Preserving an Evolving Collection: "On-The-Fly" Solutions for *The Chora of Metaponto* Publication Series

Jessica Trelogan
Institute of Classical Archaeology
University of Texas at Austin
3925 W. Braker Lane
+1 (512) 232-9317
j.trelogan@austin.utexas.edu

Maria Esteva
Texas Advanced Computing Center
University of Texas at Austin
J.J. Pickle Research Campus
+1 (512) 475-9411
maria@tacc.utexas.edu

Lauren M. Jackson
Institute of Classical Archaeology
University of Texas at Austin
3925 W. Braker Lane
+1 (512) 232-9322
lmjackson@utexas.edu

ABSTRACT
As digital scholarship continues to transform research, so it changes the way we present and publish it. In archaeology, this has meant a transition from the traditional print monograph, representing the "definitive" interpretation of a site or landscape, to an online, open, and interactive model in which data collections have become central. Online representations of archaeological research must achieve transparency, exposing the connections between fieldwork and research methods, data objects, metadata, and derived conclusions. Accomplishing this often requires multiple platforms that can be burdensome to integrate and preserve. To address this, the Institute of Classical Archaeology and the Texas Advanced Computing Center have developed a "collection architecture" that integrates disparate and distributed cyberinfrastructure resources through a customized automated metadata platform, along with procedures for data presentation and preservation. The system supports "on-the-fly" data archiving and publication, as the collection is organized, shared, documented, analyzed, and distributed.

General Terms
Institutional opportunities and challenges; Infrastructure opportunities and challenges; Frameworks for digital preservation; Preservation strategies and workflows.

Keywords
Archaeological data; database preservation; collection architecture.

1. INTRODUCTION
In archaeology, as in many disciplines, digital scholarship continues to transform the research process at every stage, from the collection of primary data on site, through post-excavation study and analysis, to the final interpretation and publication of results. A major effect of this transformation is the drive to publish full data collections in addition to print (or electronic) books. The printed monograph, traditionally considered the ultimate goal and the "definitive word" of any academic archaeological project, is giving way to an open, online, and interactive model that reflects a larger continuum of interpretation and reinterpretation. To represent and preserve archaeological research in this way, complex technical infrastructures and services are needed to support and provide fail-safes for data and multiple, simultaneous functions throughout a project's lifecycle. Storage, access, analysis, presentation, and preservation must be managed in a non-static, non-linear fashion within which data evolve into a collection as research progresses. In this context, data curation happens *while research is ongoing*, rather than at the tail end of the project, as is often the case. Such data curation may be accomplished within a distributed computational environment, as researchers use storage, networking, database, and web publication services available across one or multiple institutions.

Ongoing data curation can be burdensome and costly, and, until recently, there has been little professional incentive to do it [1]. Facilitating long-term access to a project's full set of primary data along with evidence for the processes of data collection, analysis, and interpretation promotes reproducibility and data reuse, but is not a trivial goal [2][3]. Whereas print publications end up in a library's custody, in this new model, maintenance and preservation not only of a project's data, but also of its mode of presentation falls, in many instances, to the research unit, requiring a post-custodial approach [4]. This is especially so when data publication requires more sophisticated technical resources than the average institutional repository can provide. This can include, as in the example we present here, web services and database and GIS technologies. Such requirements imply the backdrop of a solid infrastructure and a commitment to its long-term maintenance, and can require researchers to rethink data-intensive projects, reach out for expertise, cobble together adequate resources, and to implement more than one digital preservation strategy.

The Institute of Classical Archaeology (ICA) [5] is in the midst of a major program of study, synthesis, and publication related to long-standing field projects in the chora (countryside) of Metaponto [6]. For this initiative, a dispersed, multidisciplinary, and international team needs access to the legacy collection, a place to incorporate and share up-to-date versions of current work, a stable technical platform for managing data, and a space for continuing dialog throughout. With the Texas Advanced Computing Center (TACC) [7], which provides computational resources and expert data services to the University of Texas System and at the national level, we have implemented an infrastructure solution to accomplish those goals, while facilitating data curation tasks that will ensure the collection's preservation. In addition to storage, preservation, and computational resources at TACC, we leverage file sharing services provided by the University's Academic Technology

Support (ATS) group [8] and web services provided by Liberal Arts Instructional Technology Services (LAITS) [9], which hosts ICA's websites, including Wordpress-based digital companions to the print books (see below). We call this distributed infrastructure a "collection architecture." It integrates domain-specific technical resources and procedures customized to represent ICA's specific research processes and results.

ICA's collection is actively evolving simultaneously in different development stages of active research, publication, and archiving. Acknowledging that data in active projects are most vulnerable to disorganization and loss, and recognizing the importance of prompt archiving, access, and reuse, we consider preservation to be a constant activity that starts from the moment data are created and lasts throughout the collection's continuum.

2. THE ICA COLLECTION

ICA's data collection represents over forty years of research activities carried out since its establishment in 1974. Like any archaeological collection with such a long history, it reflects a dizzying number of technological, methodological, and theoretical changes that have influenced the field of archaeology and associated disciplines since the mid-1970s. It includes many types of data from a multitude of disciplines, from scans of analog photography, original drawings, and field notes to GIS data, born-digital imagery, full publications, and complex relational databases, each with its own set of methods, research questions, and technological requirements.

Currently ca. 5TB in size, the collection consists of data from more than twelve multi-year field projects in southern Italy and Ukraine, and a full range of associated specialist studies. It is growing rapidly as ICA's large physical archive is digitized and as new studies are conducted in support of the publication series. It is also riddled with duplication and redundancy [10], reflecting the recordkeeping habits and collected data silos from a huge, revolving team of people.

3. COLLECTION ARCHITECTURE

Over the course of the last six years, ICA and TACC developed the collection architecture presented here (Figure 1), which leverages existing storage, computing, cloud, and networking resources at the University of Texas at Austin [11][12]. The system enables data sharing and archiving "on-the-fly," as the collection is organized, documented, and analyzed during study and publication. These activities happen in parallel and behind the scenes in the collection architecture, which is distributed across major computational resources within the University. We have implemented services that include a GIS server and a set of web-based databases and Wordpress sites associated with each of ICA's archaeological projects. Metadata—extracted automatically where possible—fulfills data integration and preservation roles, and multiple preservation strategies assure data integrity and security throughout research stages and infrastructure components.

Mapped onto the collection architecture, an overview of our workflow is as follows. Messy legacy and new incoming data are first sorted by ICA research staff into broad categories in hierarchically labeled folders (the recordkeeping system), within a networked file share that functions as a staging area. These general categories (see Figure 2) provide basic descriptions, provenance, and context to data objects and help sift the collection into manageable chunks that relate to specific sites or specialist studies. Roughly organized data are then moved to a secure, geographically replicated storage resource (Corral with iRODS), where they are given unique identifiers. Thus, notably, data are archived at the outset, before further value is added to them through specialist study. From the archive, data objects are shared with the rest of the research team via web services through a web-based, domain-specific, GIS-enabled database (see ARK section, below). From here, the team studies the fully contextualized collection and adds further descriptions and connections as interpretations develop. The architecture allows the archaeological team to focus on research and publication activities, while metadata integration and preservation happens simultaneously in the background. To facilitate data sharing and to complement the print publication series, the Wordpress sites provide a guided entry point for unfamiliar users to navigate the data collection within the database. In addition, they provide access to original field notebooks and intermediary grey literature that cannot be presented in print and are beyond the scope of the database. Thus, each component of the architecture has a unique function, described in detail below, and all the data are preserved.

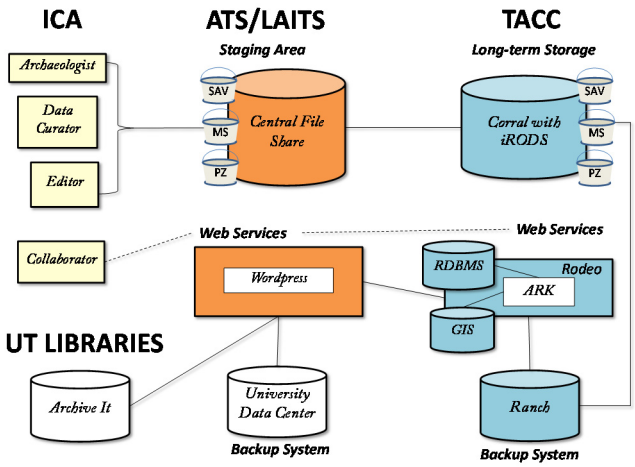

Figure 1. Collection architecture.

3.1 Staging Area and Recordkeeping System

Incoming data are moved into the collection architecture after being roughly sorted in the recordkeeping system within the central file share hosted by ATS. This recordkeeping system consists of a hierarchical file structure and naming conventions for various data types (Figure 2), which entail a neutral set of categories that are general enough to preserve vestiges of old recording methods and technologies, but also descriptive enough to make the collection navigable and reusable. The system is considered as a set of "big buckets" [13], the labels of which are used as descriptive metadata. In turn, the label terms have been mapped to the Dublin Core metadata standard [14] and are automatically extracted for every file as it moves from the file share to the storage resource, Corral with iRODS [15]. To preserve the integrity of the collection in terms of the fundamental archaeological principles of context and provenance, relationships between data objects and the sites and artifacts they represent are automatically captured from the recordkeeping system and recorded as metadata within Corral/iRODS.

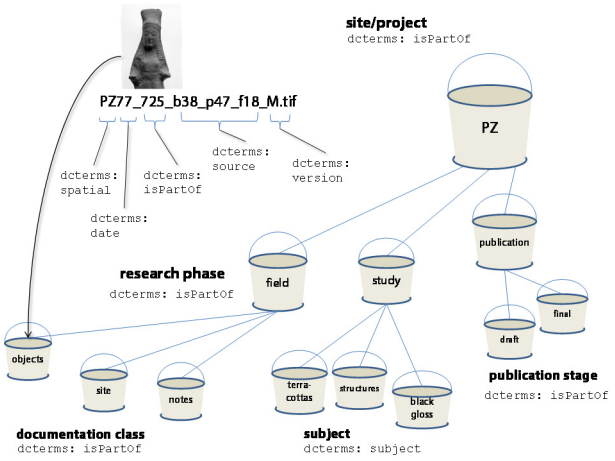

Figure 2. Recordkeeping system, implemented within the staging area and mirrored for long-term storage in Corral.

3.2 Corral with iRODS

Corral is a high performance storage system, geographically replicated, continuously monitored for security and failure, and available 24/7. It is part of the University of Texas System Research Cyberinfrastructure Initiative [16], which provides for its maintenance and expansion and subsidizes its cost. It is available to researchers in the UT System, who may have an initial allocation of 5TB of data for free. Corral uses iRODS as a data broker and rule engine, through which we enable—at ingest—automatic extraction of technical metadata along with descriptive metadata embedded in the file and recordkeeping system folder names. A checksum is also calculated for each file as part of the ingest process. This metadata gets registered in the iRODS iCAT metadata catalog for each file and is also formatted as a METS/Dublin Core/PREMIS file, stored along with the data object in Corral/iRODS. This automation provides documentation for every data object, its provenance, and relationships with other data objects and concepts *without any manual data entry* by the curators [15][12]. The data storage provides a long-term preservation solution for the primary data, which we refer to as the "archival instance" of the collection. Data are deposited here, documented, and preserved "on-the-fly," independent of their selection for further study or publication. The metadata gathered at this instance are preserved and integrated into ARK, the online database described below, to help users navigate the collection during study and make data reuse possible in the future. It also ensures the collection's integrity and helps reduce duplicated effort by providing a system of version control and tracking for each individual data object. This archival instance ensures the preservation of individual data objects and their metadata, acting as a fail-safe should any of the other components of the architecture (e.g., the online publication component) fail.

3.3 ARK (the Archaeological Recording Kit) and Rodeo

From the "archival instance" on Corral/iRODS, data objects and their Dublin-Core-mapped metadata are ingested into a web-based database built on the Archaeological Recording Kit (ARK), a pre-fabricated, open-source system [17] that required little extra investment in web development. ARK resides in Rodeo [18], TACC's cloud computing resource. Rodeo hosts a variety of databases and web services for the UT community in Virtual Machines (VM), allowing for fully customized computational environments and easy access to stored data from any location.

ARK's customizable structure and interface can be easily deployed for all of the varied archaeological projects[1] that are part of the Metaponto series—including excavation, survey, conservation projects, and museum exhibits—facilitating collaborative study and providing a central location for the international team to add details and make additional connections between related objects (Figure 3).

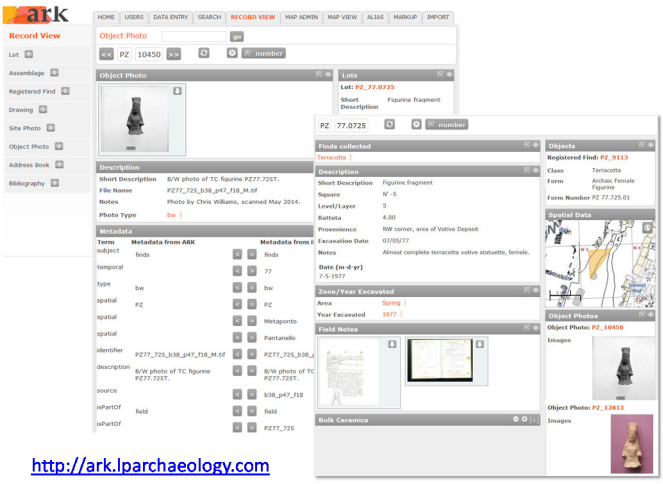

Figure 3. ARK screenshots: photograph stored in Corral/iRODS, metadata extracted from the recordkeeping system, the artifact's context within an excavation unit.

This part of the collection, which we refer to as the "study and presentation instance," also feeds directly into publication workflows by allowing the publication team direct access to artifact and site data as well as high-quality, original photographs and illustrations. More detailed metadata (dating, quantifications, typologies, etc.) can be entered here throughout study and pushed back to the persistent metadata storage system on Corral, so that at any point within the system, there is a full and up-to-date metadata record for each digital object. The evolving archive is thus constantly advancing, providing the basis for related studies, but is always secure. Once a project is complete and published, the ARK database is opened for public access and, via a persistent identifier (DOIs), the organized and fully-documented collection is ensured a permanent home for future access and further inquiry. For the presentation instances of *The Chora of Metaponto* series, we have configured one implementation of ARK per archaeological project. Each of these may have its own particular mode of presentation and contains its own set of data tables in ARK's database.

[1] "Projects" in this case may refer to any of ICA's excavation or surface survey campaigns. Each of these projects may contain more than one excavated site and may refer to more than one print monograph. ARK's flexibility means allows for a different configuration within each ARK instance, depending on the main unit of inquiry (e.g., the "site" in a surface survey, or the "stratigraphic unit" and "artifact" in an excavation).

3.4 Ranch

Ranch is TACC's massive tape-based, long-term storage system. Within our collection architecture, it is used as a high-reliability backup system for the study and publication instance of the collection. Here, we store routine backups of the ARK code base and custom configurations (see Preservation Strategies section below). Across Corral and Ranch, the entire collection architecture is replicated for high data availability and fault tolerance.

4. PUBLICATIONS INFRASTRUCTURE

4.1 Print Publications

The collection architecture functions as the data resource used during the publication process. Thus, specialist studies and interpretations, informed by and incorporated into ARK, either culminate in monographs within *The Chora of Metaponto* publication series, or appear as stand-alone articles, presentations, or grey literature reports. Since overall site interpretation relies upon the primary field documentation as well as dating and contextual information provided by multiple authors, constant access to full and up-to-date data via ARK expedites the creation of an accurate manuscript that reflects a cohesive understanding of the site or project.

4.2 Online Publications

A set of Wordpress-based websites serve as digital companions to the print publication series and as a portal to the data collections housed in ARK [19]. This service is hosted by LAITS as part of their remit to support faculty and staff research projects. The websites can either stand alone as a guided entry point to the data collection or to expand and complement interpretations presented in print. They also provide space to share full-resolution scans and transcripts of field notebooks, grey literature, and specialist reports related to the project. The blog platform's comment section permits immediate discussion and questions that can be directly connected to the original narrative in print, allowing the static interpretation to evolve with further research and input.

5. PRESERVATION STRATEGIES

Preservation is a key function requiring the implementation of more than one preservation strategy across the different infrastructure resources.

5.1 Integration of Data Objects and Metadata

All primary data objects are preserved in Corral/iRODS along with complete technical and descriptive metadata extracted at ingest. These are referred to via URIs within the ARK system, so that if users request a download of the original object, it comes directly from the archival instance on Corral along with its associated METS/PREMIS/DC record. When selected objects are called from the archive into ARK, a thumbnail is generated and descriptive metadata from the iRODS iCAT database populates basic information fields for that record. In turn, if extra descriptive metadata is added through the ARK interface during study, it is pushed back into the iCAT database. Thus, all the primary data and complete metadata are geographically replicated in case of failure of either component in the architecture.

5.2 Databases and Virtualization

While the complete Rodeo system that hosts the databases and the web code is backed up on a daily basis, such backups do not account for the specific workflows, data entry, and usage of individual projects. Thus, we implemented a customized database security and preservation strategy that could handle our ongoing publication production workflows and interfaces. To lower security risks, ARK's database is on one virtual machine, and its web code on another. By separating the database from the public access system we intended to avoid malicious breaches to the site's security. We created an automatic script to initiate daily SQL dumps of the ARK database tables, which are kept in a cascade: one a day for a week, one a week for a month, one a month for a year, and then one a year after that [20]. Additionally, virtualization was implemented as a preservation strategy in which the entire ARK database system running on the VM in Rodeo has a snapshot taken every night at 10 pm. This includes the accumulated SQL files that are produced earlier in the day. The resultant zip file is sent to the backup system in place on Ranch (see Ranch section, above) where we keep three days in a row and two months of backup files. This redundant approach avoids risks such as, for example, the unlikely corruption of files that could result from database writes happening at the same moment the database is snapshotted.

5.3 Wordpress Sites

LAITS provides cascading backups for files stored in the central file share and of the content of the Wordpress sites, with the latest versions discarded after 90 days. This type of backup is designed for disaster recovery as opposed to preservation of evolving interpretation. For this, we use the Archive IT service [22], sponsored by the UT Libraries, to archive snapshots of the Wordpress sites over time. At this time and until the publication is finalized we have scheduled monthly snapshot of the sites (e.g., http://wayback.archive-it.org/5446/20150508134828/ http://metaponto.la.utexas.edu/#).

6. CONCLUSIONS

Archaeological data are inherently vulnerable. Not only is excavation a destructive process, leaving the documentation the only remaining evidence of a site as it is uncovered, but archaeological collections can present serious data preservation challenges during and after a project. These collections tend to be accumulated and studied over decades, are especially large and complex, and reflect a huge range of technical sophistication.

In this project, we perceive preservation as an ongoing activity, which happens throughout the research process and continues well beyond a project's lifecycle into long-term maintenance of the published datasets. Data that are well organized, well documented, and authenticated from the beginning of the project are less vulnerable. We use a distributed set of diverse resources within which we are able to organize, describe, integrate and share data while archiving behind the scenes. In this system, raw, in-progress, and finalized data and publications are constantly secured using a variety of preservation strategies relevant to the different functions and technologies supporting the collection.

The solutions presented here have gone a long way toward streamlining ICA's publication and data sharing efforts and have ensured that a vulnerable collection is archived from the earliest stage possible. By leveraging existing University resources and expertise, the ICA team has been able focus on what it does best—archaeological research—and on enhancing the presentation of our results to provide more sophisticated interactive experiences to our target audiences. The next phase of our work will focus on issues of data reuse. The University of Texas Library supports the use of DOIs and ARKs (archival resource keys) [21], which we have begun minting for our data collections.

The administration and maintenance of the systems through TACC, ATS, and LAITS, are handled by people with the appropriate expertise. Nevertheless, implementing and maintaining this distributed infrastructure required extensive involvement and a learning curve for the domain expert data curators. For similar projects with large legacy collections in a push to publish a backlog of material, an "on-the-fly" approach like the one we present here can help alleviate the burden involved in making data comprehensible and reusable, while simultaneously preserving it as research progresses.

A major challenge that arises with the post-custodial approach adopted here, especially for grant-funded units like ICA, is to find an institution that can commit to maintain the fully functioning and dynamic set of ARK databases and the associated Wordpress sites for the long term. At the same time, thanks to this post-custodial approach, we know we have created a sustainable, well-documented platform that will make it easy to transfer once we do find such a host. Meanwhile, the metadata-ready archive can be deposited in an archaeological repository or at the UT Libraries as a static collection.

7. ACKNOWLEDGEMENTS

We would like to thank the Packard Humanities Institute (PHI) for enabling this research and for its continued generous support of *The Chora of Metaponto*, *The Chora of Croton,* and *Chersonesan Studies* publication series.

8. REFERENCES

[1] Kansa, E. C., and Kansa, S. W. 2011. Toward a Do-It-Yourself Cyberinfrastructure: Open Data, Incentives, and Reducing Costs and Complexities of Data Sharing. In *Archaeology 2.0: new approaches to communication and collaboration*. E. C. Kansa, S. W. Kansa, and E. Wattrall, Eds. 57–92. Los Angeles: Cotsen Institute of Archaeology Press.

[2] Borgman, C. L. 2012. The conundrum of sharing research data. *J. Am. Soc. Inf. Sci.* 63: 1059–78. DOI= http://dx.doi.org/10.1002/asi.22634.

[3] Frank, R. D., Yakel, E., and Faniel, I. M. 2015. Destruction/reconstruction: preservation of archaeological and zoological research data. *Journal of Archival Science* 2015-01-11: 1–27. DOI= http://dx.doi.org/10.1007/s10502-014-9238-9.

[4] Henry, L. J. 1998. Schellenberg in Cyberspace. *American Archivist* 61.2: 309–27.

[5] *Institute of Classical Archaeology (ICA)*. Accessed 10 August 2015. http://www.utexas.edu/cola/ica/.

[6] Carter, J. C, ed. 1998–2014. *The Chora of Metaponto*. Vols. 1–5. Austin: University of Texas Press.

[7] *Texas Advanced Computing Center (TACC)*. Accessed 13 August 2015. https://www.tacc.utexas.edu/home.

[8] *University of Texas at Austin Academic Technology Support (ATS)*. Accessed 15 September 2015. https://www.utexas.edu/transforming-ut/shared-services/ats.

[9] *College of Liberal Arts Instructional Technology Services (LAITS)*. The University of Texas at Austin. Accessed 15 September 2015. http://www.utexas.edu/cola/laits/.

[10] Arora, R., M. Esteva, and J. Trelogan. 2014. Leveraging High Performance Computing for Managing Large and Evolving Data Collections. *International Journal of Digital Curation* Vol. 9, No. 2: 17–27. DOI= http://dx.doi.org/10.2218/ijdc.v9i2.331.

[11] Esteva, M., Trelogan, J., Rabinowitz, A., Walling, D., and Pipkin, S. 2010. From the site to long-term preservation: a reflexive system to manage and archive digital archaeological data. In *Archiving 2010. Proceedings of the Archiving Conference, Vol. 7 (Den Haag, the Netherlands, June 1–4, 2010)*. 1–6.

[12] Kulasekaran, S., Trelogan, J., Esteva, M., and Johnson, M. 2014. Metadata Integration for an Archaeology Collection Architecture. In *Proceedings of the International Conference on Dublin Core and Metadata Applications (Austin, Texas, 8–11 October 2014)*. 53–63. Retrieved from http://dcpapers.dublincore.org/pubs/article/view/3702.

[13] Cisco, S. 2008. Big buckets for simplifying records retention schedules. *ARMA International's Hot Topic 2008*: 3–6. Retrieved 9 May 2014 from http://www.emmettleahyaward.org/uploads/Big_Bucket_Theory.pdf.

[14] Dublin Core Metadata Initiative. DCMI Specifications. Accessed 20 April 2015. http://dublincore.org/specifications/.

[15] Walling, D., and Esteva, M. 2011. Automating the Extraction of Metadata from Archaeological Data using iRods Rules. *International Journal of Digital Curation* Vol. 6, No. 2: 253–64. DOI= http://dx.doi.org/10.2218/ijdc.v6i2.201.

[16] The University of Texas System Research Cyberinfrastructure (UTRC). Accessed 13 April 2015. http://www.utsystem.edu/offices/health-affairs/utrc/storage.

[17] Eve, S., and Hunt, G. 2008. ARK: a developmental framework for archaeological recording." In *Layers of Perception: Proceedings of the 35th International Conference on Computer Applications and Quantitative Methods in Archaeology (CAA) (Berlin, Germany, April 2–6, 2007)*. A. Posluschny, K. Lambers, and I. Herzog, Eds. Kolloquien zur Vor- und Frühgeschichte Band 10. Bonn: Dr Rudolf Habelt GmbH. Retrieved from http://proceedings.caaconference.org/paper/09_eve_hunt_caa2007/.

[18] Rodeo. 2014. Retrieved 14 August 2014 from https://www.tacc.utexas.edu/resources/data-storage/#rodeo.

[19] Institute of Classical Archaeology. The Chora of Metaponto: a digital companion to the publication series. Accessed 20 April 2015. http://metaponto.la.utexas.edu.

[20] Preserving Relational Databases. 2015. Retrieved 20 April 2015. http://digital.humanities.ox.ac.uk/Support/PreservingDatabases.aspx.

[21] California Digital Library. EZID. Accessed 13 April 2015. http://ezid.cdlib.org.

[22] Archive-It -Web Archiving Services for Libraries and Archives. Retrieved 29 April 2015. https://archive-it.org/.