NATIONAL ENDOWMENT FOR THE
Humanities

DIVISION OF **PRESERVATION AND ACCESS**

# Narrative Section of a Successful Application

The attached document contains the grant narrative of a previously funded grant application. It is not intended to serve as a model, but to give you a sense of how a successful application may be crafted. Every successful application is different, and each applicant is urged to prepare a proposal that reflects its unique project and aspirations. Prospective applicants should consult the NEH Division of Preservation and Access application guidelines at http://www.neh.gov/divisions/preservation for instructions. Applicants are also strongly encouraged to consult with the NEH Division of Preservation and Access staff well before a grant deadline.

Note: The attachment only contains the grant narrative, not the entire funded application. In addition, certain portions may have been redacted to protect the privacy interests of an individual and/or to protect confidential commercial and financial information and/or to protect copyrighted materials.

Project Title: Beyond Management: Data Curation as Scholarship in Archaeology

Institution: Alexandria Archive Institute

Project Director: Sarah Whitcher Kansu

Grant Program: Research and Development

## 3.1 SIGNIFICANCE

The Alexandria Archive Institute (AAI) seeks Tier II advanced implementation funding for applied research that will facilitate digital preservation of, and access to, archaeological field collections. To build upon prior successes, the current project will conduct methodologically rigorous qualitative studies in researcher data creation and reuse practices. Better understanding of the relationships between data creation and preservation and reuse will guide development of data collection tools and methods and new publishing services for Open Context (http://opencontext.org), an open access data publishing venue for archaeology. Understanding researcher data needs will guide development of new Open Context publication services that encourage the sharing, debate and reuse of data creation methodologies and data organization systems (controlled vocabularies and ontologies). Rather than imposing arbitrary technical standards that may constrain a researcher's intellectual freedom, these publishing services will help situate "data management" as an integral aspect of scholarship. In doing so, it will encourage wider intellectual investment in fundamental challenges of archaeological data preservation and reuse.

Many granting programs, including NEH (digital humanities) and NSF (archaeology), now require data management plans as part of applications. These policy changes reflect technological advances in data capture and storage, as well as increasing recognition of the strategic need to share, preserve and reuse research data. Unfortunately, many archaeologists lack awareness of the downstream research uses of digital data. Thus, they lack understanding of what "good" data management means in terms of their own research practices. A host of complex issues, including costs, technological capabilities, data documentation challenges, professional incentives, and legal considerations all hinder the ability of archaeologists to better manage, make use of, and share their data.

Regardless of the research question or theoretical approach, methodological rigor must underpin all archaeological endeavors (among many, see Hodder 2001; Flannery 2006). As digital data increasingly play a key role in all forms of archaeological observation and recording, professional practice must increasingly emphasize rigorous and effective data management. Several recent papers have called for more serious theoretical engagement in archaeological data (Kansa 2015; Shott 2014; Dallas 2015; Huggett 2015a, b). Unfortunately, without examples of how standards, metadata, and data quality impact research outcomes from sharing data, field archaeologists will have little motivation to improve their data creation and management practices. An overly Taylorist focus on "incentivizing" repository deposit with data citation metrics and making data management a condition of funding will do little to encourage more intellectual investment in data (Kansa 2014b, 2015). We need approaches that inspire and motivate greater intellectual engagement with data so that data becomes more than a bureaucratic compliance concern of only secondary or tertiary importance to core research goals.

This project will improve the practice of archaeological data management by developing cost-effective strategies to align data creation with reuse and understanding. This project builds upon existing best practice guidance by considering how data creation practices impact downstream reuse of data. In doing so, it takes a holistic approach to data, considering every aspect of its lifecycle, including planning, creation, use, dissemination, and preservation. Empirical study of both data creators and data consumers will inform understanding of needs across the entire data lifecycle. Only by considering how data flows in a research information ecosystem, before the tip of the trowel even touches the ground or a survey begins (Austin 2014), can we better meet the demands of data-intensive, 21st century research programs.

### 3.1.1 PROJECT OBJECTIVES AND OUTCOMES

This project aims to expand our understanding of best practices in data management in order to improve the quality and research impact of data now filling digital repositories. Through empirical qualitative study of researcher needs in reusing data and through the publication and study of multiple comparable datasets, this project will build upon prior work and aim to meet the following interconnected objectives:

1. Consider the entire data lifecycle when developing systematic approaches to align data creation and field data management practices with preservation, dissemination, and reuse requirements.
2. Identify at what point in the data lifecycle challenges encountered during reuse are introduced and what changes can be made to minimize them.
3. Identify how differences in tools, data characteristics, and disciplinary practices in a range of geographic locations and time periods affect data collection, management, and documentation.
4. Enhance the research value of archaeological data by better integrating them with other datasets and publications created by colleagues inside and outside of this specific discipline.

The products of this work over three years will include the following:

1. Extend Open Context's services (see below) to enable researchers to define, publish, network, and reuse controlled vocabularies and data models needed to organize information. Rather than imposing arbitrary standards, this approach will enable researchers to share common ways of organizing data in an iterative, contestable, and "bottom-up" manner.
2. Offer "context aware" reconciliation services so that researchers and other data managers can relate their own data to controlled vocabularies authored by selected peers and other authorities.
3. Create and disseminate high-quality, open archaeological datasets as exemplars of good practice.
4. Build tailored, web-based guidance (a "DMP-adviser") and "recipes" for scholars to create higher quality and more widely usable research data, thus widening participation in Linked Open Data.
5. Institutionalize professionalism in data creation, management and curation in Institute for Field Research (IFR) field programs across the world and in Open Context's data publication programs.

Investigating how data creation practices impact reuse represents a new approach to data management. By systematically analyzing data quality and modeling needs in a variety of settings, this project will identify methods and practices that apply to multiple temporal periods and geographic regions. This will make data management policies and investments more effective, and promote greater professional recognition and intellectual investment in the creation and use of data. Such incremental changes in professional attitudes and practice will make archaeology more rigorous, open, and inclusive.

### 3.1.2 TOWARD MEANINGFUL PRESERVATION AND ACCESS

The emerging discipline of "data curation" has a growing body of literature documenting scientific data practices (among others, see Yakel et al. 2013c; Faniel et al. 2012; Rolland & Lee 2013). Until recently, most of the data curation literature has focused on research data archiving needs and practices (Borgman 2007; Richards 1997; McManamon & Kintigh 2010). Many assume that structured data mainly needs to be "archived" with repositories. In other words, a researcher's main responsibility toward data centers on preservation. This emphasis on data preservation with repositories represents a normative best practice.

While data archiving has recently attracted funding and assumed greater policy importance, calls to archive data may not sufficiently motivate changes in professional practice needed to make shared data widely useful and used for new research. Archaeologists still invest little professional discussion or scrutiny in data modeling and documentation. Surveys and interviews conducted by the DIPIR project[1] reveal that archaeologists often ignore or see extant guidelines as unhelpful (Faniel et al. 2013). These attitudes limit the impact of "best practice" guidance for data modeling and creation, such as the guide co-published by the Archaeology Data Service (ADS) and Digital Antiquity.[2] Furthermore, while the NEH, NSF, and other funders require data management plans, they currently provide no specific guidance on the management of data, leaving both the development and review of data management plans in the hands of people who often lack guidance or expertise in what constitutes a good data management plan. To help fill this void, several university libraries and disciplinary repositories have come together to give the

[1] See description below in section **3.3 History, Scope, and Duration**
[2] See: http://guides.archaeologydataservice.ac.uk/

research community better guidance in grant-mandated data management, such as the DMPTool[3], an online system to aid the creation of project-specific data management plans. However, while useful in general terms, the DMPTool does not offer discipline-specific standards or modeling help.

Best practice guidance without reference to concrete and specific examples of research applications and outcomes may seem too abstract and irrelevant to the priorities of many practicing field researchers. Thus, archaeologists lack motivation to seek out and follow rigorous approaches that bridge data modeling, data creation, archiving, dissemination, reuse and integration needs. Currently, archaeologists usually move through each of the steps in the research data lifecycle in an *ad hoc* and piecemeal manner. The failure to align data management with research needs and outcomes undermines the point of data preservation. The data curation literature notes that the actual reuse of data remains rare in many fields (Wallis et al. 2013; Wallis 2014; Peer et al. 2014). Addressing the issue of data reuse has assumed greater urgency, given the substantial investments flowing into repositories (Faniel et al. 2013; Faniel & Jacobsen 2010a). As case studies in data reuse are still rare, applicants E. Kansa and S. Kansa, with their colleague B. Arbuckle recently won a "best paper" prize for their study of workflows, complexity, and costs in data reuse in archaeology (Kansa et al. 2014). This study highlights the importance of regarding data as more than a "residue" of research needing archiving. To be usable by a wider community, data require substantive intellectual investment in modeling and validation (see also Kratz & Strasser 2014).

## 3.2 BACKGROUND OF APPLICANT

Archaeology is inherently interdisciplinary, involving collaboration among many specialists worldwide. The Web has emerged as the key medium for the dissemination and reuse of research data across virtually every discipline (Kuhn et al. 2014; Groth et al. 2013) including archaeology (Wells et al. 2014; Isaksen et al. 2014; Niccolucci & Richards 2013; Kansa 2012, 2014a) and related fields of ancient studies such as art history, numismatics, geography, epigraphy (see Elliott et al. 2014). As such, archaeological data management needs to work toward good practices, especially Linked Open Data[4] methods and standards, in using the Web for data archiving, data publishing, and integration programs.

The AAI, the lead institution on this application, works at the interface of archaeology and the Web. Our research over the past decade has explored many aspects of the data lifecycle, including data acquisition workflows, data integration with Linked Open Data, accessibility, citation, and reusability (see titles of grant-funded projects in **Section 7- History of Grants**). Several of the co-investigators on this proposal have already demonstrated successful collaboration (see Faniel et al. 2013). Our past work also highlights that successful data sharing projects leverage well-established relationships among colleagues (Kansa et al. 2014); thus, we partner with leaders who can bring those types of connections to this project. Our interdisciplinary team brings expertise in data publishing, field excavation, information science, qualitative data acquisition and analysis, archiving, and database and interface design.

The AAI was incorporated as a non-profit in 2001 and has secured continual funding from government grants, private foundations, individuals, and consulting toward its mission of enhancing scholarship through use of the Open Web. Our research and development efforts aim to develop professionally-vetted, comprehensive, and open access scholarly resources, specifically through the open access data publishing platform, Open Context. An early adopted of Creative Commons licensing and data citation, Open Context is now referenced by the NSF and NEH as an option for data management. Open Context's data publications, which are all open access, include 50 projects worldwide, representing 350+ researchers. Significant datasets include UNESCO World Heritage sites (Petra, Catalhöyük, Giza) and the Digital Index of North American Archaeology (DINAA), the largest set of data documenting ancient settlement

[3] See: https://dmp.cdlib.org/

[4] The W3C recommends "Linked Data" (http://www.w3.org/standards/semanticweb/data) for Web-based data sharing. Linked Data uses stable Web addresses (URL/URIs) to identify concepts in datasets, allowing data publishers to share metadata based on common standards and cross-reference (integrate) data across the Web.

in North America, currently with 350,000 site records. The AAI partners with the California Digital Library (CDL) for data archiving, the Mozilla Science Lab for digital humanities training, the German Archaeological Institute for mirror hosting, and the Cotsen Institute of Archaeology Press at UCLA and a growing list of publishers for publishing datasets linked to conventional publications.

The AAI's work has a global reach, including scholars who collect primary data, those who wish to discover Web data, and publishers linking to web-published datasets. Since 2013, the AAI has been working with several organizations to develop data management, self-archiving policies, and open access policies. In 2013, Open Context's developer and Co-I on this application, Eric Kansa, was recognized by the White House as a Champion of Change in Open Science.

### 3.2.1 FACILITIES & EQUIPMENT

All project activities will be conducted on standard desktop, laptop, tablet, and networked computers available to the project's key personnel. Project participants will work from their home institutions and at the three field sites (A. Austin). The AAI operates Open Context from Google's commercial cloud-hosting services and German Archaeological Institute hosts a mirror of the site. As discussed below, Open Context facilitates data preservation by accessioning all data into a separate institutional repository managed by the University of California system's California Digital Library (CDL). The CDL preserves data (independent of Open Context's continued operation) and provides stable identifiers (DOIs, ARKs) for Open Context content. The project will provide 3 tablets for data collection at each of the 3 partner archaeological field sites. These will be used during Years 2 and 3 of data collection after initial observations in the field of the projects' existing data collection strategies.

## 3.3 HISTORY, SCOPE, AND DURATION

Our multi-institutional team includes the Institute for Field Research (IFR)[5], an internationally-recognized non-profit offering field research courses at archaeological sites worldwide; directors of archaeological excavations in Peru, North Africa, and Europe; and OCLC[6], addressing challenges facing libraries and archives in the rapidly changing 21st century information technology environment. Participants bring expertise and successes in various aspects of this project: data organization, data dissemination, development of tools, and application of standards (Kansa and Kansa); qualitative data acquisition and analysis (Faniel and Yakel); archaeological field research planning and documentation (Boytner and field projects); and development and user experience assessment of field-based documentation tools (Austin).

A key aspect of this project involves extending Open Context, which will provide long-term, open access to the project's outcomes (datasets, data models, and controlled vocabularies). Open Context now publishes and archives over 1.2 million records from projects worldwide, a scale comparable to that of a major museum (for instance, the online collection of the Metropolitan Museum of New York makes some 407,000 records available). Open Context publishes a wide variety of data, ranging from archaeological survey data to excavation documentation, artifact descriptions, chemical analyses, and detailed descriptions of bones and other biological remains found in archaeological contexts. Open Context adapts a "publishing" metaphor for data sharing to help set expectations about labor and intellectual investments needed for meaningful data dissemination (Kansa and Kansa 2013). "Data sharing as publication" helps convey the idea that data dissemination involves co-production, where data authors and specialized editors work collaboratively, contributing different elements of expertise and taking on various professional responsibilities. A publishing metaphor is widely understood by the research community, helping to convey the idea that data publishing implies efforts and outcomes similar to conventional publishing. We also hope that offering a more formalized approach to data sharing can also promote professional recognition, helping to create the reward structures that make data reuse less costly and more rewarding, both in terms of career benefits and opening new research opportunities in reusing shared data.

[5] https://www.ifrglobal.org/
[6] http://www.oclc.org/home.en.html

Part of our team recently demonstrated the value of "data sharing as publishing" in a collaborative study using integrated data sets to explore the dispersal of early domestic animals in the Neolithic (Arbuckle et al. 2014). Through large-scale integration and comparative analysis of data collected by 34 archaeologists working at 15 sites in Turkey, this study improved archaeological models of the initial dispersal of agropastoral economies in the Near East. Creating these data required great investments in training and fieldwork, years of expert observation, and laboratory facilities—a financial investment far greater than the $33,000 award enabling digital publication, preservation, and face-to-face collaborative analysis. In making these data freely accessible to future researchers, this study highlighted how data reuse represents a huge economic efficiency gain in the practice of archaeology (see also Beagrie & Houghton 2013).

Project Co-I Faniel led the project "A Cyberinfrastructure Evaluation of the Network for Earthquake Engineering Simulation (NEES)", with the objective to evaluate NEES, a large cyberinfrastructure initiative, part of which was sharing, managing, and reusing data via a digital data repository. Faniel's evaluation of data sharing and reuse helps set the stage for the current project. Research outcomes highlighted the need for more research on large scale data sharing and reuse (Faniel & Zimmerman 2011). Increasing the supply and access to data via repositories is not sufficient for data reuse; there is a need to provide information about the context of data's production in order for researchers to decide whether it is reusable as well (Faniel & Jacobsen 2010b). The NEES project produced several outputs which have impacted research and practice related to the success and challenges of cyberinfrastructure initiatives. The report has been downloaded 550 times since it was deposited in Deep Blue at the University of Michigan. The work related to data sharing and reuse was used to define a research agenda for large scale data sharing and reuse (Faniel & Zimmerman 2011), which has been cited 37 times, and resulted in one of the early studies of data reuse practices (Faniel & Jacobsen 2010b), cited 52 times.

As one of the early studies of data reuse, the framework inspired the research program Faniel launched with funding from the Institute for Museum and Library Services for the Dissemination Information Packages for Information Reuse (DIPIR) project, which examined data reuse practices in social science, archaeology, and zoology to identify how contextual information about data that supports reuse can best be curated and preserved. Thus, our current proposal's holistic consideration of data within a broader research lifecycle has a firm foundation of prior research. This project builds upon the DIPIR project's qualitative research methodology examining archaeologists' data reuse practices, how they construct trust in repositories, and how repositories manage changes to data over time (Kriesberg et al., 2013; Yakel et al. 2013c; Daniels et al. 2012; Faniel et al. 2013). Our findings highlight archaeologists' concerns over methods and sampling procedures as well as how archaeologists use records generated by their colleagues and by third parties, such as museums and repositories (Faniel et al. 2013; Faniel 2014).

This previous work has shown actions that take place in one part of a lifecycle can create challenges or facilitate work in another part of the lifecycle (Faniel & Yakel 2014; Kansa et al. 2014). Although repository processing benefits archaeologists who share and reuse data, it cannot reverse archaeologists' collection and documentation practices that take place in the field. Actions that take place in the field when archaeologists initially model and document data play key roles in shaping later data reuse. After-the-fact data curation (clean-up, metadata documentation) cannot undo sampling decisions, data modeling approaches, or the application of underspecified standards (Frank et al. 2015; Yakel et al. 2013a; Faniel & Yakel 2014). Thus, improving data creation will set better conditions for reuse. Similarly, the condition of the data (e.g. software format, whether it is coded or decoded, etc.) also has implications for curation and reuse (Faniel & Yakel 2014; Kansa et al. 2014). Such seemingly trivial details greatly impact the success of later data preservation efforts and the time and effort needed for reuse.

## 3.4 METHODOLOGY AND STANDARDS

Current "best practice" guides (i.e. ADS/Digital Antiquity) have already defined invaluable technical standards for data preservation and documentation of different file types and content types. However, semantic standards needed to model archaeological data for discovery and interoperability remain far

more contentious. The focus on this project centers on situating such semantic standards within larger intellectual agendas and needs. Open Context's model of data publication underscores how digital data involves intellectual effort and creativity. While Open Context facilitates repository data preservation, it offers a very different approach to information organization than most repositories. Conventional repositories strive to preserve digital files and make them discoverable with some metadata documentation. The main object of search and citation in digital repositories centers on digital files (spreadsheets, image files, relational databases, etc.). In contrast, Open Context makes researcher-defined "entities" (records of sites, potsherds, bones, coins, and classification terms and properties) the main objects of search and citation. Each such entity represents a "micro-publication" that can be individually indexed, retrieved (in HTML and machine-readable JSON-LD and CSV open formats), and cited (see section **5.4 Data Management Plan** for more specifics on the technical and semantic standards).

Though more labor intensive than simple repository deposit, Open Context's approach enables more granular access and citation of specific items of interest defined by the contributing researcher. This makes the complexity and scale of archaeological data more manageable. For example, an ancient coin may be represented in several tables and files in a project's documentation (such as a finds catalog, a context inventory, a photo log, several digital images, and a table of XRF results). Thus, information about the coin may be scattered across thousands of records and in many tables and files, all organized with different schema. This highlights the limitations of repository archiving, which makes the "file" the main object of discovery and citation. Files can be arbitrary and opaque containers for items of interest (such as coins). Thus, if we limit data curation practices to archiving digital files, citation (or even discovery) of specific archaeologically meaningful entities (like a coin) becomes nearly impossible.

Referencing specific archaeological items plays an important role in interoperability. Linked Open Data, the current best-practice for data sharing and interoperability, centers on relating data across the Web by referencing stable Web URIs (a URI is a URL that is also a globally unique identifier). Since Open Context mints stable URIs for each item of archaeological interest (as defined by the data author), it is possible to use Linked Open Data with much more granular and specific information than feasible with a typical repository, where one can only link to an aggregate of information encoded in a file. This makes it possible for anyone to reference precisely specified sites, coins, potsherds, or even individual categories in a researcher's typology. One can link those items with items published anywhere else on the Web. This last point highlights a key advantage of greater granularity in Linked Open Data applications. Open Context's granularity helps it to network data and cross-reference with other cultural heritage information systems. For instance, Open Context already cross references data with tDAR[7], the Catalhöyük Living Archive[8], Arachne[9], Pleaides[10], and other systems despite differences in data models and software.

This discussion shows how interoperability therefore is not an all-or-nothing issue. Linked Open Data can help cross-reference relevant parts of different datasets in a relatively simple "loosely coupled" manner. As demonstrated by our prior successes in zooarchaeology (Arbuckle et al. 2014) and by the Pelagios's Project's[11] success in linking several hundred thousand historical geography data points (Isaksen et al. 2014), research reuse of data does not require total semantic harmonization. Instead, the simple formalisms of explicit Web identification and cross-referencing promote interoperability at a grand scale.

While Linked Data can help network together diverse data, many areas of archaeology have yet to develop the vocabularies needed as common points of reference. Our project will explore ways to encourage the research community to author and publish such vocabularies while still promoting autonomy in defining and describing materials. To do so, we will use interviews and participatory

[7] http://tdar.org
[8] http://catalhoyuk.stanford.edu/
[9] http://arachne.dainst.org/
[10] http://pleiades.stoa.org/
[11] http://pelagios-project.blogspot.co.uk/

observation studies to gather qualitative evidence on how and why researchers create data, and how their data creation practices articulate with reuse. In studying these issues over 3 field seasons in 3 different research contexts, we will be able to guide better services for research data management.

### 3.4.1 TENSIONS WITH STANDARDS

The term "standards" refers to many different issues, including research methods, recording practices, technical formats, data models, and controlled vocabularies and terms. All of these issues impact meaningful interoperability. For example, one widely recognized standard is the CIDOC-CRM[12], a widely-used (especially in Europe) domain ontology for cultural heritage data interoperability. However, even using the CIDOC-CRM for something as seemingly simple as representing Munsell colors, shows the complexity of how standards interact with research practices (Kansa 2014a). A Munsell color reading can be considered as a measurement, making the CIDOC-CRM property "P43F has dimension" appropriate for representing Munsell color values. However, in practice, many researchers take Munsell readings because they vaguely think they should, and then do not adequately control for several factors (lighting, dampness). Since the practice of taking Munsell values often lacks formal controls, using the CIDOC-CRM property "P3 has note," a concept meant for informal description, may be a better choice. Thus, recording practices and methods impact the appropriate use of information standards. Imposing common data standards without consideration of behind-the-scenes data collection methodologies makes interoperability superficial and interpretively suspect. Neither Open Context's use of Linked Open Data methods (see also Kansa et al. 2014), nor tDAR's user-generated ontology mappings (see Spielmann & Kintigh 2011), can yield analytically meaningful results without consideration of data creation practices.

Similarly, chronological periods, though central to archaeological practice, illustrate how the theoretical challenges of attempts to standardize metadata (Rabinowitz 2014). Researchers sometimes reference periods to advance certain interpretive arguments. For example, the past two decades have witnessed ongoing controversies over "High", "Middle", and "Low" chronologies for the Eastern Mediterranean (Manning et al 2001; Coldstream and Mazar 2003; Finkelstein and Piasetzky 2003; Sharon et al 2007; Plicht et al 2009; Fatalkin et al 2011). These different chronologies reflect different understandings of a variety of historical and social changes including the end of the Mycenaean palatial system, emergence of the "Sea Peoples", and political developments in Biblical Israel (recently reviewed by Joffe 2007 and Boaretto 2015). Simply defining a standard chronology, even if scoped to a given geographic region, therefore could obscure important interpretative issues and debates.

To further complicate matters, archaeological recording practices, methods, and research designs evolve and must be tailored to specific research questions and field conditions (Schloen 2001; Kansa 2005, 2009; Kintigh 2006). For example, Open Context is now publishing the Pyla-Koutsopetria Archaeological Project (PKAP) a dataset documenting an archaeological survey near Larnaka, Cyprus, led by William Caraher. In this project, Caraher and his colleagues defined the "chronotype" system for classifying very fragmentary surface finds gathered in the survey (Caraher et al. 2006; Tartaron et al. 2006). Because body sherds make up the majority of the finds collected in survey, Caraher's team needed an alternative to ceramic typologies based on vessel forms and decorations. The chronotype system helps organize survey pottery to explore questions about diachronic patterns in settlement in the survey area. Again, this organizational scheme reflects the close relationship between research methods and classification.

### 3.4.2 FORMALIZATION RATHER THAN STANDARDIZATION

The above examples illustrate how archaeology can be described as an artisanal craft (Shanks and McGuire 1996), and why many archaeologists would reject attempts to "mass-produce" standardized and highly fungible data. The key need for the discipline is *not* to standardize what archaeologists say or cannot say about the past. Rather, we should aim for data management practices that make modeling and classification, including definition of new classification schemes, more formal and explicit. If

[12] See: http://www.cidoc-crm.org/

archaeologists want to meaningfully reuse and compare datasets from multiple field projects, and if they do not want to accept standardized recording practices, then they must accept greater responsibility in formally and precisely documenting and modeling their own "customized" approaches to organizing data.

The PeriodO project[13] illustrates the value of formalization rather than standardization (Shaw et al., in press). PeriodO models the geographic and temporal scope of a period, including information about the authority that defined the period. Because each PeriodO period has a computationally explicit definition, datasets annotated with these periods can be aggregated and compared. Furthermore, because PeriodO documents the authority that defined a given period, it provides some clues about interpretive perspectives. This enables, for example, use of a High, Middle, or a Low chronology version of the period "Iron Age I." Since preference for a High or Low chronology marks one's position in a theoretical camp, PeriodO helps to document an important element of scholarly context.

PeriodO illustrates the value of publishing research-defined classification systems using computational formalism. It does not demand agreement where agreement does not exist. We propose to extend this overall approach to other areas of archaeological data. This project will develop services and venues to help researchers to define and publish their classification schemes formally and explicitly using W3C (official Web) standards like Web Ontology Language (OWL[14]) and Simple Knowledge Organization System (SKOS). This strategy retains interpretive freedom while promoting interoperability. More formal approaches to modeling and classification will make it easier to reference, reuse, extend and adapt data and classification systems in a transparent manner. Given the research significance of explicit data modeling, this project will extend Open Context to better document data creation metadata and to better serve as a publication venue for SKOS and OWL vocabularies.

### 3.4.3 STUDY STRUCTURE

Formal modeling will facilitate the networking of data in a manner that promotes interoperability while still accommodating differences and continued evolution of archaeological practice. While archaeologists urgently need ways to formally and explicitly publish, reference and adapt controlled vocabularies and data models, the vast majority of field practitioners lack the needed technical expertise. Open Context can provide assistance to appropriately model and publish researcher-defined vocabularies and data models (with SKOS and OWL). However, such publication services will fail to meet researcher needs if there is not a clear understanding of how data creation practices relate to later data reuse. For example, Faniel et al (2013) noted that researchers interested in reusing data from a repository expressed a great deal of concern over sampling bias issues. To reuse a dataset with confidence, researchers needed adequate documentation about data collection and sampling methodologies. These issues are rarely discussed in archaeological data modeling or metadata proposals. Thus, development of new publishing services for Open Context should be guided by a firm understanding of how researchers actually use data in practice.

We will document data management practices through interviews and observations with archaeologists working on 3 excavations, identify data collection and management tools archaeologists use in the field, and interview archaeologists who represent potential reusers of the data from the 3 excavations. The specific practices deployed follow along with details about the research methodology we will use to inform the practices. Furthermore, we plan to take advantage of several different, but complementary, methods (structured interviews, field observations, and review/audits of archaeological datasets) in order to triangulate the observational data for stronger results (Creswell 2009; Denzin 1997; Yin 2003).

#### *3.4.3.1 Develop baseline data management best practices*

Archaeologists choose a host of software and data management tools. In many cases, they rely upon widely used proprietary and commercial office suites (database and spreadsheet applications), GIS, and

[13] http://perio.do
[14] "OWL" is not a typo, but the acronym convention for this particular standard.

image management tools. While not optimized for field archaeology, their ubiquity and polished user interfaces offer lower barriers to use and make them popular. Since this condition is not likely to change much, we propose to develop and promote simple and practical approaches to make more effective and rigorous use of these "off-the-shelf" tools. These approaches include promoting techniques for validating data, using controlled vocabularies (see above), and reliably using identifiers (organizational keys needed to relate different data, images, and other content together).

**A.** Conduct 5 interviews in person or by phone with archaeologists at each of the 3 excavation sites for a total of 15 interviews. Findings will establish a baseline that describes current data collection/ management needs and challenges during excavations, including the features and functions of the tools they use versus those they wish were available for use.

**B.** Conduct 25 interviews in person or by phone with archaeologists who have reused data and would be potential reusers of data collected at the 3 excavations (i.e. have interests in the time period, region, and specializations at each project). These 25 interviewees will be identified by the excavation directors and solicited via their professional networks. The interviews will focus on the problems archaeologists experience during reuse, particularly when attempting to integrate data from different sources. The findings from these interviews will be used to develop interview and observation protocols for Phases 2-4 (see **3.6 Work Plan**).

**C.** The interview process has two goals: 1) establish baseline data quality, modeling, and documentation requirements that will help inform better practice at the three field sites; and 2) recruit additional researchers to participate in data sharing and collaborative analysis. The interviews will help by identifying wider networks of collaborators who may have additional related data available for dissemination and comparison. Collaborative data analysis and integration with wider datasets will help multiply archaeological research impacts.

#### *3.4.3.2 Coordinate field data creation*

Archaeology is an inherently collaborative practice, with many participating researchers working in the field or in the lab. Creating a well-organized and reliable system for capturing and recording data contributed by many users, often working in challenging environments (heat, moisture, sun-glare, dust, etc.) represents a key requirement. The FAIMS project[15], now in its 4th year, offers a comprehensive open source field data management system to meet these needs. The FAIMS system will be evaluated and compared with more *ad hoc* approaches that emphasize commercial off-the-shelf data capture and management tools more widely deployed by archaeologists.

**A.** Create a feature/function list for the data collection/management tools used at the 3 excavation sites based on the previous interviews and continue to evaluate their performance at the 3 sites with interviews and observations during the 3 field seasons. In seasons 2 and 3, implement suggested changes in data collection using off-the-shelf software, including tablets for digital data collection *in situ*. Compare tools against the FAIMS system by evaluating the time required for data management, limitations to the types of data/data structures collected in the field, and the quality of data recorded in situ. This evaluation will guide import, modification, and re-use of data models and vocabularies defined by other researchers and published by Open Context in the FAIMS system.

#### *3.4.3.3 Explicit modeling of controlled vocabularies*

The development of archaeological typologies and classification systems is a fundamental aspect of archaeological data recording. However, such typologies are rarely published in a systematic way. It is even rarer for archaeologists to publish classification systems in a manner suitable for computation using interoperable open standards such as OWL or SKOS. Documenting and expressing archaeological typologies using such standards helps to describe the meaning of archaeological datasets. Making such

[15] See: https://www.fedarch.org

vocabularies open for reuse and wider scrutiny can refine typologies and improve overall data quality. Refined through years of development, open source tools, such as Protégé[16], allow researchers to easily author vocabularies. In each year of the project, controlled vocabularies used at each field site will be modeled by E. Kansa and A. Austin. With participating researchers, Austin will identify problems in the use of controlled vocabularies, and track the history of changes and refinements with GitHub.

#### *3.4.3.4 Publication of controlled vocabularies*

The implementation and use of these tools in archaeology needs to be tested. Because we anticipate some reluctance among archaeologists to use a specialized tool like Protégé to define their typologies, this project will extend Open Context to help researchers (with editorial assistance) publish typologies, implicitly or explicitly used in their databases, as SKOS vocabularies. Extracting implicit vocabularies will require modification (supported by this project) of Open Context's import and publishing processes. Once digitally published, vocabularies can be referenced, reused, and extended by future researchers to improve the creation and semantic compatibility of future datasets. To facilitate such reuse, we will make controlled vocabularies downloadable in several formats for more specialized Linked Data applications (JSON-LD, Turtle), and for more general users in simple tabular formats (CSV). Open Context will also put each controlled vocabulary into GitHub for version control, where they can be refined over time through GitHub's issue tracking and version control features. They can even be "forked" if researchers need to adapt a vocabulary for specific projects. Finally, context-aware entity reconciliation services (see below) will help even researchers lacking programming expertise to selectively use controlled vocabularies created by their peers and by other authorities. Thus this effort can facilitate greater interoperability without limiting a researcher's agency in crafting data creation strategies.

This project will build a data management plan advisory tool (DMP-adviser), aimed at disseminating data models, vocabularies, and good data creation workflows. The DMP-adviser, which Open Context will offer in addition to its current data management guidance[17], will inform researchers planning new field work about relevant controlled vocabularies and data modeling approaches, and offer links to specialized data creation tools (ArkDB[18], OpenDig[19], FAIMS), all tailored in response to a user's input. Our qualitative user needs studies will inform us about what sorts of advice to emphasize. This is a way of distilling best practices and sharing them in a way that's more immediately useful and relevant to a given researcher (as opposed to having them read through generalized best practice guides). As archaeologists publish more controlled vocabularies using open standards, the DMP-adviser can inform users of relevant vocabularies in more areas of topical specialization. Thus, vocabulary and data sharing can feed back to inform future data creation, and shared vocabularies and data will see greater scholarly impact. Using this tool will in no way require use of Open Context as a repository. Users may take what they learn and archive data with other repositories, including tDAR.

#### *3.4.3.5 Data preservation and dissemination*

Excellence in field documentation should facilitate analysis, interpretation, and the publication of interpretive syntheses though peer-reviewed venues, thus showing how data management aligns with professional incentives. Thus, a key goal of this project is to encourage the scholarly reuse of well-managed and well-documented data publications.

**A.** Each summer, Austin will observe and interview 15 archaeologists at the 3 excavations (5 at each site). She will spend two weeks at each site observing and identifying problems managing and documenting data. Follow-up interviews will document post-excavation processing and use of data. Findings from season 1 will help develop and introduce new or refined approaches, good practices

[16] http://protege.stanford.edu/
[17] http://opencontext.org/about/estimate
[18] http://ark.lparchaeology.com
[19] http://opendig.org/

for tool utilization, and tools for season 2. Findings from field season 2 will be used to develop and introduce new or refined approaches and tools for field season 3. Findings across the three field seasons will be used to guide enhanced data publishing services and data management guidance.

### 3.4.4 FIELD SITES

We chose 3 field sites that represent a wide regional and chronological range of archaeological data creation practices. We also chose them based on the availability of comparative data. In addition to publishing the corpus of data from the 3 participating sites (see below), Open Context will also publish data related to each of these sites in order to explore issues of data integration and comparative reuse. The related datasets are: the tophet at Carthage, Tunisia (see J. Greene's letter of commitment), related to the Zita project; an osteological dataset from the Huaura Valley, Peru (see L. Jahnke's letter of commitment), related to the Vitor project; and several Iron Age and Roman datasets already available in the Archeology Data Service (ADS), related to the Poulton project. Availability of these relevant comparative datasets will improve our understanding of how data creation practices impact data integration studies.

These field sites allow us to focus on data creation workflows, whereas in some other contexts, contested perspectives on intellectual property, religious views, and problematic relationships between some stakeholder communities and archaeologists raise difficult questions about ethical data management (Nicholas & Bannister 2004; Kansa et al. 2005; Kansa 2012; Nicholas et al. 2010; Christen 2012). These are important topics, and NEH and IMLS have already invested in projects to address information privacy issues (such as Mukurtu.org). Empowering communities with respect to digital cultural heritage involves a host of issues beyond access controls and intellectual property claims. Methods to define and promote alternative metadata and ontological systems for organizing information can empower and help bridge indigenous communities and academic communities. In that sense, this project will complement Mukurtu and other projects that seek to better align digitized cultural heritage with community needs.

In addition, research into these field sites complements DINAA, an ongoing project hosted by Open Context that focuses on public archaeology and cultural resource management (CRM) in North America. The NSF funded DINAA project is developing a gazetteer of North American site file identifiers from data contributed by state historical preservation offices, state site file administrators, and tribal historical preservation offices. DINAA involves broad stakeholder engagement, workshops, and collaboration with public archaeology and CRM professionals (Wells et al. 2014). Outcomes and experience from DINAA help will inform work on this project.

#### *3.4.4.1 Field Site 1: Zita, Tunisia*

The Zita Project is the first United States-Tunisia archaeology and ethnography partnership since the Arab Spring. Jointly run by scholars from Tunisia and the US, intensive survey, mapping, and excavation have been carried out over the past two summers. Excavations at Zita have succeeded in identifying the remains of a Roman Forum (imperial administrative structure), a Neo-Punic tophet (Carthaginian child sacrifice precinct), and a metallurgical precinct. Data collection and documentation methods include stratigraphic excavation based on a resolution of locus, pottery bucket, and artifact, with all data entered into a FileMaker database that has been constructed specifically for the project. The data are entered offline after daily excavation activity, then uploaded and shared with other project members via Dropbox.

#### *3.4.4.2 Field Site 2: Vitor, Peru*

Since 2009, the Vitor Archaeological Project has been conducting multidisciplinary research in the Vitor Valley of Southern Peru. Located 40 km southeast of the modern city of Arequipa, Vitor has served as a nexus between the coastal and highland societies. Research focuses on the issue of material cultural meaning, what is authentically local and what are imported design and cultural traits. We extend this question to the bioarchaeological realm, contextualizing human remains by studying biological diversity and origin in relations to specific material cultural traits. Systematic recovery of scattered material permits crucial osteological and cultural studies that may help to determine the group's biological,

cultural, political and economic affinity with much larger cultural groups outside Vitor. Data from excavations have been recorded in specially-designed forms, written by hand *in situ*. Survey data have been collected digitally, using Total Station, Differential GPS and recorded using ArcGIS.

#### *3.4.4.3 Field Site 3: Poulton, United Kingdom*

Excavations at Poulton, UK are investigating a multi-period landscape that has seen occupation from the early Bronze Age to the medieval period. Research focuses primarily on the late Iron Age to Roman transition, particularly on the social and economic changes that were entailed in a shift from high-status Iron Age settlement to Roman industrial use, associated with the potential construction of a villa. The primary research aim is to identify and characterize archaeological remains associated with high-status Iron Age and Romano-British occupation, and recover artifact and environmental samples to qualify the economy and use of structures on the site. Whilst retaining a paper records system, Poulton is also developing digital recording systems that can be synchronized with the national excavation database held in the UK by Online Access to the Index of archaeological investigations (OASIS), based at the University of York. These will be trialed alongside paper records to test the robustness of the system.

#### *3.4.4.4 Combining Depth with Breadth*

Archaeological explorations span human history and prehistory, using a broad range of research methods in diverse environments. Given budget and time constraints, we can only sample archaeology's immense disciplinary breadth. In doing so, this project will take a "T-shaped" approach. It will examine field data collection practices broadly but focus in-depth on the collection of specific classes of data centering on human and animal remains. This focus has a number of advantages:

- *More readily comparable data:* Well-established common documentation frameworks for human remains (Buikstra & Ubelaker 1994; Steckel & Rose 2002) can facilitate comparison of practices across each field site. Similarly, zooarchaeologists record taxonomic and skeletal elements and use several common recording methods (Driesch 1976; Payne 1973). However, the implementation of such standards can be highly variable. As identified by Kansa et al. (2014), despite widespread use of a common system for scoring tooth eruption and wear (using Payne 1973), resulting data could not be aggregated because zooarchaeologists recorded complex observations of multiple teeth as free-form text in comments fields. Common recording standards do not necessarily lead to comparable data, especially in cases that require sophistication in data modeling. Thus, study of how data recording practices relate to reuse even in these areas where we expect more comparable data will help inform problems in data management more broadly.
- *Multimedia documentation:* Like many areas of archaeology, human and animal bone documentation involves the creation and management of digital data in multiple media. Structured databases (usually tabular or relational data), GIS, drawings, photographs, and sometimes photogrammetry and 3D models all play important roles (Levy et al. 2010). These various media need to be managed, cross-referenced, archived, and disseminated. File formats, file sizes, and associated metadata requirements all factor in downstream preservation and reuse.
- *Global Significance:* Bioarchaeology and zooarchaeology represent key sources of evidence about status hierarchies, gender and other identity dynamics, interpersonal violence, ideologies, demographics, health and nutrition, diet and economy.

## 3.5 STAFF

Our interdisciplinary project team includes specialists in archaeology, anthropology, archiving, and informatics. Our core team has collaborated on several publications and conference presentations (Yakel et al. 2013b, 2013c; Faniel et al. 2013), and a major data integration and reuse study (Kansa et al. 2014; Arbuckle et al. 2014). The qualifications and time commitment of the key project participants is below. Table 1 lists all team members (for reference in the next section, **Work Plan**).

***Table 1:*** *Project Team Members*

| **Key Personnel** | |
|---|---|
| **Sarah Whitcher Kansa (SK)**, Alexandria Archive Institute / Open Context<br>**Ixchel Faniel (IF)**, OCLC Research | **Eric C. Kansa (EK)**, Alexandria Archive Institute / Open Context<br>**Anne Austin (AA)**, Stanford University |
| **Other Personnel** | |
| **Ran Boytner (RB)**, Institute for Field Research<br>**Elizabeth Yakel (EY)**, University of Michigan, School of Information<br>**Zita Excavation Directors**: Brett Kaufman (Brown), Hans Barnard & Rayed Khedher (UCLA), Ali Drine (Institut National du Patrimoine, Tunisia) | **Vitor Excavation Directors**: Maria Cecilia Lozada (University of Chicago), Hans Barnard (UCLA), Augusto Cardona Rosas (Centro de Investigaciones Arqueológicas Arequipa)<br>**Poulton Excavation Directors**: Ben Edwards, Seren Griffiths (Manchester Metropolitan U., UK) |

**Project Director Sarah Whitcher Kansa** is the Executive Editor for Open Context, where she manages the full cycle of data publication, from solicitation and management of submissions to archiving with the California Digital Library. She is also a practicing zooarchaeologist with experience in the U.S., Europe, and the Middle East. Her domain expertise in zooarchaeology will complement Postdoctoral Researcher Austin's bioarchaeology expertise. S. Kansa will dedicate .5 FTE to project management, data editing/ publishing, dissemination, and development of the DMP-adviser. Half of her cost is provided by the Alexandria Archive Institute through other funding sources. No salary escalation is requested.

**Technology Director (and Co-I) Eric Kansa** is Program Director for Open Context. His role in this project combines intellectual leadership and technology implementation of required standards and features. He will develop, test, and expand Open Context's suite of Web services and Linked Data services that form the basis of interoperability and data portability for the project. He will also oversee development of the DMP-adviser. In addition, E. Kansa will program new features for Open Context to enable publication of vocabularies created by researchers using SKOS and OWL standards. He will respond to feature and interface improvement requests to Open Context, where feasible. Finally, he will insure that all data published by Open Context are accessioned by the CDL for long-term preservation, access, and curation. He will dedicate .3 FTE to this project. No salary escalation is requested.

**Co-I Ixchel Faniel** is an Research Scientist for OCLC Online Computer Library Center, Inc. OCLC is a worldwide library cooperative working to improve access to the information held in libraries around the globe, and find ways to reduce costs for libraries through collaboration. Faniel brings expertise in qualitative and quantitative research methodologies and user behavior research. She has studied data sharing, management, and reuse and the role of digital data repositories within academic communities, including archaeology. Faniel will dedicate .10 FTE to this project. Her salary is provided through cost-sharing by OCLC. She is requesting travel funds for data collection and analysis activities (one week in each of Years 1 and 2 to collaborate with Postdoctoral Researcher Anne Austin) and the dissemination of project results annually at conferences. She will develop data collection instruments, conduct interviews with study participants, analyze data, and disseminate results through publications and presentations. As a Co-I on the project Dr. Faniel will work closely with Postdoctoral Researcher Austin to train her to be the primary team member responsible for data collection, management, and analysis.

**Senior Person Ran Boytner** is Director of the Institute for Field Research (IFR). He is the liaison between this project and the three archaeological excavations participating in this study, as the field projects are managed by IFR. Boytner will provide oversight for the field-based project activities, namely, Austin's field-based data collection during project phases 2, 4, and 5. IFR will charge no additional costs (other than room and board) to Austin for her work at the excavation sites. Boytner will build data management recommendations into IFR excavation review processes. He is requesting travel funds for face-to-face meetings in San Francisco and travel to one conference/year to disseminate project results.

**Postdoctoral Researcher Anne Austin** (Stanford University) will commit 5 months/year over the course of the 3-year project to carrying out the interviews and field research. Austin has a Ph.D. in Archaeology (UCLA, 2014). She has expertise in mortuary archaeology and bioarchaeology, as well as database design and the development and application of standards to improve data documentation practices. She created OsteoSurvey, an open-source series of XML files for collecting bioarchaeological data on Android-based mobile devices. These skills, in addition to her extensive field experience, make her well qualified to manage the field-based data documentation activities for this project. Austin will work with all team members to develop an interview protocol, and with Faniel and Yakel to conduct and analyze interviews. She will travel to all 3 field sites annually during summer months, spending at least two weeks at each site observing data documentation strategies and interviewing project participants. She will work with E. Kansa and S. Kansa to identify and extend effective field-based data documentation methods. Austin will undertake the proposed work during Years 1 and 2 as part of her current postdoctoral research. In the third year of the project, she is requesting compensation for five months of work.

**Consultant Elizabeth Yakel** is Professor of Information and Associate Dean for Research and Faculty Affairs at the University of Michigan, School of Information. She is requesting travel funds to support research design and data collection and analysis efforts (with Faniel and Austin), as well as dissemination of project results. She is not requesting any salary from NEH.

One **Research Assistant** (RA) will be hired to work 5 hours/week for 30 weeks/year and 20 hours/week in the summer months for the duration of the project. The RA will have archaeological and/or database management experience. The RA's responsibilities will include: processing transcripts (in / out with Scribie, etc.); checking transcripts after transcription; possibly participating in interviewing (data collection) and coding (data analysis); assisting with data management; collecting background information about sites/interviewees; and assisting in the functional analysis of the technologies. While working on the project, the RA will maintain at least weekly contact with Austin and monthly contact with S. Kansa. The RA will undergo responsible conduct of research training. The RA salary is $15/hour.

## 3.6 WORK PLAN

Activities will take place in five phases. All expenses and work will occur over a three-year period.

**Phase 1 (Months 1-10):** Develop Phase 1 interview protocol (whole team); Conduct baseline interviews in person or by phone with archaeologists participating in the three field projects (AA, IF); Review initial transcripts and effectiveness of protocol (AA, IF, EY); Create a features/function list for data collection/management tools used in the field for each project (AA, EK, SK); Introduce data collection/management tools for projects that do not have existing tools (AA, EK, SK); Conduct interviews in person or by phone with archaeologists reusing data (with interests in the time period, region, specializations at each project) (AA, IF); Analyze interviews and share findings with core team (AA, IF, EY); Draft field interview and observation protocols for Phases 2-4 (all team members); Meet with core team to finalize interview and observation protocols for Phases 2-4 (whole team); Start software development for publication of controlled vocabularies and reconciliation services (EK).

**Phase 2 (Months 11-14):** Interview and observe archaeologists at the 3 sites during the first field season (AA, with oversight by RB in the field and IF and EY virtually); Initial SKOS modeling of controlled-vocabularies / typologies (EK, AA); Start development of DMP-adviser tool (SK, EK).

**Phase 3 (Months 15-23):** Analyze field interview and observation data (AA, IF, EY); Review findings during meeting with core team; Introduce new or refined documentation approaches/tools based on interviews (EK, SK); Present mid-project findings at conferences (whole team); Revision of SKOS modeling of controlled-vocabularies / typologies (EK, AA); Conclude software development for Open Context publication of controlled vocabularies (EK); Refine DMP-adviser tool (SK, EK).

**Phase 4 (Months 24-27):** Interview and observe archaeologists at the three sites during the second field season (AA, with oversight by RB in the field and IF and EY virtually); Conduct mid-project interviews with excavation teams to gauge success of data collection approaches and tools (AA, IF); Final revision of SKOS modeling of controlled-vocabularies / typologies (EK, AA); Start publishing datasets and

controlled vocabularies from field-sites and other contributing researchers using Open Context (SK, EK); Refine DMP-adviser tool and reconciliation services and "recipes" (SK, EK).

**Phase 5 (Months 28-36):** Analyze interview and observation data (AA, IF, EY); Develop guidelines for best practices in data management (whole team); Create white paper (whole team); Interview and observe archaeologists at the three sites during third field season (AA, with oversight by RB in the field and IF virtually); Disseminate findings at conferences and in publications (whole team); Conclude data and controlled vocabulary publications with Open Context (SK, EK); Finalize development of DMP-adviser tool, reconciliation services and "recipes" (SK, EK).

### 3.6.1 Software Development Methods

This project will involve continued open-source software development with Open Context, including: the Django (Python) framework, a Postgres datastore, Apache Solr for indexing, and Bootstrap, Leaflet, and jQuery for client interfaces. Open Context offers a powerful and flexible API and publishes researcher defined "predicates" (descriptive properties, linking relationships) and "types" (concepts in controlled vocabularies) and can model these concepts with other concepts published elsewhere on the Web using SKOS. However, Open Context needs additional software development to make researcher defined data models and concepts easier to document, support multi-lingual labels and annotations, discover and use via the API, and cite as coherent scholarly works. This project will support software development to enable open-access publication of data models and controlled vocabularies.

These software development efforts will continue to use GitHub for version control, issue tracking, documentation and distribution. GitHub will play a key role in overall software project management, including planning, feature requests, debugging, and deployment instructions. Because software development will extend a functioning system with additional features and because this project involves wide qualitative user-needs research, we are well-positioned to adapt agile and user-centered design methodologies. Essentially this will involve iterative deployment and enhancement of features and interfaces in response to user feedback (mediated by interviews, email, GitHub).

## 3.7 Sustainability and Evaluation

### 3.7.1 Open Data Publishing to Make Archaeology more Sustainable

Publishing data reduces wasted effort associated with neglecting data. It helps ensure that money and effort invested in archaeological field work yield greater returns, ultimately helping to make archaeology more financially sustainable (Kansa 2012). For example, the 5-year Kenan Tepe excavations and analysis required roughly $800,000 in direct costs. Publication of this large, complex dataset in Open Context cost just $15,000. If we can finance costly excavations, we must finance better stewardship and dissemination of those excavation results. Moreover, because archaeological research methods are often destructive, every dataset represents unique work that can never be replaced, making data publishing a low-cost and worthwhile investment for preserving unique elements of our cultural heritage.

### 3.7.2 Interoperability and Archaeology's Information Ecosystem

Open Context is not the only system publishing archaeological data with such granularity. ArkDB, Heurist[20], FAIMS, the Çatalhöyük Living Archive, Arches[21], and other specialized systems, especially various systems hosted by the American Numismatic Society[22], similarly offer high granularity access to data. These systems all have different organizational schemas and approaches to data modeling. Such information diversity should be seen as a feature, rather than a bug. As a discipline, archaeology should encourage such diversity and experimentation because the theoretical and methodological challenges inherent in making sense of data are as rich as any other research program.

[20] http://heuristnetwork.org
[21] http://archesproject.org/
[22] See examples: Mantis: http://numismatics.org/search/ and Nomisma: http://nomisma.org

The need to respect and encourage continued thought, experimentation, and intellectual freedom in creating and using archaeological data underlies choices in Open Context's design. A key aspect of Open Context's approach to technology centers on interoperability, the capacity of an information system to efficiently exchange data with other information systems. The Web now boasts a tremendous wealth of cultural heritage data and talent invested by institutions world-wide. To promote collaboration with these distributed efforts, Open Context focuses technical developments in two key areas:

- Application Program Interfaces (APIs): APIs enable people with some technical skills to easily combine information from different online sources to use in novel user interfaces, visualization and analysis. They offer flexibility and customization so that data are not trapped in one website (see Kansa & Kansa 2011). APIs also allow us to combine different information systems together in a Lego-like manner. For example, Open Context uses APIs to archive data with the University of California's California Digital Library (CDL), thus facilitating long-term digital preservation. Enhancement of Open Context's APIs, especially for entity reconciliation (see 3.8 below), will greatly multiple the impact of this project.
- Linked Open Data: As discussed above, Linked Open Data represents current best practice to communicate the meaning of data on the Web. While APIs enable information to flow across systems, LOD uses links to further define data. For example, Open Context links to an online gazetteer to note that a certain coin was minted in the ancient city of Rome and not the modern town of Rome, Georgia. Open Context editors work with contributors to use SKOS properties like "close match" and "has broader" to model correspondences between a given researcher's own project-specific terminologies and concepts used by wider communities.

### 3.7.3 Evaluating Field Data Management Tools

This project will explore data modeling and recording practices used for diverse and evolving research designs and with different typology and terminology systems. In meeting data modeling challenges, archaeologists organize data using commercial "off-the-shelf" data management tools (spreadsheets, relational databases, GIS) and data management software specifically designed for archaeology. Archaeological informatics specialists designing discipline-specific data management tools tend to implement very different underlying data models than typically used with commercial database applications deployed informally by field archaeologists (see Schloen 2001). For instance, many archaeologists informally model their data in tabular structures specific to their own recording systems in spreadsheets like Excel or in relational database systems (Filemaker, Access) or GIS applications (ArcGIS, qGIS). In contrast, special purpose archaeological data management systems typically implement more abstract, formalized, and flexible data models to accommodate a wider range of recording systems. These more abstract data models can be implemented on a variety of software platforms, including relatively new "noSQL" databases, special linked data-stores (also called "triple stores"), and standard relational database applications. ArkDB, OpenDig, Open Context (Kansa & Kansa, 2011), Heurist, OCHRE[23], and FAIMS are all open source systems now using abstract data models to manage archaeological data of great diversity and widely varying conventions for description.

This project will help document the advantages and disadvantages of off-the-shelf databases versus more special-purpose data management strategies. Advantages of systems specifically design for archaeology may be lost because of practical factors. Only ArkDB, OpenDig and FAIMS directly support in-field data collection. Even so, using these systems requires deployment, configuration, and application-specific learning. In addition, highly abstracted data models are not as intuitive for some researchers, many of whom expect to work with familiar tabular structures like Excel. Many archaeologists may feel more comfortable with common commercial database management and office suite software. Since commonly used commercial software will likely play an important role in archaeology for some years, an important

[23] See: https://ochre.uchicago.edu/

goal of this project will be to investigate good workflow practices applicable to commonly used applications. No matter what database organization is used, archaeologists also need workflows to reliably manage data in relation to other digital content including images archives, field notes, geospatial data, remote sensing, and instrument outputs (XRF and the like) (see Levy et al. 2010).

### 3.7.4 Survey Implementation and Evaluation

Austin's primary tasks will be collecting, managing, and analyzing all the data related to the project. In order to complete these tasks, she will undergo appropriate human subjects training at Stanford University. Faniel will work closely with Austin on data collection, management, and analysis. Austin's tasks will include developing the project's Institutional Review Board (IRB) application, developing interview and observation protocols, conducting interviews and observations, sending audio recordings out to be transcribed, reviewing transcripts for accuracy, analyzing data, and disseminating results. Faniel will provide Austin with hands-on training in data analysis techniques and related qualitative data analysis software (e.g. NVivo). Austin will help develop a codebook for analyzing interview and observation transcripts, calculate inter-rater reliability, code interviews and observations, and run queries in NVivo to identify patterns and themes in the data. She will be trained in existing data vocabularies and structures for each research project, including Open Context, Filemaker Pro, the OASIS archaeological index, and the Vitor database. Revisions to ontologies and Linked Data standards will be documented in GitHub.

### 3.7.5 Long-term Support and Sustainability

Institutional support for grant products is provided by the CDL (data archiving, preservation, and migration) and the German Archaeological Institute (mirror hosting). Use of a GitHub repository provides open tracking of software developments. Updates and revisions to Open Context data publications are recorded in GitHub and are noted on the dataset. Previous versions of datasets are archived, and the current version is shown in Open Context (similar to editions of a book). All data publications in this project will be archived with the CDL, using Open Context's well established data archiving protocol. All data are open access and are annotated with Linked Open Data to facilitate their use, interoperability, and longevity. For more information about data standards, data formats, access to humanities collections, and long-term preservation of grant products (including survey results), see section 5.4 (Data Management Plan) and section 3.4.1.5 (Data preservation and dissemination).

Continuity and sustainability of Open Context's operations represents a long-term challenge. Open Context receives financial support by charging fees for grant data management. Additional financial support for the AAI comes from grants and from technology and information management consulting services. This mixed financial strategy, coupled with low institutional overhead and costs, has supported 12 years of continual operation. Nevertheless, even if Open Context ceases publishing activities, all content will still be freely available thanks to archiving with the CDL's institutional repository.

## 3.8 Dissemination and Intended Audience

### 3.8.1 Dissemination

This project brings together expertise in data curation and reuse with expertise in field data creation. The overall goal is to demonstrate and promote practices that streamline data creation so that data collected in the field can be better understood, adopted and reused by a wider community of archaeologists. This will provide guidance for data management plans, and thus offer both substance and means for individuals and projects to implement this important but poorly understood documentation required by many funders. Project outcomes will be disseminated in the following ways (see additional details on project outcomes and their evaluation in the **Dissemination Plan** in section 5.3):

- A **white paper** providing considerations for data management plan development and review in archaeology. [Evaluation: Web impact metrics and more formal academic citation of the guidelines]
- **Build data management plan adviser tool** (DMP-adviser tool), which will offer advice on specific technical standards and data modeling approaches, and offer links to specialized data management

tools (ARK, OpenDig, FAIMS), all tailored in response to a user's input. This new DMP-adviser will use search APIs (application program interfaces) from Open Context, tDAR, and others, to query for controlled vocabularies relevant to a user's research interests. As archaeologists publish more controlled vocabularies using open standards, the DMP-adviser can inform users of relevant standards in more areas of topical specialization. Thus, data and vocabulary sharing can feed back to inform future data creation, and shared data and vocabularies will see greater research impact. Using this tool will in no way require use of Open Context as a repository. Users may take what they learn and archive data with other repositories, including tDAR. The DMP-adviser tool will be available on Open Context's website. The controlled vocabularies the tool points to are editorially-curated and available either in Open Context (related to published projects) or across the web from various projects and institutions (such as vocabularies to describe coins, curated by the American Numismatic Society). [Evaluation: Web usage metrics, new datasets referencing recommended standards]

- **Controlled vocabularies** expressed in open computational standards (SKOS and OWL), which will help document data and develop the basis of standards needed for semantic alignment and integration of data. These vocabularies can be used by anyone on the Web, not just Open Context. As is the case with datasets, each researcher defined controlled vocabulary published by Open Context will be a citable scholarly work (with DOI assignment) with impact tracked via citation bibliometrics. [Evaluation: Web usage metrics, development of linked datasets referencing recommended standards, academic citation]
- **FAIMS reuse:** Researcher-defined data models and controlled vocabularies will be published with Open Context and loaded to the FAIMS project, allowing FAIMS users to reuse and adapt descriptive systems created by others. [Evaluation: # of FAIMS projects using Open Context descriptions]
- **Conventional publications and presentations** to communicate this work among the archaeology, information science and data curation communities. [Evaluation: Peer-review; citation impacts]
- **High-quality, open archaeological data** published by this project in a variety of open and widely-used formats (JSON-LD, CSV) will add to larger bodies of related, comparable data already available openly through Open Context and related open data and linked data initiatives worldwide. DOI assignment via the EZID system will enable bibliometric impact tracking. [Evaluation: Web usage metrics, citation of published data sets]
- **Innovative uses of open data, vocabularies, and APIs** developed during this project. As Linked Open Data continues to grow, we anticipate increasing uses of the content this project will produce. Some examples of current uses include: (1) rOpenSci sponsored an R statistical package that uses Open Context's new RESTful API (https://github.com/ropensci/opencontext); (2) CDL uses Open Context's API to ingest data for archiving; (3) archaeologist Shawn Graham's recent topic modeling of field notes published in Open Context used the new API (http://rpubs.com/shawngraham/79365). [Evaluation: Number of outside projects drawing on data, vocabularies, and APIs]
- **Context-aware entity reconciliation services** represent one of the most significant outcomes of this project. Entity reconciliation involves finding records in a data set that cross reference with entities in another data source. Entity reconciliation can improve overall data quality and interoperability by making cross-references in datasets explicit. For example, Open Context already supports entity reconciliation of North American archaeological site records published by the DINAA project. Researchers can query Open Context to get URIs associated with Smithsonian trinomials (an identifier system widely used in N. America). By getting URIs for Smithsonian trinomials, they relate their own data to DINAA data, thereby associating their data with DINAA-curated metadata, including site type, chronological, and geospatial information[24]. Open Context is ideally suited to offer powerful entity reconciliation services. Open Context publishes datasets and controlled vocabularies annotated with geospatial, temporal, and author metadata. These metadata provided context needed for greater precision and reliability in entity reconciliation. Such context is important.

[24] To protect site security, we only offer geospatial data at low levels of precision, see Wells et al. (2014).

For example, a user wanting to reconcile the term "sheep" with URIs for appropriate biological taxa would need to get different results depending on context. In data documenting N. America before European contact, "sheep" should link with the URI for *Ovis canadensis* (big horn sheep), not *Ovis aries* (Old World sheep). Similarly, "red-slipped pottery" can relate to several different types in different typological systems around the world, making context essential to successful entity reconciliation. Finally, author metadata can similarly play a role in contextualizing reconciliation services. For example, a researcher may choose to link their ceramic typology dataset with the controlled vocabulary created by one research and not another. Developing such context-aware reconciliation services for Open Context will be invaluable for data creators, data re-users, and data managers alike. [Evaluation: User feedback, number of uses]

- **API entity reconciliation "recipes"**[25] will multiply the impacts of shared data and vocabularies, if they are easy to understand and use. This project will develop "recipes" to use the Open Context API for entity reconciliation with Open Refine, a popular open-source data cleanup tool. Such recipes will guide users without programming backgrounds in reconciliation. Our qualitative user needs studies will inform us about what sorts of reconciliation services and recipes to emphasize. [Evaluation: User feedback, number of uses of API recipes]

### 3.8.2 AUDIENCE AND IMPACTS

This project's results have a global reach, including archaeologists and scholars in related fields (museums, libraries and archives) who collect cultural heritage data, those who wish to discover data, and publishers linking to web-published datasets. One of the unique aspects of this project is that it works to not only publish data, but also data models and controlled vocabularies. Widening participation in data modeling and semantics broadens intellectual engagement in communicating and understanding data. However, we recognize that most researchers will not create SKOS vocabularies on their own. Our promotion of expert professional services to aid data management highlights key needs in many disciplines. Our project offers much-needed institutional support for researchers engaging with data.

By involving the wider research community in improving data collection practices and collaborating by sharing datasets and data models, this project will help more firmly establish data sharing as a regular part of professional practice. Access to comparative data published by this project will create new research opportunities unavailable to researchers considering datasets in isolation, thereby helping to motivate more data sharing. Moreover, Project Director S. Kansa, as Vice President of the International Council of Archaeozoology (ICAZ) and member of the publication committees for both the American Schools of Oriental Research (ASOR) and the Society for American Archaeology (SAA), can widely promote project outcomes in key professional venues. Technology Director E. Kansa also participates in Web standards development (GeoJSON-LD) on location-based service design with the W3C. Exploring the data modeling challenges of archaeology advances geospatial informatics, more generally. This benefits wider communities, including the commercial and open-source technology sectors, including:

- *Educational Opportunities*: The data created at the three field sites will all be published open access, free of copyright restrictions, and using widely-accepted open standards and formats. These data can be used without restriction by researchers, students and the public for future studies, training in data analysis methods (bridging the artificial divides between the humanities and STEM fields), and translation to local languages, exhibition, and other forms of public engagement.
- *Professional Development*: This project's promotion of better data creation practices will guide archaeologists in improving practice, and will enhance the quality of data management plans.
- *Graduate Training*: Open access dissemination of good data management guides will also improve graduate education in archaeology, particularly in the digital humanities.

[25] http://opencontext.dainst.org/about/recipes

- *Informatics Mentoring*: As discussed, archaeology faces tremendous information management challenges. The project will offer excellent interdisciplinary mentorship opportunities for postdoctoral researcher A. Austin to closely collaborate with library and information science researchers (I. Faniel, E. Yakel) and leading practitioners in digital archaeology (S. Kansa, E. Kansa).
- *Field Training*: The field school programs run by IFR (led by Boytner) will provide excellent "hands-on" training opportunities in good data-management at both the graduate and undergraduate level.

Promotion of excellence in practice also motivates the Institute for Field Research's (IFR) participation in this project. IFR is a non-profit organization that offers archaeology field research courses at sites around the world. IFR requires extensive peer-review of field methods and practices of projects seeking IFR affiliation. IFR sees data management as a key area of need. In identifying good data management practices, this project will help IFR improve peer-review processes with respect to data management and long term data curation. These efforts will increase the prestige of data management and further promote its professional recognition, which will in turn motivate continued and sustained intellectual investment in data management and curation. These efforts will make data best practices less abstract and more closely aligned with professional goals. Understanding and promoting data creation and long term preservation practices that better promote effective data reuse will advance the mission of both IFR and Open Context.

Open Context has played an important leadership role in research data management. In advancing a model of "data sharing as publication," use of GitHub for dataset version control, approaches toward data modeling, Open Context serves as a case study for RECODE[26], a major project developing research data management policies for the European Union. The archaeological community itself recognizes these achievements, as demonstrated by Co-I E. Kansa presenting a keynote address to the 2013 Computer Applications in Archaeology conference in Perth, Australia. Similarly, our team collaborates (both technically and on advisory boards) with a network of allied projects, including the FAIMS project (discussed above), ARCS (an NEH-funded project to share legacy field notes in archaeology), the German Archaeological Institute (DAI) IANUS project (developing a national archaeological data repository for Germany), PeriodO (an NEH-funded project developing linked data around time periods), and tDAR. Thus, outcomes of this project will help inform archaeological data preservation and access across many allied efforts, both nationally and internationally.

Widening the community of scholars skilled in creating high-quality data, authoring ontologies, and using these information systems, will mean more users and research impact for systems across the entire "ecosystem" of archaeological information. Archaeology needs a thriving information ecosystem with many teams engaged in innovative projects. A diversity of perspectives should be seen as a "feature" rather than a "bug" because archaeological data management issues involve significant theoretical, practical and technological challenges. These intellectual challenges are as rich and deep as any other archaeological research question, necessitating a wide variety of perspectives and experiments. Though Open Context may grow more slowly than conventional repositories (in terms of numbers of datasets), the data and vocabularies it does publish can have more immediate value. Open Context data are immediately ready for linking, interoperability, and entity reconciliation, which can augment quality and interoperability broadly for users of other datasets and information systems. Thus, Open Context's "value-added" investments in certain datasets will make it easier to enrich other data on and off the Web.

This project will help Open Context meet longer-term goals of promoting greater professional recognition for, and intellectual engagement with, digital data in archaeology. Open Context's model of "data sharing as publication" tries to better situate data sharing with professional rewards and incentive structures (Kansa 2012; Kansa & Kansa 2013). Ultimately, data dissemination needs to lead to research outcomes in order to sustain interest and investment by the scholarly community. Demonstrating how data excellence leads to excellence in research outcomes will help to build needed professional recognition.

[26] http://recodeproject.eu/