Original Study **Open Access**

Keith W. Kintigh

Extracting Information from Archaeological Texts

Abstract: To address archaeology's most pressing substantive challenges, researchers must discover, access, and extract information contained in the reports and articles that codify so much of archaeology's knowledge. These efforts will require application of existing and emerging natural language processing technologies to extensive digital corpora. Automated classification can enable development of metadata needed for the discovery of relevant documents. Although it is even more technically challenging, automated extraction of and reasoning with information from texts can provide urgently needed access to contextualized information within documents. Effective automated translation is needed for scholars to benefit from research published in other languages.

Keywords: synthesis; digital repositories; natural language processing; automated translation; automated reasoning

DOI 10.1515/opar-2015-0004
Received December 19, 2014; accepted February 11, 2015

If we are to know what are the grand challenges of digital archaeology, we must know what are digital archaeology's objectives. I believe that the overarching goal of digital archaeology is to advance the use of computational and formal methods to serve the goals of archaeology more broadly—which is to say to achieve systematic understandings of human societies based on the material remains they leave behind. Moving forward from that premise, digital archaeology's challenge is to develop and apply computational tools that allow us better to address archaeology's most important questions. So, what are they?

I recently led an effort to define grand challenges for archaeology. We crowd-sourced suggestions through email requests and listserve postings by the major North American and European professional organizations. Through a workshop at the Santa Fe Institute, a group of 15 distinguished scholars with diverse interests and orientations then augmented, refined, and prioritized the crowd-sourced suggestions, yielding a set of 25 grand challenges. These challenges were published in *American Antiquity* [1], and for the wider scientific community in the *Proceedings of the National Academy of Sciences* [2].

The resulting grand challenges are not unique to archaeology; rather they are social science questions whose answers require knowledge on temporal and spatial scales that only archaeology can provide. They focus on the dynamics of cultural processes and are based on a conviction that understanding the cultural dynamics we observe today demands deciphering the long-term histories that produced them. While one might legitimately quibble with the particulars of the list, it is quite clear that transformative progress in archaeology demands a stronger focus on synthetic research by archaeologists.

A number of factors conspire to frustrate synthetic research. They include the problems of discovery and access to archaeological data, the difficulty of integrating data from diverse sources, and the problem

Article note: This article is a part of Topical Issue on Challenging Digital Archaeology.

***Corresponding author: Keith W. Kintigh:** School of Human Evolution & Social Change, Box 872402, Arizona State University, Tempe AZ 85282-4002 USA, Office: +1 480 965 6909 Email: Kintigh@asu.edu

of extracting usable data, information, and knowledge from text. Here, I will focus on the last of these, the extraction of knowledge from archaeological texts.

Discovery, Access, and Digital Repositories

Enormous quantities of archaeological information and knowledge are embedded in often-lengthy reports and journal articles. A substantial fraction of journal content is now available digitally through JSTOR and commercial publishers. While articles are important sources, vast amounts of critical information reside only in hundreds of thousands of gray literature reports, of which only a tiny fraction is digitally accessible. These reports often constitute the only available documentation of the excavation of important sites that are now thoroughly excavated, destroyed, or otherwise unavailable. In the US alone, approximately 45,000 cultural heritage field investigations are conducted annually [3] at a cost of about a billion dollars [4]. Each of these produces a report, ranging from a short letter to a multi-volume report with a more than 1000 pages. These reports are filled with data tables, descriptions and interpretations of archaeological contexts and finds. They have tremendous scientific potential that must be leveraged to advance our knowledge and understanding of the social world.

Addressing archaeology's substantive research challenges demands an ability to extract knowledge from these textual sources. First, of course, these reports must be in digital form and they must be readily accessible. Furthermore, a substantial corpus related to the subject of interest is necessary to successfully apply many of the automated approaches discussed below. While the challenges of acquiring archaeological texts (and other data) in digital form, preserving them, and making them accessible are daunting, the problems are primarily social and economic. In fact, we know a lot about how to do these things [5] and have established useful digital repositories for those purposes (e.g., ADS, tDAR, and DANS). However, an enormous amount of work, requiring large-scale funding, remains to be done to make important legacy reports digitally accessible. As a start, the institutions funding or permitting the work should require that newly created reports be moved into one of these digital archives.

Having searchable documents accessible, of course, does not entirely solve the problem. It must be possible to exclude irrelevant documents and to identify relevant documents and extract useful information from them. To leverage the knowledge embedded within those texts, we must exploit existing and emerging technologies to allow sophisticated discovery of and information extraction from archaeological texts.

Word Searches

Simple word searches of full text documents are useful to the extent that they can substantially focus investigation on a much smaller corpus. For example, a colleague's current study of Southwest-Mesoamerica connections is seeking to understand the distribution within the prehistoric Southwest US of pyrite mirrors that originated in what is now Mexico. A tDAR search for documents from the Southwest yields 98,000 citations and 1055 documents. Further investigation indicates that about 550 of the 1000+ documents have searchable text. Searching tDAR for "pyrite mirrors" yields 6 reports. To be a bit more cautious we might search separately for "pyrite" (23 hits) and "mirror" (152 hits, the larger number probably reflecting the use of "mirror" as a verb). Realistically, the search reduces our investigation from 550 to fewer than 25 reports that need to be examined. Despite its obvious value in some instances, the limitations of word searches are readily apparent to anyone who has used an Internet search engine. Word searches would not be very helpful, for example in looking for descriptions of excavated room floors found with *in situ* ceramics (because the key terms are all extremely common in archaeological reports).

Digital Humanities Tools

Tools employed in digital humanities and qualitative data analysis research, such as NVIVO may be used to characterize and further investigate a corpus of documents. Word frequency analyses [6] or analyses of citations through time might usefully illuminate theoretical trends in the discipline. However, these analyses are directed to different sorts of research objectives than I am attempting here to address.

I believe that the development and application of sophisticated tools for natural language understanding will be essential for the synthetic research needs of the discipline. Unfortunately, natural language understanding has turned out to be a much more difficult problem than computer scientists anticipated in the 1960s and we still lack reliable general- purpose tools. However, we can capitalize on the substantial progress that has been made in natural language processing research.

Automated Classification

Richards and his colleagues [7] summarize current progress on automated classification of articles and reports. The idea is to use natural language processing tools to extract basic indexing information— what, where, and when—from the text of an article or report, in order that a large corpus of digital documents that lack human-generated metadata can be effectively searched for subject, location, and temporal period. In order to do this effectively, it is necessary to discriminate the topics, places and times that describe the thrust of the text, from those that are simply referred to in the text.

As Richards et al. indicate, both rule-based and machine learning approaches have been used with considerable success [8, 9]. However, it should be emphasized that there are no "plug and play" tools for this task. Recent applications have been serious research efforts involving substantial efforts by teams of archaeologists working with computer scientists. Different approaches have different demands, but they variously require the development of hand-crafted rules, training sets of human-annotated documents, and development of specialized glossaries, gazeteers, and thesauri of relevant terms.

Knowledge Extraction

Automated extraction of knowledge from texts and reasoning with that knowledge is, of course, even more difficult than automated classification, though it builds on some of the same technologies. I can offer no examples of successful applications in archaeology, but there is promising work on this topic, for example, in biology [10, 11]. In the remainder of this article I will attempt to indicate what might be possible and some of what might be required to get there.

I suspect that the preponderance of scientific value will derive from published articles and longer documents, including multi-volume reports with hundreds or even thousands of pages. Let's say we want to find in the literature descriptions of:

> 12th century excavated pitstructures from New Mexico or Arizona that have a southern recess and are associated with above-ground pueblos with 10 or more rooms

My query thus has several requirements:

Date: AD 1100-1199
Location: Arizona or New Mexico
Other terms: pitstructure with southern recess, excavated, associated with a pueblo with 10 or more rooms

Given that longer reports commonly describe multiple architectural units within several sites, it should be clear that a Google-like search is not going to cut it.

Most statements of interest, such as those relating architectural details (Figure 1) or radiocarbon dates are of scientific interest *with respect to a particular context*. For archaeologists, the context is the set of salient characteristics of the place in which the object was found, how it was recovered, and what other

entities were collocated with it. Thus, the same date derived from charcoal flecks found in the sediment filling a prehistoric canal will be of quite different interest from one that derives from burned maize found in a fire pit of a thousand year old house. In these examples, what is being dated are the use of a canal and the end of occupation of the house, respectively.

Main Chamber Bench. Southern Bench, and Vent System. The main chamber was encircled by a bench or shelf located 1.00 m above the floor. Use of the term bench is qualified because this main chamber bench is only 0.17 m wide; so while it might have been used as a storage area for small items, it was not large enough to sit or lie upon. This bench or shelf might have functioned more as an architectural feature, a buttress, stabilizing the upper 1.00 m of walls by setting them back slightly from the lower walls.

On the south end of the structure was a deep bench that measured 2.05 m wide where it interfaced with the mainchamber, 2.88 m at the back of the bench, and 1.00 m from the front to the back of the bench. The surface of the bench was 1.00 m above the floor of the main chamber. Benches such as these with outward-flaring walls located on the south side of the pitstructure are often termed a southern recess (McKenna and Truell 1986:197) and are common features in pitstructures dating from the late AD 1000s to the mid-1100s. Pitstructures that have this deeply recessed southern bench are often termed "keyhole shaped" and have been interpreted in the past, perhaps erroneously, as a Mesa Verdean architectural characteristic (McKenna and Truell 1986:224). A discussion of pitstructure shape is presented in greater detail in Chapter 16.

Figure 1: Sample Section of Report Text from Howell [12].

A key challenge to factual knowledge extraction in archaeological reports is that the *scope conditions* of a natural language statement are rarely contained in or even proximate to the statement but must instead be inferred from a hierarchy of chapters and multiple levels of section headings. Many key relationships may never be stated directly in words but are essentially inherited. Figure 1 shows a section of text in a report and Figure 2 lists the five levels of superior headings that contextualize these paragraphs at this sixth evel (page 101). Apart from the semantic complexity of the text itself, to simply know that this architectural description is *of* Unit 6 which is *part of* site NM:12:K3:101, it is necessary to understand the heading hierarchy. The table of contents, of course, is helpful but it includes only the top four of the six heading levels that are operative here. Note further that while the sites will have unique numbers, every site will have one or more horizontal subdivisions called "units" (such as this structure), each starting with Unit 1.

The Archaeology and Ethnohistory of Oak Wash, Zuni Indian Reservation, New Mexico	
II. Site Descriptions	49
Site NM:12:K3:101 Description	93
Pitstructures	96
Unit Descriptions	98
Unit 6	98
Main Chamber Bench, Southern Bench, and Vent System	101
Dates	103
Site NM:12:K3:102 Description	116

Figure 2: Hierarchy of section headings in the report [12] contextualizing Figure 1.

Other sections of this site's chapter will identify key relationships of the site (such as the fact that this particular site was "partially excavated" and that it has an above-ground "pueblo" with about 20 "rooms"). Since site NM:12:K3:101 contains the architectural unit described above, those characteristics further contextualize Unit 6. Going the other direction, later in the report, dates of AD 1101-1131 are reported for Unit 6. Although these dates are specific to Unit 6, since Unit 6 is a part of this NM:12:K3:101, the site's occupation must contain those dates but is not restricted to them.

Recalling our question, "What 12th century excavated pitstructures from New Mexico or Arizona that have a southern recess and are associated with above-ground pueblos' with 10 or more rooms?" we can see that all of the relevant information is present in the report but it is structured in complex ways. To properly represent the knowledge requires not just recognition of *nested relationships* but also substantial *reasoning* to properly assess the information.

Obviously, this is difficult. However, computer scientists are developing ways in which these problems can be addressed. One approach with considerable promise requires preprocessing the text, to translate knowledge written in natural language into a state-of-the-art knowledge representation language. Once the factual knowledge in the text is represented in a knowledge representation language, it can be queried by machine reasoning based on formalized basic principles of archaeology (*generic knowledge*).

This generic knowledge is also expressed in the literature, for example

> ... three basic functional room types found in modern pueblos: habitation, storage, and ceremonial. Habitation and storage rooms showed distinct differences in architectural attributes, such as room size, the location of openings, and the presence of non-portable facilities, like hearths and mealing bins. In contrast, masonry benches, subfloor vents, and hearths with deflectors were used to identify ceremonial or religious rooms [13].

To apply this technology we would need to develop a formalized base of generic knowledge about archaeology (and probably regionally specific variants). At the time a digitized report or article was ingested into the system, it would be automatically translated into a knowledge representation language. This translation would have to account for the structure of the text, which in substantial measure is conveyed by hierarchical headings, as described above. Finally, when a query is received, the system would query the factual knowledge base of the texts using machine reasoning based on the generic knowledge base.

Question Answering

IBM's Watson Ecosystem [14] has demonstrated impressive capabilities in reading, understanding, learning, and reasoning with unstructured natural language text. It is now possible to experiment with an instance of Watson delivered through the Cloud. Experimenting with Watson's question-answering capabilities once it has "read" a corpus of archaeological literature would make it possible to assess the practical, present-day usability of this technology and to get a sense of its strengths and weaknesses.

Leveraging Text in Multiple Languages

To this point, I have not dealt with the problem of documents in languages other than English. Thoroughly addressing many of the important questions in archaeology will require that we take advantage of literature in more than one language. The first step, of course, is for digital repositories and associated text-processing tools to adequately support the use non-English (including Asian) characters (see http://www.unicode.org/).

Automated translation, of course would greatly aid human users, both directly and by effectively extending the application of automated classification, knowledge extraction and other text processing tools to include documents in additional languages. Turning this around, to the extent that we have good semantic representations of text in any language (as discussed under Knowledge Extraction, above), it would greatly facilitate high quality automated translation to other languages.

Discussion

Extracting information from archaeological texts using natural language processing is a major challenge for digital archaeology. Because so much of the discipline's knowledge is codified exclusively in reports and articles, the ability to extract information from text and to reason with it is essential for the synthetic research needed to address archaeology's most pressing substantive challenges. As we consider new investments in archaeology's computational and social infrastructure, the development of digital archives of archaeological literature and work on natural language processing should be an important priority for digital archaeology.

Acknowledgments: The author is grateful to Chitta Baral and Julian Richards for their productive collaborations on this issue and their comments on a draft of this article. Thanks are also due two anonymous reviewers who provided helpful suggestions.

References

[1] Kintigh, K.W., Altschul, J.H., Beaudry, M.C., Drennan, R.D., Kinzig, A.P., Kohler, T.A., et al., Grand challenges for archaeology, American Antiquity 2014, 79(1), 5-24
[2] Kintigh, K.W., Altschul, J.H., Beaudry, M.C., Drennan, R.D., Kinzig, A.P., Kohler, T.A., et al., Grand challenges for archaeology, Proceedings of the National Academy of Sciences, 2014, 111(3), 879-88
[3] Departmental Consulting Archeologist, The Secretary's report to Congress on the Federal Archeological Program, comparable SRC data 1985-2012, by year. Archeology Program, National Park Service, Washington, DC, http://www.nps.gov/archeology/SRC/data.htm, 2015
[4] Altschul, J. H., Patterson, T.C., Trends in employment and training in American archaeology. In: Ashmore, W., Lippert, D., Mills, B.J. (Eds.), Voices in American archaeology, Society for American Archaeology, Washington DC, 2010
[5] Archaeology Data Service, Center for Digital Antiquity, Caring for digital data in archaeology: a guide to good practice. Oxbow Books, Oxford, UK, 2013 http://guides.archaeologydataservice.ac.uk/
[6] Michel, J., Shen, Y.K., Aiden, A.P., Veres, A., Gray, M.K., the Google Books Team, et al., Quantitative analysis of culture using millions of digitized books, Science, 331, 176-1812, 2011, DOI: 10.1126/science.1199644
[7] Richards, J., Tudhope, D., Vlachidis, A., Text Mining in archaeology: extracting information from archaeological reports, In: Barceló, J.A., Bogdanovic, I., (Eds.), Mathematics in archaeology, Science Publishers, Boca Raton, Florida, (in press)
[8] Jeffrey, S., Richards, J.D., Ciravegna, F., Waller, S., Chapman, S., Zhang, Z., The Archaeotools project: faceted classification and natural language processing in an archaeological context. In: Coveney P., (Ed), Crossing boundaries: computational science, e-Science and global e-infrastructures, Philosophical Transactions of the Royal Society A, 2009, 367, 2507-2519.
[9] Tudhope, D., May, K., Binding, C., Vlachidis, A., Connecting archaeological data and grey literature via semantic cross search, Internet Archaeology 2011, 30, DOI: 10.11141/ia.30.5
[10] Baral, C., Chancellor, K., Tran, N., Tran, N.L., Joy, A, Berens, M., A knowledge based approach for representing and reasoning about signaling networks, Bioinformatics 2004, 20, no. suppl 1 (2004), i15-i22.
[11] Tran, N., Baral, C., Nagaraj, V.J., Joshi, L., Knowledge-based framework for hypothesis formation in biochemical networks, Bioinformatics, 2005, 21, supplement 2, ii213–219.
[12] Howell, T.L., The archaeology and ethnohistory of Oak Wash, Zuni Indian Reservation, New Mexico, Zuni Cultural Resource Enterprise Report, 2000, 644, Zuni, New Mexico
[13] Clark, T.C., Assessing room function using unmodified faunal bone: a case study from east-central Arizona, Kiva, 1998, 64(1), 27–51.
[14] IBM What is Watson? 2014 http://www.ibm.com/smarterplanet/us/en/ibmwatson/what-is-watson.html