologists deal with are artifacts and all data types collected are essentially annotations for artifact description. ELRAP field workers collect artifacts in two ways simultaneously using an on-site GIS recording system. One collection strategy collects artifacts in bulk within a predefined spatial unit. These "bulk" artifacts are then sorted and processed as a part of later analysis. The second strategy records the precise location that an important find is located along with an on-site diagnostic description with a wide variety of different designations possible. The metadata that is stored as a part of annotation provide the means to anchor artifacts in space and time, to characterize digital data assets beyond brute force recorded data

and reevaluate the data and its validity at a later time. We distinguish within the data two different ways of identifying difference between artifacts: material difference and meaningful difference. The concepts of material difference and meaningful difference are modeled on the idea of genotype and phenotype in genetics. This model works to show how archaeologists evaluate the meaning behind observed variation in archeological material.

Material difference is derived by analysis that represents a distinct set of operations using microscopic and diagnostic tools (i.e. XRF, FTIR) that provide objective metrics for differentiating between objects. The results from such tools provide objective results that are directly comparable using mathematical functions. For instance using XRF analysis the relative elemental composition of any given material can be represented and compared to other samples allowing a variety of conclusions to be drawn regarding materials used to produce a given archeological sample. Chronological variation in material composition can be observed by comparing the materials from different strata or layers from a given site. Additionally, the provenience of archeological material can be detected by comparing materials with their sources. This kind of analysis follows a methodology that looks only at abstract traits of materials in a quantitative way, but ignores other important features of any object that inform qualitative comparison of archeological material. This could include comparing formal or stylistic similarities between artifacts made of different materials that might not be obvious using mechanistic techniques.

The identification of meaningful difference between artifacts that cannot be captured digitally is also taken into account within the database. Meaningful difference is constructed of features that are derived and most efficiently characterized through human interaction and observation representing multiple dimensions of any given artifact. This would include an understanding of functional description, formal-typological description, and other contextual information that is meant to help lead diagnostic analysis. Using annotations based on meaningful difference we can be sure that we are always comparing apples to apples as opposed to apples to pears. Given the aforementioned variety in the interpretation of ancient material culture, flexibility in understanding what makes material difference significant is important, and it is assumed that the difference observed through diagnostic analysis should validate any conclusions. In this way observations of difference and meaningful difference work as a part of a reflexive system bridging different ways of understanding the ancient past.

The material difference and meaningful difference distinction could be compared to the semiological etic and emic distinctions; however this would ignore two important mediating data types fundamental to the system: time and geography. By tying our data to absolute dating techniques by using geographic information systems, our data organization system offers a framework for interpreting the conventional archaeological classificatory structures using a combination of both visual and traditional query techniques.

Spatial annotation is fundamental to how we deal with these aspects of difference. Over the course of an excavation, while unique finds are individually recorded, there are too many "bulk" artifacts (large numbers of bones, lithics, pottery sherds, etc.) collected to give each one its own spatial context. Thus, we divide the artifacts into two categories: classes and singularities. A class is a group of artifacts taken from a similar spatial context and grouped according to basic material and functional category. This means that the artifacts that are made of ceramic are grouped together while the artifacts that are made of flint are given their own grouping and so on. Artifacts within classes are broadly categorized according to features deemed important and some artifacts may be separated and analyzed as diagnostic based on formal or constructive markers. Additionally, in the field, artifacts can also be identified as having unique attributes and despite belonging to a class are identified as singularly important finds and are recorded as such in order to maintain tighter contextual control for those artifacts. These artifacts identified as diagnostic either in the field or as a part of lab analysis of the "bulk" artifacts are singularities. These singularities are the artifacts that are the primary source of descriptive diagnostic description to characterize ancient social activity (i.e. mining, exchange, etc.). By isolating the material into defined spatial units we are able to help guide the understanding of what constitutes meaningful difference and make further distinctions in other dimensions and the broader categories.

5. Technical aspects of the system

In the technical architecture of our system, a data repository server to aggregate our data plays a central role. The server components consist of a database server, application server, and web application. Various open source software were used to make our system independent of proprietary technologies.

The database server is PostgreSQL with PostGIS extension, which is a reliable relational database software and can handle spatial data—critical for archeological research. Additionally, it allows us to have multiple users accessing and manipulating the data at once, necessary given the interoperability of each dataset. The application server is Jetty, and it is common for running Java-based web applications. We implemented the data system as a web application to efficiently interact with the database through the web interface provided by the application server. Additionally this provides a ready application for sharing our data with other researchers in the future. We use the Grails web application framework to realize maintainable and productive software development. It provides a way to define database schema with a simple object-relational mapping model and auto-generate templates of application to manipulate database. The client components consist of any web browsers, GIS clients that can connect to the database, and other visualization environments. We are currently using Quantum GIS (QGIS) as a GIS client for intensive GIS applications because it is both open source and user friendly.

One of the major challenges that we faced in developing the system was dealing with implementation for field use. We cannot access our servers on campus in California while at our remote Jordanian field site. The ability to sync months of collected data seamlessly upon return is fundamental to our operating objectives. Therefore, portability of the system and the ability to sync with any data left behind have been stressed as a fundamental organizational part of the system. The present system is running on virtualized Linux environment (CentOS), but any operating system that supports the software components described above will work. We employ virtualization technology to use the data system both on the campus and excavation sites. By using virtualization technology, the whole environment of the data system can migrate from a computer to another computer with just copying the disk image. This is useful for field research because we can carry the virtualized server environment anywhere loaded on portable computer hardware, while it runs on larger-scale server hardware on the campus. By effectively using the same server both on campus and in the field, despite no internet connectivity, allows us to use a single map for how our data is organized through the system (see Fig. 2). When we return from the field we can take the virtual machine and sync it back with the data on campus. Currently this synchronization is one way from the field server to the campus server because new data is not added during excavation. One of our planned projects is to make synchronization between the field system and the system that remains in San Diego possible (see Fig. 3).

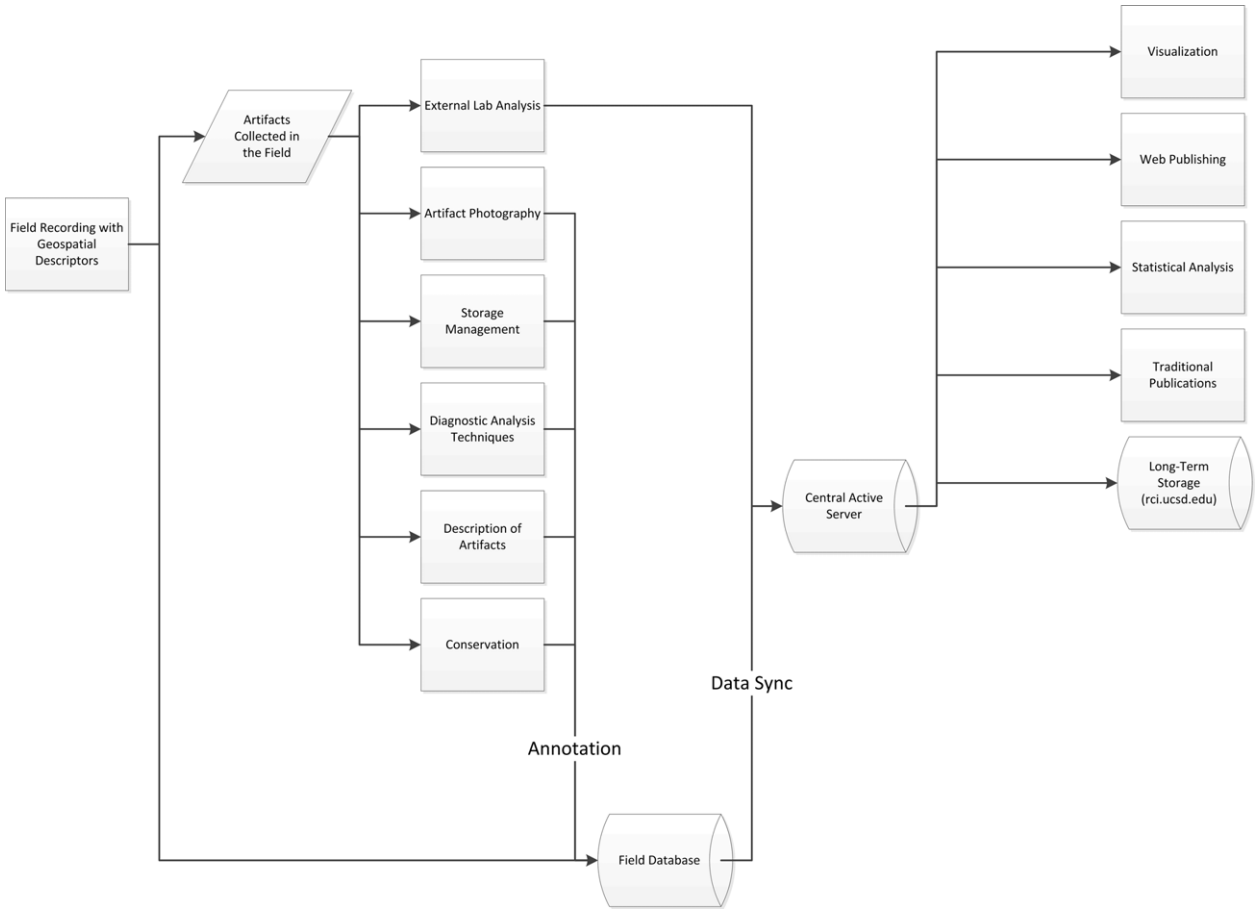


Fig. 2. A simplified flowchart illustrating how artifacts are recorded and stored in the database for later retrieval.

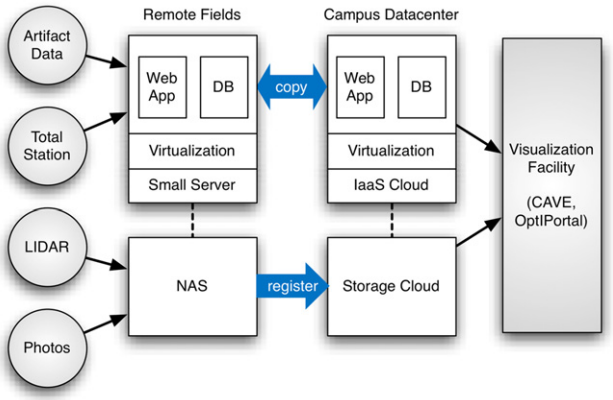


Fig. 3. Flowchart illustrating the system components and their relationship for ArchaeoSTOR.

6. Current use of the system

When building our database we needed to be sure to develop around anticipated pipelines for data ingestion in field-like conditions. One aspect of this was to begin the initial testing of the data system with the narrow goal of showing the feasibility of incorporating datasets that show both difference and meaningful difference. For this we decided to incorporate a traditional artifact analysis methodology (sorting pottery according to formal variation), with one of the newer diagnostic technologies that we have adopted, XRF. When we began our field tests we also included a utility to visualize spatial data and other associated data. The benefit of this tool (discussed in the next section) is that archaeologists can instantly grasp visual overview of sites and artifacts after coming back from excavations. It is also advantageous that archeological data stored then can be directly published and shared on the web using spatial and SQL interfaces.

When dealing with every assemblage of material there are a number of organizing principles used to classify the data. In the case of pottery, basic descriptive categories are assigned based on formal and general characteristics. However as already mentioned, the categories that are important to a ceramicist vary according the spatial context of where the material was found within the excavation. For the purpose of class based analysis there are standards that are employed for the purpose of analysis and form an easy to identify lowest common denominator for the purpose of the database. In order to manage each individual data type within these different structures is a unique Id is assigned that can be tracked through the system; not only within the database, but also physically using a bar coding system.

The implementation of data collected from XRF is a much greater challenge to fit into the database. XRF is an important tool for identifying the relative quantity of different elements that constitute a given material. However, different settings on the XRF machine are required to measure different element groups, which results in varied results derived from methodological differences. Any time a material is analyzed using different settings the parameters for recording the relative frequencies of elements changes. This means that depending on what aspects of a given material are being analyzed the results can vary considerably. While XRF is useful for a variety of analytical purposes, we need to be very careful to record the methods used for analysis in order to maintain data consistency. Additionally, consistency in how samples are analyzed is important as the software calculates quantities of a given sample

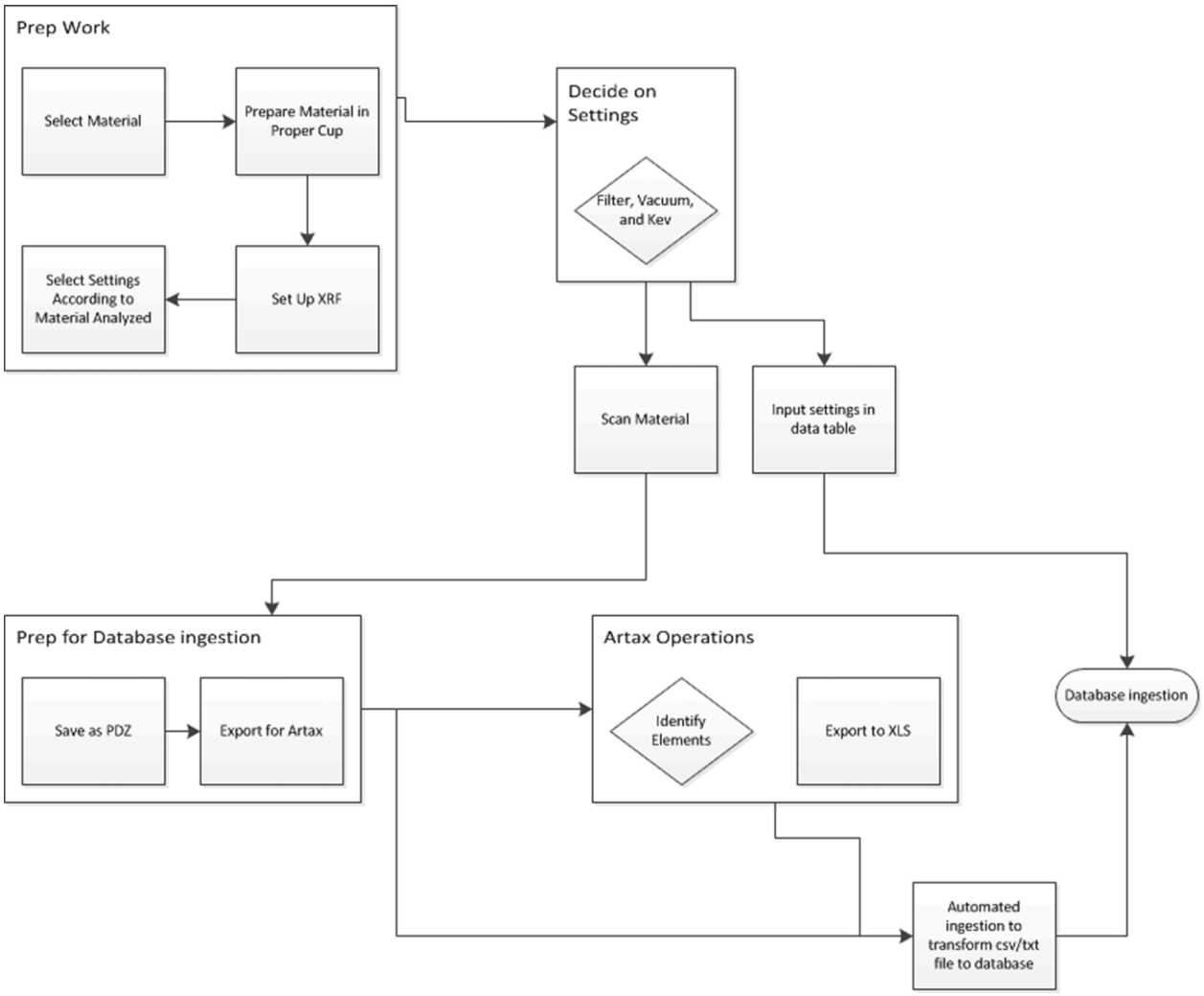


Fig. 4. The workflow for processing XRF data for ingestion into ArchaeoSTOR.

differently according to the total samples run. In order to deal with these problems we have broken down data storage into two different components, the raw data and processed data, both of which we can pull from depending on the analysis that is desired.

Additional consideration has to be made for the manufacturer of the machine used for analysis. Each manufacturer has different specifications for how data is recorded, stored and processed. These specifications apply both to technical differences in the quality of the data and the data formats used to store the raw data. As a result the raw data created by different manufacturer's instruments is not readily interchangeable until full processing to ppm measurements is performed. Given our aim of using the XRF data before PPM calculations are made, we account for this problem by specifying the machine used and the methods implemented to process the data. However, currently we organize the data specifically for the brand of XRF that we employ (Brucker).

By processing the XRF data in two stages we can better use the data as it is processed. The uptake of data at two different points of collection for the XRF introduces another problem due to the variety of file types and variability in how data is collected. The files used by the software associated with the XRF are proprietary and need to be converted to a format that can be easily stored in an open-source PostgreSQL database. This is a process that is not difficult, but it means that as a part of the recording/import process care needs to be taken to make sure that all of the settings are also recorded as metadata for data consistency (See Fig. 4). The software embedded in the data system can parse the XRF data file, store parsed data in the database, and export the data in its original XRF data format. Assuming that consistency is maintained then the database will benefit users by allowing the ability to query for similar samples in order to calculate results from larger groups of material that may have been scanned for many different projects. Other archeological data including site information, survey feature information, and artifact information can be manually input with data entry web forms that our new system provides (see Figs. 5 and 6).

Within the server several workflows are in place to put field data into the database system according to different typological categories. Regarding geospatial data, measuring devices like total stations produce shape files, which are standard geospatial vector data format for geographic information systems, and others like dGPS devices produce raw data in their original format. Our data system can import shapefiles into PostGIS database by converting their formats. For the data collected with the dGPS we preprocess the raw data into shape files for import. This function is more important for dealing with ingesting either legacy or outside research groups' data. For our field collection we use our own program that can sync directly with our server after returning from the field site.

It is with the goal of testing the database system in 'field-like' conditions that the application of our database management system was applied and tested with ceramics from the ELRAP site

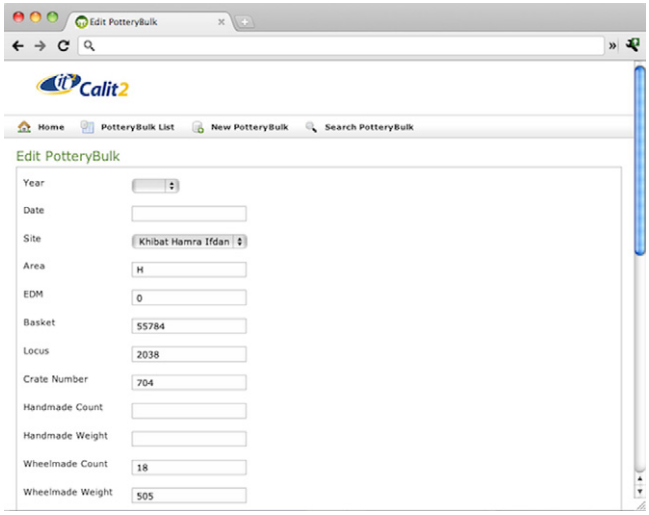


Fig. 5. Basic web interface for data input. Most data input takes place using a form designed for the different material types. Contextual data is automatically entered into the database by relating the table to geospatial data as it comes in from the field.

Khirbat Hamra Ifdan (KHI) in southern Jordan [21]. The important thing was to test the functionality of the data collection system, process it using the built web interface and then visualize it within the interface of a GIS program like QGIS. Each of these components came together over time to the point that visualizing the data collected was not a problem. As the organization of the data became a seamless process, by scaling the database to larger numbers of artifacts and other data-types, we have put in place an integrated database system capable of handling the full range of artifacts to be recorded in our next field season.

7. Spatial data queries

One new feature that we have begun to implement in the data organization system is the ability to query the data spatially using a web application. Already we noted our use of QGIS to generate maps used to illustrate relationships between data at sites. However, GIS programs, like QGIS, offer static interactivity as a feature of design as a powerful map-making tool, offering a low degree of simpler click-and-use interactivity. For more interaction with spatial data while performing other tasks, a click-and-use interface is preferable. This tool has a number of applications for archeological research that transcend traditional GIS applications. Firstly, it allows for rapid recall of the spatial relationships between excavated units. Additionally, archeological data is visual by nature so spatial queries are a more natural way to interact with data over the course of research in and out of the field. Lastly, we can associate other kinds of visual metadata, like photos, as a part of the query tool enhancing the abstract geospatial data collected for rapid recall.

In order to visualize the data within a web browser we are using a program called GeoServer [22]. As an opensource software server, GeoServer allows us to publish a wide variety of data types and implement some of our own tools for outputs that are useful for the needs of archaeologists from a web-based server outlined above. We also use OpenLayers [23] to visualize geospatial data. OpenLayers is an open-source JavaScript library that allows web applications to communicate with geospatial data servers and to render map data on web browsers.

The ability to spatially query data is an important part of archeological research to help match the excavated material record with the space from which it came. The importance of maintaining spatial relationships in the metadata has already been discussed especially in the contexts of using GIS to visualize that data. The process of opening up a GIS program and loading all of the data for that space is inappropriate in some instances due to the extra time it would take to access data in a GIS program. For instance, often times when doing material analysis using the data management system, the ability to click on an archeological context and see a set of images and other associated data instantly is exactly what the researcher needs while working with material, in addition to then being able to visually query neighboring contexts (see Fig. 7).

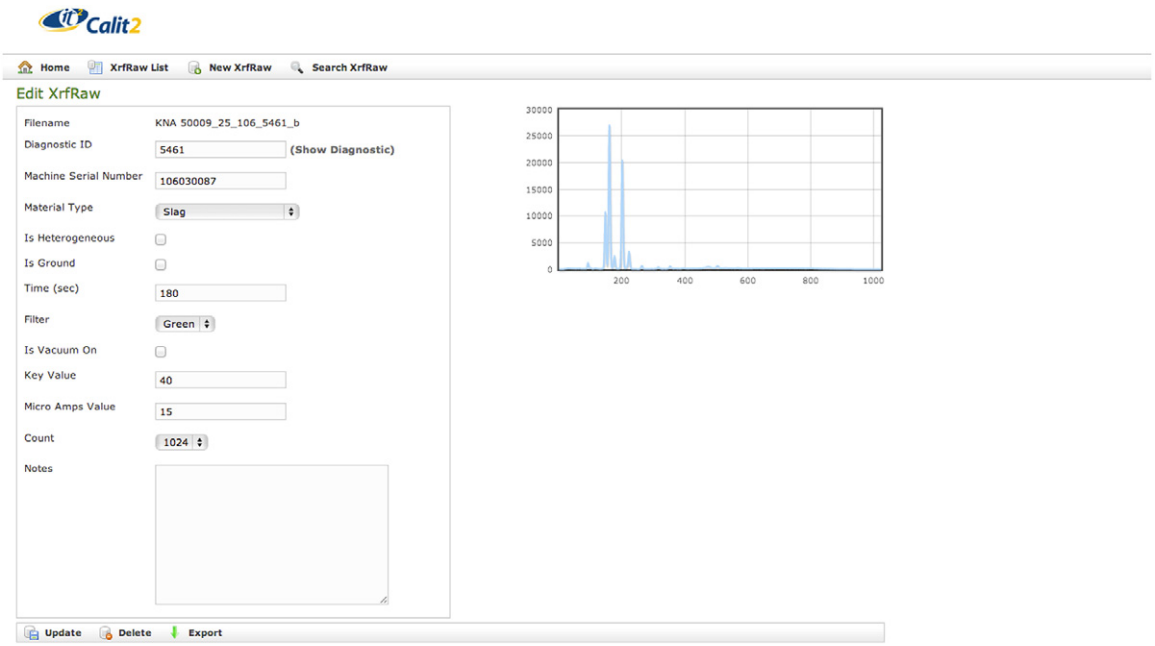


Fig. 6. XRF data input interface with graphic showing XRF sample spectrum.

Fig. 7. In the figure, lines form shapes of sites, and points represent locations where artifacts are found. When artifact points are clicked on the map, detailed artifact information and photos are retrieved and shown.

This application of spatial queries also offers a great platform for sharing archeological data with the general public and other researchers. By generating a curated set of data specifically for public consumption, researchers can share a fully interactive map with can be accessed directly on the web without using a stand-alone GIS program or special plug-ins beyond Java. The experience of directly selecting relevant data on a map is particularly useful in the context of collaboration that otherwise would require sending large datasets, over which there might be copyright issues.

8. Preliminary study results—thoughts from the most recent field season

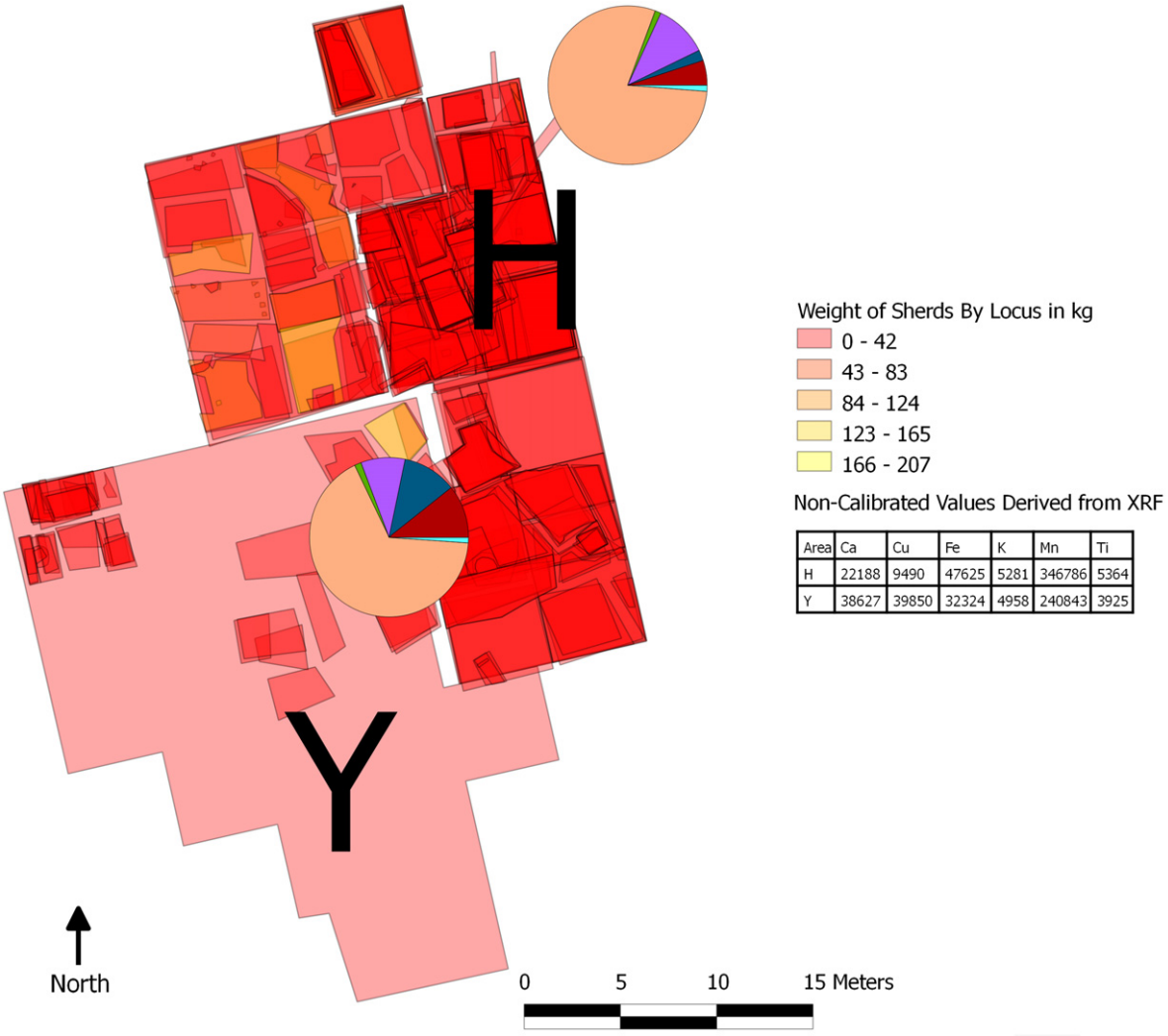
In order to test the system we used it first in one of the active laboratories at UCSD and then deployed the system during the 2011 field season in Southern Jordan. For the preliminary test of the new data ingestion and visualization system we decided to use legacy data from a previous excavation at KHI. KHI was excavated in 1999 and 2000 and is the site of one of the largest copper manufactures in the Southern Levant from the later half of the Early Bronze Age (2500–2000 BCE). There is a large collection of unpublished artifact data that has only undergone preliminary analysis. In using the material from this site we were able to test two intended applications of the system in development: ingesting legacy spatial data while taking inventory of excavated material and using the legacy data as a model for ingesting outside data. Based on how labor intensive it was simply handling legacy data we decided to limit our current goals to the ingestion of legacy data. From the data stored in the database we created a map (see Fig. 8) using QGIS that illuminates how data might be used to analyze datasets in the GIS environment to illustrate the feasibility of the project going forward.

The large quantity of legacy data produced by ELRAP at KHI needs to be ingested into the integrated database described here. Dealing with legacy data is a time consuming process because over the years the methods for data collection changed with the implementation of new technologies. This means that applying the old data to new standards takes some time to manipulate into a format that is workable. The most time consuming part of this process is making sure that the spatial data collected in the past can adequately be used in the new system and then correlated with other types of data within GIS software.

For field archeology projects, taking inventory of daily-excavated material is one of the most fundamental activities. Therefore we decided to simulate the inventory of material stored in San Diego as though we were in the field in Jordan. The original inventory of the artifact material was never adequately stored digitally during the earlier excavations, so we were able to explore how the workflow could be best managed with the new digital inventory system. We approached the method of analysis as would be expected in the field to evaluate issues of performance and usability, adding and moving features as we became aware of changes needed. Over time we added a number of features that helped functionality while looking at the data. These include: easier export of data, querying ability, and bar-coding for rapid recall of data that has been processed.

Our work with XRF data consisted primarily of making sure that the import/export feature worked seamlessly with the proprietary software provided by the manufacturer. Once the database was capable of ingesting and exporting data for raw processing we queried the data from previous analysis using the new system and processed it for storage in the new data system. This storage mechanism allows for us to have clear metadata describing the circumstances for any processing of the raw XRF data when used for future comparison. After making necessary adjustments to workflow and other problems that were apparent during initial testing we loaded the virtual machine on a Mac Mini Server for use in our field laboratory.

The field application of the data system allowed us to test a number of functions beyond Early Bronze Age Pottery. The most significant changes that we made to the workflow for field use were reactions to the quick recognition of missing data elements



Area	Ca	Cu	Fe	K	Mn	Ti
H	22188	9490	47625	5281	346786	5364
Y	38627	39850	32324	4958	240843	3925

Fig. 8. A map of two adjoining excavation areas at KHI using open source GIS, Quantum GIS, to visualize the relative quantity of pottery and an XRF measurement of a slag sample. The fills for the loci are transparent to reveal the weights at lower levels. The pie chart shows the relative measurements of different elements identified in an ancient metallurgical slag sample.

in the database. We refined the data management protocols and fields to materials beyond pottery and to time periods beyond the Early Bronze Age. This meant adding a number of material classes that we had not previously thought of, and refining how we consider diagnostic materials to include ecofacts (ancient ecological remains), which are always collected for diagnostic analysis to help build contextual analysis of excavated units. The technology most responsible for the recognition of ecofacts was our new implementation of a workflow for storage of FTiR raw data. This was the first season that we used FTiR as a part of the excavation process and more work remains to be done in order to fully integrate it into our excavation workflows. Lastly, we connected ArchaeoSTOR to a new system developed to record geospatial data for this past field season. This new system stores all of the field data in a PostgreSQL database with its own set of barcodes. We linked the databases to make data entry move faster, compared to the manual entry currently required for legacy data.

In short, we have been able to show the functionality desired from the data management system described. By using this new database, we were able to ingest spatial data used to describe the context of the finds, describe the artifacts and associated materials and finally, query the results for publication. For every process we were able to have multiple clients manipulating the same data allowing for real-time observation of the data as it was being produced.

Additionally, the new database allows for easier compliance with current data sharing protocols that are now becoming common for the social sciences. The scope of this project will facilitate fulfillment of the National Science Foundation's data sharing initiative. It is important to note that while we are trying to use language that is broadly relevant to archeological research, international researchers use different terminology to describe archeological phenomena around the world. Our Levantine data is structured in a way that is not meant to be universalizing, but it will allow for quick intake into the other larger data sharing initiatives specific to archeology that are designed to help mediate differences in ontology. Examples of these data sharing initiatives already mentioned are the DAAHL and tDAR.

9. Future efforts and conclusions—legacy data and long term storage

Developing a low-cost data management system that facilitates adherence to the guidelines set out by funding agencies for research has been one of the underlying driving points for our research. Developing efficient workflows for ingesting legacy data

into our data system, especially to facilitate public access to completed research, remains one of the hurdles that we are actively seeking to pass as a part of our research. Not only is public access to our data important for other researchers, but it also fundamental to archeological research to have this data readily at hand for comparative analysis.

The implementation of the digital archeology database system presented here illustrates a step towards a new frontier in data interoperability within archeology. Our project is one of many that signify the beginning of efforts to take the material culture record and digitize it without loss of fidelity. The process of data transformation described here facilitates comprehensive analysis rather than intuitive assumptions and basic descriptions of material culture that are typically made in archeological investigations. The new database provides the diagnostic tools to objectively understand the complexities in the material record. The logical next step in this process is to create a portal through which other researchers can interact with our data freely to test our conclusions and to allow them to freely draw on the full breadth of the data that we have collected. This can be achieved through a number of avenues including the DAAHL project and the associated Pottery Informatics Query Database [9,24] that our team is actively developing. This 'portal science' approach has been successfully developed by the NSF GEON project for geosciences [25]. In combination these tools offer a powerful suite to generate, process, and disseminate data digitally.

One way that we are looking to develop alternative models for data sharing and long term storage is through a program, UCSD Library's Research Data Management and Curation Pilot Program, at UCSD in conjunction with the San Diego Supercomputer Center and the UCSD library system [26]. We see this model of data sharing as complementary to using other data sharing initiatives specific to archeology by using the library's platform to share our data with a wider public. The UCSD library's data sharing program is linked directly to other digital repositories that hold digital material from other disciplines. We hope that these kinds of connections of digital data will allow for future interdisciplinary research using archeological data. Additionally, by using the UCSD library to share our digital data, we are able to take advantage of digital object identifiers (DOI) to maintain appropriate copyright control of our data.

In the near future we plan on adding a number of other technologies to the database. Firstly, we need to continue to expand the toolkit to the full range of archeological materials. As already mentioned, in regards to ecofacts, there are still archeological material and data types that we have not had a chance to fully integrate into the system. Constructing the kinds of data tables that specialists of different material cultures can find useful is a long and difficult process because all material culture elements are represented differently from sub-discipline to sub-discipline. After we feel comfortable that we have accounted for the basic kinds of data that are a part of archeological excavation we will begin to implement other tools for diagnostic analysis.

Data from photography from the site and three-dimensional scanning of artifacts are our first priorities for diagnostic tools. We look to develop a number of workflows in conjunction with our work with LiDAR in order to enhance the usability of the LiDAR dataset as a skeleton on which to layer the images. Additionally, we want to integrate the results from analysis of three-dimensional scans of artifacts into our framework [27]. The combination of these tools. More immediately, this project operates as the backend that allows for more complex scientific visualizations that take advantage of the three-dimensional recording of data collected in the field using the GPS and total station methods described above by providing necessary metadata to provide meaningful analysis. As we begin to integrate other sources of diagnostic data into the visualization system in development by our colleagues we anticipate using our framework to generate dynamic visualizations of the archeology as it is excavated. These more complex visualizations can be used both analytically and to better communicate our research as it is in motion [10]. The ability to easily disseminate results in a systemic fashion is a step forward for archeological research away from simply publishing long form monographs that only provide a select picture of excavation results and methodologies.

Acknowledgments

Thanks to Ramesh Rao, Director, Calit2 San Diego Division; Ziad Al-Saad, former Director General, Department of Antiquities of Jordan; Mohammad Najjar, UCSD Levantine Archeology Lab; and the friends and patrons of CALIT2/CISA3 for their support. We also thank the undergraduates from the UCSD Department of Anthropology for their assistance in data entry for this project.

This research was supported by the National Science Foundation under IGERT Award #DGE-0966375, "Training, Research and Education in Engineering for Cultural Heritage Diagnostics," and the UCSD Calit2 Strategic Research Opportunities (CSRO) grant program. Thanks also to the staff the UCSD Research Cyberinfrastructure initiative.

References

[1] L. Isaksen, K. Martinez, G. Earl, Archaeology, formality & the CIDOC CRM, in: Paper Presented at Interconnected Data Worlds: Workshop on the Implementation of the CIDOCCRM, Berlin, Germany, 23–24 Nov. 2009. Retrieved from http://eprints.soton.ac.uk/69707/ (accessed 20.02.12).
[2] S. Jeffrey, J. Richards, F. Ciravegna, S. Waller, S. Chapman, Z. Zhang, The Archaeotools project: faceted classification and natural language processing in an archaeological context, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 367 (2009) 2507–2519.
[3] Woods hole oceanographic institution, Data Management for Marine Geology and Geophysics: Tools for Archiving, Analysis, and Visualization, 2001 [Online]. Available: http://www.whoi.edu/cms/files/WorkshopReport_27505.pdf.
[4] S.P. McPherron, H.L. Dibble, Using Computers in Archaeology: A Practical Guide, McGraw-Hill, Mayfield, Boston, 2002.
[5] http://www.lparchaeology.com/cms/services/ark-archaeological-recording-kit.
[6] B. Frischer, J.W. Crawford, D. Koller (Eds.), Making History Interactive. Computer Applications and Quantitative Methods in Archaeology (CAA). Proceedings of the 37th International Conference, Williamsburg, Virginia, United States of America, March 22–26. BAR S2079, Archaeopress, Oxford, 2009.
[7] K.W. Kintigh, The promise and challenge of archaeological data integration, American Antiquity 71 (3) (2006) 567–578.
[8] M. Baker, C. Fisher, E. O'Riordan, M. Grove, M. Fulford, C. Warwick, M. Terras, A. Clarke, M. Rains, VERA: a virtual environment for research in archaeology, in: 4th International Conference on e-Social Science, National Centre for e-Social Science, Manchester, 2008, pp. 18–20.
[9] T.E. Levy, et al., On-site digital archaeology 3.0 and cyber-archaeology: into the future of the past–new developments, delivery and the creation of a data avalanche, in: M. Forte (Ed.), Cyber-Archaeology, Archaeopress, Oxford, 2010, pp. 135–153.
[10] V. Petrovic, A. Gidding, T. Wypych, F. Kuester, T. DeFanti, T. Levy, Dealing with Archaeology's Data Avalanche: Computational Tools for the Analysis of LiDAR Datasets from A Cyber-Enabled Field Site, IEEE Computer, 12 May 2011, IEEE Computer Society Digital Library, IEEE Computer Society. http://doi.ieeecomputersociety.org/10.1109/MC.2011.161.
[11] http://adaa.ucsd.edu/ArchField/.
[12] http://www.opendig.org/.
[13] http://daahl.ucsd.edu/DAAHL/.
[14] http://pleiades.stoa.org/.
[15] http://ochre.lib.uchicago.edu/.
[16] E. Gay, D. Cooper, B. Kimia, G. Taubin, D. Cabrini, S. Karumuri, W. Doutre, S. Liu, K. Galor, D. Sanders, A. Willis, REVEAL intermediate report, in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPRW, 13–18 June 2010, pp. 1–6. http://dx.doi.org/10.1109/CVPRW.2010.5543548.
[17] http://iadb.org.uk/.
[18] http://www.tdar.org/.
[19] http://archaeologydataservice.ac.uk/.
[20] http://opencontext.org/.

[21] T.E. Levy, R.B. Adams, A. Hauptmann, M. Prange, S. Schmitt-Strecker, M. Najjar, Early Bronze Age metallurgy: a newly discovered copper manufactory in southern Jordan, Antiquity 76 (2002) 425–437.
[22] http://geoserver.org/.
[23] http://www.openlayers.org/.
[24] N.G. Smith, A. Karasik, T. Narayanan, E.S. Olson, U. Smilansky, T.E. Levy, The pottery informatics query database: a new method for mathematic and quantitative analyses of large regional ceramic datasets, Journal of Archaeological Method and Theory (2012) 1–39.
[25] http://www.geongrid.org/.
[26] http://rci.ucsd.edu/.
[27] A. Karasik, U. Smilansky, 3D scanning technology as a standard archaeological tool for pottery analysis: practice and theory, Journal of Archaeological Science 35 (2008) 1148–1168.

Aaron Gidding is a Ph.D. candidate at the University of California, San Diego in the Department of Anthropology, He holds a MA from University College London in the Archeology of the Eastern Mediterranean and Middle East. He is also affiliated with the California Institute for Telecommunications and Information Technology and Scripps Institution of Oceanography. His research interests include data management systems, ancient economic systems, and archeo-magnetism.

Yuma Matsui is a research engineer at Canon Inc. and visiting scholar at the California Institute for Telecommunications and Information Technology (Calit2) at the University of California, San Diego. He received his M.S. degree in Information Science and Technology from the University of Tokyo in 2005. His research interests include data management systems, scalable cyber-infrastructure, software engineering, and interdisciplinary computing.

Thomas E. Levy is Distinguished Professor of Anthropology and holds the Norma Kershaw Chair in the Archeology of Ancient Israel and Neighboring Lands at the University of California, San Diego. A fellow of the American Academy of Arts and Sciences, Levy is a Levantine field archaeologist with interests in the role of technology, especially early mining and metallurgy, on social evolution from the beginnings of sedentism and the domestication of plants and animals in the Neolithic period to Islamic times. Levy directs the cyber-archeology laboratory at the California Institute of Telecommunications and Information Technology (Calit2).

Tom DeFanti, Ph.D., is a research scientist at the California Institute for Telecommunications and Information Technology, University of California, San Diego, and a distinguished professor emeritus of Computer Science at the University of Illinois at Chicago. He is principal investigator of the NSF International Research Network Connections Program TransLight/StarLight project, the NSF GreenLight Instrument project and the KAUST Calit2 OptIPresence Project. He is recipient of the 1988 ACM Outstanding Contribution Award and was appointed an ACM Fellow in 1994. He shares recognition along with EVL director Daniel J. Sandin for conceiving the CAVE virtual reality theater in 1991.

Falko Kuester is the Calit2 Professor for Visualization and Virtual Reality and an associate professor in the Department of Structural Engineering at the Jacobs School of Engineering at UCSD and serves as the director of the Calit2 Center of Graphics, Visualization and Virtual Reality and director of the Center of Interdisciplinary Science for Art, Architecture and Archeology. His research in the field of cyber-archeology is focused on creating innovative new techniques and methodologies for cultural heritage diagnostics and preservation, including diagnostic and analytical imaging as well as visual and cultural analytics in collaborative digital workspaces. He received a Ph.D. in computer science from the University of California, Davis.

Future Generation Computer Systems 29 (2013) 2115–2116

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Editorial

Preface

This special issue contains extensions of work presented at the 2011 e-Science conference, held in Stockholm, Sweden in December 2011. Scientific research is increasingly carried out by communities of researchers that span disciplines, laboratories, organizations, and national boundaries. The e-Science 2011 conference brought together leading international and interdisciplinary research communities, developers, and users of e-Science applications and enabling technologies. In this special issue we highlight selected contributions to the conference that demonstrate the wide applicability of e-Science methodologies and tools and how they become mainstream in many scientific areas. Authors of the best contributions to the conference have been invited to submit an extended version of their work, which then went through the normal review process of the journal. This resulted in the following 8 papers selected for this special issue:

Gidding et al. present ArchaeoSTOR [1], a data curation system for research on the archaeological frontier. Archeology is one of the areas where e-Science methods are increasingly important, particularly when it comes to management of ever increasing amounts of archeological data. The paper specifically presents a data pipeline from data acquisition, tagging, and characterization as well as novel methods for querying archeological data.

Benson et al. studied effective locations and densities of low cost sensors that connect volunteer computers across the world to monitor seismic events in the Quake-Cather Network (QCN) [2]. Using a BOINC emulator they are able to simulate and study diverse sensor densities and seismic scenarios under different geographical and infrastructural constraints.

The role of social networks in e-Science is analyzed by Bubendorfer et al. [3]. They identify two approaches, first, using social networks as an overlay to facilitate collaborative work, targeting both the exchange of information but also the creation of ad-hoc e-Science infrastructures. Second, they identify social networks as a tool to reach out to non-technical users and to encourage them to contribute their computational resources in a volunteer computing style infrastructure.

Bentley et al. [4] discuss the usage of e-Science data infrastructures in Heliophysics, an area that has an increased need to federate data from many different observatories and communities to understand and predict highly energetic events on the Sun. This includes help in locating data and understanding the relevance of data as well as data integration, web portals, and workflows.

Another aspect in data handling is data provenance and the paper by Asuncion [5] presents a new methodology for automated in situ provenance tracking for MS Excel spreadsheet – a tool used by many scientists. Case studies with atmospheric and fishery research groups demonstrate the usefulness of their approach.

Zaki et al. [6] also considered data provenance and developed a user-orientated Electronic Laboratory Notebook (ELN) that allows the automatic capture of metadata during the modeling process. A case study with an atmospheric chemistry community validates their approach.

The high performance computing aspects of e-Science are being tackled by Aguilar et al. [7]. They present a performance characterization and new methods to overcome the identified bottlenecks in the Dalton quantum mechanics/molecular dynamics code. They provide a case study for performance optimization that can be useful to other e-Science applications, too, and present novel hierarchical parallelization strategies.

Parallel computing techniques are also in the center of the paper by Ocaña et al. [8] who designed parallel workflows for phylogenetic analyses. Particularly, they extend the SciHmm workflow used for multiple sequence alignment to phylogenetic analysis. They also analyzed the cost/benefit ration when executing these kinds of workflows on cloud-based environments.

The conference, as well as the selected papers, has demonstrated the wide uptake of e-Science methodologies and tools in many scientific areas and that increasingly tools developed for a specific area are being re-used for related problems in different domains. This clearly shows how e-Science has become an essential core component of modern scientific processes andis seamlessly embedded in many scientific domains.

References

[1] A. Gidding, Y. Matsui, T.E. Levy, T. DeFanti, F. Kuester, ArchaeoSTOR: a data curation system for research on the archeological frontier, Future Generation Computer Systems (2013) http://dx.doi.org/10.1016/j.future.2013.04.007.
[2] K. Benson, S. Schlachter, T. Estrada, M. Taufer, J. Lawrence, E. Cochran, On the powerful use of simulations in the Quake-Catcher Network to efficiently position low-cost earthquake sensors, Future Generation Computer Systems (2013) http://dx.doi.org/10.1016/j.future.2013.04.012.
[3] K. Bubendorfer, K. Chard, K. John, A.M. Thaufeeg, eScience in the Social Cloud, Future Generation Computer Systems (2013) http://dx.doi.org/10.1016/j.future.2013.04.003.
[4] R. Bentley, J. Brooke, A. Csillaghy, et al., HELIO: discovery and analysis of data in heliophysics, Future Generation Computer Systems (2013) http://dx.doi.org/10.1016/j.future.2013.04.006.
[5] H.U. Asuncion, Automated data provenance capture in spreadsheets, with case studies, Future Generation Computer Systems (2013) http://dx.doi.org/10.1016/j.future.2013.04.009.
[6] Z.M. Zaki, P.M. Dew, L.M.S. Lau, et al., Architecture design of a user-orientated electronic laboratory notebook: a case study within an atmospheric chemistry community, Future Generation Computer Systems (2013) http://dx.doi.org/10.1016/j.future.2013.04.011.
[7] X. Aguilar, M. Schliephake, O. Vahtras, J. Gimenez, E. Laure, Scalability analysis of Dalton, a molecular structure program, Future Generation Computer Systems (2013) http://dx.doi.org/10.1016/j.future.2013.04.013.

0167-739X/$ – see front matter
http://dx.doi.org/10.1016/j.future.2013.05.011

[8] K.A.C.S. Ocaña, D. de Oliveira, J. Dias, E. Ogasawara, M. Mattoso, Designing a parallel cloud based comparative genomics workflow to improve phylogenetic analyses, Future Generation Computer Systems (2013) http://dx.doi.org/10.1016/j.future.2013.04.005.

Erwin Laure*
PDC & SeRC, KTH, Sweden
E-mail address: erwinl@pdc.kth.se.

Sverker Holmgren
Department of Information Technology & eSSENCE, Uppsala University, Sweden

Available online 24 May 2013

* Corresponding editor.