Mediterranean Archaeology and Archaeometry, Vol.14, No 4, pp. 109-116

OPENDIG: CONTEXTUALIZING THE PAST FROM THE FIELD TO THE WEB

Matthew L. Vincent*[1,2], Falko Kuester[2] and Thomas E. Levy[1,2]

[1]Levantine and Cyber-Archaeology Laboratory, University of California, San Diego
[2]Qualcomm Institute, University of California, San Diego

Received: 20/01/2014
Accepted: 08/05/2014
Corresponding author: Matthew L. Vincent (mlvincent@ucsd.edu)

ABSTRACT

Data recording is one of the primary requirements of any archaeological project. Some projects rely on the traditional pen-and-paper methods, while others have begun to employ field data recording applications through mobile computing platforms. The former method relies on later transcription of the data, while the later passes over this step, integrating the data from various devices at some later point. Many rely on commercial solutions to solve their data recording needs. Well-known platforms, which have had a long and successful track record with databases, are now being employed for archaeological databases. Although these robust platforms provide straightforward solutions, they are expensive and not easily extensible.

OpenDig was developed with a focus on open source frameworks, with the idea that future expansion would be important for any archaeological database. By utilizing open source tools that were born in the World Wide Web, *OpenDig* provides a complete framework for archaeological data from the field and post-excavation studies. The three main tools that make up the *OpenDig* framework are: 1) a field recording application for describing archaeological contexts, associated photos, geospatial data, and find; 2) a lightweight data reader and editor for deployment in field laboratories; 3) a full web application for a more complete tool set for reviewing, analysing and disseminating these data acquired from the field. Three tools, on their own, may not seem very different from other solutions available to archaeologists today. However, *OpenDig* demonstrates the viability of using open source tools and open source data to create a complete system for data recording, analysis and dissemination. The future of archaeological data lays in finding ways to link disparate data sets from various projects and being able to make sensible comparisons. This can only be achieved by providing open access to these data and creating common interfaces that allow archaeologists to link their data with others.

KEYWORDS: OpenDig, data recording, field archaeology, Cyber-archaeology, archaeology cyber-infrastructure

1. INTRODUCTION

As archaeology is the 'science of destruction,' it depends on careful and meticulous data recording in order to preserve the archaeological contexts where artifacts are found for analyses and modeling. These data become the metadata that describe the context of artifacts, samples and geographic data. Yet, for many projects, the recording of these data takes a secondary place in the field as well as the publication process. Data recording is often done on paper forms that later must be digitized, which is often a tedious process that ends up being neglected. Others have adopted digital field recording methods, but have not found ways to publish them to the Web – an essential process to share and disseminate data. Instead, these data are disseminated to the public only through the lens of interpretation in the form of journal articles, books and other written publications. Although these interpretations are still some of the most important methods for disseminating this information, the primary data must find a place in the publication process, allowing others to use these data as part of their research as well.

This paper looks at *OpenDig*, a framework for archaeological meta-data, and the role it plays in the recording, analysis and dissemination of archaeological data. Furthermore, it seeks to encourage the rapid publication of primary data through machine-readable formats, opening up new avenues of research through open data. Finally, this article encourages an agile approach to archaeological data. Rather than trying to conform disparate archaeological datasets to a single schema, ways should be found to link datasets using Web-based standards proven in the field today.

2. OPENDIG: AN ARHCAEOLOGICAL FRAMEWORK

OpenDig has its foundations in the Madaba Plains Project, namely Tell al-'Umayri, where a recording system was developed with a focus on holistic verbal descriptions that attempt to detail the archaeological context as much as possible (Brower 1989, Clark 2011). The forms include detailed information that describe the type of sediment, building style, related colors and other information (see figure). Unfortunately, such information does not readily lend itself to tabular data, like that found in the widespread **Structured Query Language** (SQL) databases popular today. These databases rely on tables with rows and columns for data storage. Rows are related through common identifiers, allowing the user to compile together rows from various tables to connect data together. The problem presented by these forms was finding a way to correctly represent them in an electronic format, particularly when it required breaking a single document apart into multiple tables.

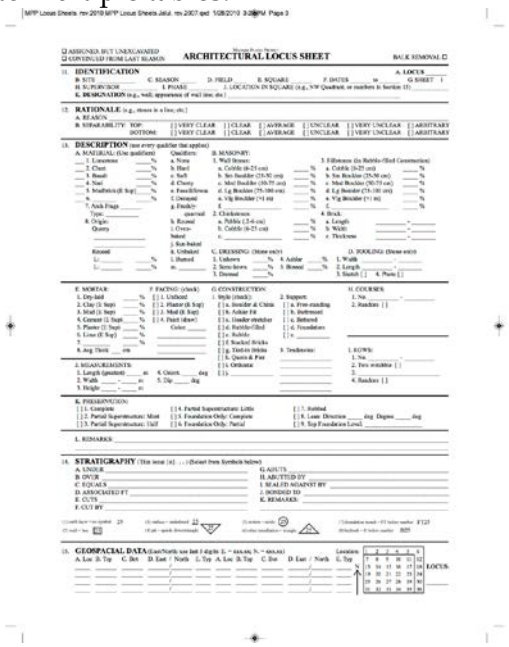
ARCHITECTURAL LOCUS SHEET

Figure 1 An example of the Madaba Plains Project paper based recording system.

2.1 Schemaless Data

Apache's CouchDB (Anderson, Lehnardt, and Slater 2010) is one of the popular NoSQL databases currently available. NoSQL databases, as the name suggests, move away from tabular recording systems and instead store data as documents. The flexible database schemas make it ideal for archaeological recording, as small variances in recording methodologies from site to site make archaeological databases difficult to implement across a varie-

ty of sites. The initial implementation of *OpenDig* across the three sites which make up the Madaba Plains Project (Tall Hisban, Tall Jalul and the aforementioned Tall al-'Umayri) revealed that each site, using the same recording methodologies, had small variations in the interpretations and application of the forms that made it impossible to use the database at each of the three sites. This problem can be solved by using a schema-less database system such as CouchDB.

CouchDB shifts away from the usual columnar databases that require a specific schema for data storage. Unlike its columnar predecessors, CouchDB is a schema-less database which calls itself a "document storage" (Anderson, Lehnardt, and Slater 2010, 4). “Schema-less” refers to xxxxx. Previously, data recorded in the field would be made by hand on a single sheet of paper, which was then digitized across several different tables. Shifting to CouchDB allows one to rethink data recording in such a way that enables the data storage to mimic the field practice; documents are created instead of tables and key-value pairs instead columns. Each document is stored in JSON (JavaScript Object Notation), which is a human-readable markup language similar to XML, but requiring significantly less overhead.

Beyond schematic differences, CouchDB offers a two perks that makes it ideal for use in archaeology. First, it has a replication layer built into the database. This makes deploying the database to any number of devices a painless process. Archaeology in Jordan often takes place in remote locations with little to no access to the Internet. Often researchers will make a copy of the database that they then take to the field as they will not be able to access their primary database back home. This local copy will handle all changes to the data while used in the field and then will overwrite the database after the project returns from the field. This means that the field copy of the database becomes the primary database and all operations at the home location must cease or complex synchronization routines need to be created. CouchDB's replication layer allows for working with multiple copies of the database at one time. The built-in synchronization layer handles any conflict management and will push for "eventual consistency" (Anderson, Lehnardt, and Slater 2010, 11-20) as it handles multiple devices at any given time.

The replication feature also acts as a distributed backup procedure. As long as every device using the database is synchronized on a regular basis, the database will effectively be backed up on each device, and assuming there is an Internet connection present, it can also be replicated to the cloud on a regular basis. This greatly reduces the risk of data loss through the use of cloud and local backups for replication.

The application has developed over time from one single application to three integrated individual applications. In order to manage these applications efficiently, a single file defines all the fields used for recording in the field. Each of the applications then draws from this file to layout the data for display as well as entry. This way, the three applications can easily be modified from one single file, as well as adapted to other projects as each project may have different recording needs.

2.2 Part I: The Web

The first *OpenDig application* was created solely for the web, and initially only to act as a way to publish the data. At that point, the data was still being entered into an Access database and then migrated into the Ruby on Rails web application whenever updates were made. It quickly became apparent that the web was the way forward, and it was decided to move all the data entry to the web, which was done in 2010 for the first time at the Tall al-‘Umayri excavations in Jordan. It was during that season, where all the data entry was done in the field, that the problems with dependable Internet connections became apparent, not to mention that the data entry was challenging for the supervisors who were then overloaded with both field reports and da-

ta curation. These problems inspired the creation of the first mobile *OpenDig* application so that data would be entered directly into the database in the field and cut out the need to transcribe the forms later. The web application also represents the most sophisticated set of tools, but this comes at the cost of a dedicated server that isn't easily deployed to the field.

2.3 Part II: The Field

The original recording system at Tall al-'Umayri depended on the former co-director, Larry Herr, entering all the data from the forms into a Microsoft Access database at the end of the season. Once the data were entered into the database, a processing usually lasting about two months, the database was then burned on to a DVD and distributed via postal mail to the various researchers working on the project. With *OpenDig*, an all digital data entry system streamlines this process by removing the need to transcribe data altogether.

Figure 2 OpenDig on the iPhone.

With paper-based recording systems, there was also a need to have one person verify the data on a weekly basis. This person would routinely check each notebook, making sure all the necessary data were recorded on the forms, and check for any erroneous data along the way. Just as this was a tedious task undertaken by a single individual, this process can be streamlined with *OpenDig* through the implementation of data validations in the application itself.

In order to achieve this, an *OpenDig* application has been developed focused on in-field data entry using native Apple's mobile devices such as the iPhone, iPad, or iPod. Of course, connectivity cannot be guaranteed in the field and therefore the database is built to synchronize with a database at the 'Dig House' after the day's excavations are completed. Data validations can be implemented to verify that correct data is being entered, while a streamlined review process can be put into place for each field supervisor to verify the necessary data has been entered for their excavation units.

2.4 Part III: The Lab

Due to the connectivity problems often faced in remote areas, it is not possible to guarantee that using a remote database would work. Instead, it is necessary to use a local database with which the various field devices can synchronize their data. However, creating a server that is easy to deploy in the field isn't an easy task. Fortunately, since CouchDB comes packaged as a "one-click" install, it is easy to deploy it on any computer in the field. Furthermore, one of the most powerful components of CouchDB is the ability to host and serve applications from within the database itself. Rather than having to setup a complicated server to host the necessary application, CouchDB acts as the server, in which a lightweight data-reader and writer can be placed allowing for staff to have quick access to the excavation data while in camp. While this system doesn't have all the tools found in the main web application, it provides researchers with the ease and comfort of accessing all the excavation data from a browser.

3. AGILE ARCHAEOLOGY

Agile software development (Beck et al. 2001, Martin 2003) has been key in reshaping the software development world. Acknowledging that software development is a fluid, rapidly-changing process (Martin 2003, 1-9) shifted the focus on how software is produced. Rather than trying to do an assembly-line production of software, it should instead be done in iterations, building the basics first and moving on from

there. Archaeology is certainly not software development, but there are lessons to learn from the agile methodologies that can improve how archaeological data are recorded, shared and published.

For years archaeologists have been arguing over the best way to record and share data, the format, the items that should be included, and other details relating to the sharing of archaeological data (Adam Matei, Kansa, and Rauh 2011, Atici et al. 2013, Kansa 2012, Kansa and Kansa 2013, Richards et al. 2011, Richards, Richards, and Robinson 2000, Richards 1997, Schloen 2001). Unfortunately, while these conversations are necessary and enable us to more effectively collaborate, the lack of agreement has also meant that we are still waiting for a standardized data format for archaeology, and instead we see the field in a state of fragmented data formats. However, this does not have to be a problem. The agile methodologies encourage iterations, and publishing our data quickly, even if it isn't in a standardized format, at least gets these data out and available to the research community. As more data becomes available, it should become clear how best we can share and link our data together.

4. CONNECTING DATA

With disparate datasets available for research today, the question is how to link these data together to create holistic datasets for more complete research opportunities. Concepts, such as the semantic web (Berners-Lee, Hendler, and Lassila 2001) provides the ideal framework for linking these disparate datasets. Richards (Richards 2006, 977) rightly points out that using the semantic web to link archaeological data (although Richards is primarily referring to publications) still requires an agreed ontology. This holds true for archaeological datasets, but these ontologies need only be a few common fields that can be found among all archaeological datasets. Furthermore, these ontologies need only be common linking descriptions such as geographical location or chronological time period. The other data can then be returned according to a search, leaving the researcher to make the final determinations regarding the relevance of these data within their greater research. As previously said, agile methodologies adapted to archaeology dictate that the primary focus is to publish our data, even if we haven't agreed on a common ontology. These ontologies can be developed and added to existing datasets as it becomes clear what these ontologies should be, particularly if we have many datasets available to see what the best ontologies might be.

4. THE UCSD CYBER-ARCHAEOLOGY ECOSYSTEM, AN EXAMPLE

The Levantine and Cyber-archaeology Lab at the University of California, San Diego has focused on a geographic-centric recording system since 1999 (Levy et al. 2001). Since then, custom software has been developed to handle the archaeological data in the field, lab as well as long term research, dissemination and publication. Because of these developments, the UCSD Cyber-archaeology Lab serves as an example of both agile archaeology as well as how we might move forward towards linked archaeological data. Two additional systems, ArchField and ArchaeoSTOR, will be highlighted here and then a brief description of how these systems are being integrated.

4.1 ArchField

ArchField (Smith and Levy 2012) is a system for the real-time recording and visualization of geographic data in the field. Connecting directly to total stations or GPSs, the system allows the field archaeologist to record data directly to a laptop or handheld device, visualizing and editing it in real-time, reducing the need for post-processing in the lab after the excavations have been completed for the day. By having these data available to the researcher directly in the field, they can see what data might be missing and what data they may need to correct before leaving the field. This allows for greater accuracy for the ge-

ographical data by reducing the time between acquisition and visualization. The system uses PostGIS, a SQL database with geospatial extensions allowing for the storage, indexing and retrieval of geographic data, to handle the geographic and metadata, which is then synchronized with a master database back at the lab at the end of the excavation day.

4.2 ArchaeoSTOR

ArchaeoSTOR (Gidding et al. 2013) grew out of the need to organize and manage artifacts, different data file formats and related data. It became apparent that the quantity of artifacts and samples being managed by the UCSD excavations was becoming increasingly difficult to manage using traditional methods. *ArchaeoSTOR* unified all these datasets in one place, allowing the team to quickly and easily locate and manage the artifacts, samples and analyses. By creating a management system, artifacts can be found easily, analytical data can be attached and studied quickly efficiently.

4.3 Connecting the Systems

ArchField, *ArchaeoSTOR* and *OpenDig* make up the three principle systems being used in the field by UCSD's excavations. Each of the three components is independent and does not depend on the other. However, in order to conduct holistic research, all of these data are being incorporated into a single system allowing access to all three datasets seamlessly. Connecting these three distinct systems isn't problematic if it is approached from the perspective of machine-readable data. Using Application Programming Interfaces (APIs), the three systems can be easily connected using common data. For example, similar data describing the locus context is common among all three systems. Therefore, one can simply query a single system, retrieving the data from all three systems, geographic data defining the where of the archaeological data; OpenDig describing the archaeological context of these data; and *ArchaeoSTOR* accessing the artifacts which rely on the previous two for the context. This was carried out on three different excavation sites for the Edom Lowlands Regional Archaeology Project (ELRAP) in 2011 and 2012 (Levy et al.)

5. CONCLUSION

The problem of connecting disparate archaeological data collected in the field does not have to be the challenge it has been made out to be. Using agile methodologies, publishing archaeological data as soon as possible to the Web, allows researchers to begin linking archaeological data right away. Perhaps it will require more work since researchers will have to find common ontologies, however it will push the field forward as primary data is published and linked, even with additional work involved at the moment. As these data are published, common ontologies will become apparent, adding to the conversation for an archaeological data standard. Furthermore, once an archaeological data standard is agreed upon, these already published data can be updated to reflect any new standards that might be adopted in the future. In the mean time, proposed standards such as ArchaeoML (Schloen 2001) or tDAR (Plaza 2013, Kansa et al.) can be adopted as a way to bridge the gap and find common ontologies.

6. ACKNOWLEDGEMENTS

Some of this work was supported by the National Science Foundation under IGERT Award #DGE-0966375, "Training, Research and Education in Engineering for Cultural Heritage Diagnostics", awarded to Matthew L. Vincent, who is also grateful to Thomas E. Levy for allocating his NSF travel funds to support his travel to Delphi for the Virtual Archaeology conference.

REFERENCES

Adam Matei, Sorin, Eric Kansa, and Nicholas Rauh. (2011) The Visible Past/Open Context Loosely Coupled Model for Digital Humanities Ubiquitous Collaboration and Publishing: Collaborating Across Print, Mobile, and Online Media. *Spaces & Flows: An International Journal of Urban & Extra Urban Studies* no. 1 (3), pp. 33-48

Anderson, Chris, Jan Lehnardt, and Noah Slater. (2010) *CouchDB: The Definitive Guide: The Definitive Guide*: O'Reilly Media.

Atici, Levent, SarahWhitcher Kansa, Justin Lev-Tov, and EricC Kansa. (2013) Other People's Data: A Demonstration of the Imperative of Publishing Primary Data. *Journal of Archaeological Method and Theory* no. 20 (4):663-681. doi: 10.1007/s10816-012-9132-9.

Beck, Kent, Mike Beedle, Arie Van Bennekum, Alistair Cockburn, Ward Cunningham, Martin Fowler, James Grenning, Jim Highsmith, Andrew Hunt, and Ron Jeffries. Manifesto for agile software development, http://agilemanifesto.org/, Accessed 5/12/2013.

Berners-Lee, Tim, James Hendler, and Ora Lassila. (2001) The semantic web. *Scientific american* no. 284 (5):28-37.

Brower, James K. (1989) Archaeological Excavation Data Management System. In *Madaba Plains Project: The 1984 season at Tell el-Umeiri and vicinity and subsequent studies*, edited by L.T. Geraty, LG Herr, OS LaBianca and RW Younker, 387-401. Berrien Springs, MI: Andrews University Press.

Clark, Douglas R. (2011) *The Madaba Plains Project : forty years of archaeological research into Jordan's past*. Sheffield England ; Oakville, CT: Equinox Pub.

Gidding, A., Matsui, Y., Levy, T. E., DeFanti, T., & Kuester, F. (2013). ArchaeoSTOR: A data curation system for research on the archeological frontier. *Future Generation Computer Systems*, *29*(8), 2117-2127.

Kansa, Eric. (2012) Openness and archaeology's information ecosystem." *World Archaeology* no. 44 (4):498-520. doi: 10.1080/00438243.2012.737575.

Kansa, Eric C, and Sarah Whitcher Kansa. (2013) We All Know That a 14 Is a Sheep: Data Publication and Professionalism in Archaeological Communication. *Journal of Eastern Mediterranean Archaeology and Heritage Studies* no. 1 (1) : 88-97.

Kansa, Eric C, Sarah Whitcher Kansa, Francis P McManamon, Keith W Kintigh, Adam Brin, and Andrea Vianello. (2010) Digital Antiquity and the Digital Archaeological Record (tDAR): Broadening Access and Ensuring Long-Term Preservation for Digital Archaeological Data. http://csanet.org/newsletter/fall10/nlf1002.html, Accessed 5/12/2013.

Levy, TE, JD Anderson, M Waggoner, N Smith, A Muniz, and RB Adams. (2001) Interface: Archaeology and Technology–Digital Archaeology 2001: GIS-Based Excavation Recording in Jordan. *The SAA Archaeological Record* no. 1:23-29.

Levy, Thomas E, Mohammad Najjar, Aaron D. Gidding, Ian W. N. Jones, Kyle A. Knabb, Kathleen Bennallack, Matthew L. Vincent, Alex Novo Lamosco, Ashley Richter, and Craig Smitheram. The 2011 Edom Lowlands Regional Archaeology Project (ELRAP): Excavations and Surveys in the Faynan Copper Ore District, Jordan. *Annual of the Department of Antiquities of Jordan (in press)*.

Martin, Robert Cecil. (2003) *Agile software development: principles, patterns, and practices*: Prentice Hall PTR.

Plaza, David Michael. (2013) The Anasazi Origins Project Digital Archives Initiative: Transferring a Legacy Dataset to a Living Document Using tDAR, http://core.tdar.org/document/391353, Accessed 5/12/13.

Richards, Julian. (2006) Archaeology, e-publication and the semantic web. *Archaeology* no. 80 (310):970-979.

Richards, Julian C, Julian Richards, and Damian Robinson. (2000) *Digital archives from excavation and fieldwork: a guide to good practice*: Oxbow Books Ltd.

Richards, Julian D. (1997) Preservation and re-use of digital data: the role of the Archaeology Data Service. *Antiquity* no. 71 (274):1057-1059.

Richards, Julian, Stuart Jeffrey, Stewart Waller, Fabio Ciravegna, Sam Chapman, and Ziqi Zhang (2011) The Archaeology Data Service and the Archaeotools project: faceted classification and natural language processing. *Archaeology 2.0: New Approaches to Communication and Collaboration* : 31-56.

Schloen, J.D. (2001) Archaeological data models and web publication using XML. *Computers and the Humanities* no. 35 (2):123-152.

Smith, Neil G, and Thomas E Levy. (2012) Real-time 3D archaeological field recording: ArchField, an open-source GIS system pioneered in southern Jordan. *Antiquity* no. 86 (331). http://antiquity.ac.uk/projgall/smith331/, Accessed 5/12/2013