CHNT 18 Proceedings

Archaeology and Analytics:

Tapping the Pulse of Social Media for Outreach, Education, and the Future of the Field

Kathryn CHEW

Cotsen Institute of Archaeology, University of California, Los Angeles (UCLA), Los Angeles, USA

Abstract: Among the various academic disciplines, archaeology is one which has held a particular romantic sway over the public imagination. Such fictional archaeologists as Indiana Jones and Lara Croft have long occupied the attention of the Western media consumer, representing archetypes of adventurers, mavericks, romantic and sexual ideals, and the relatable side of intellectualism. Myriad documentaries and fictitious works about archaeology and archaeological inquiries attest to this public preoccupation. Yet, despite this disproportionate share of the general public's attention, it is unclear how much of this attention is focused on the fruits of academic research rather than the fruits of entertainment media. By extension, the degree to which the public understands either the discipline of archaeology as it is actually practiced or the data produced through archaeological research is uncertain. I seek to shed light on the apparent disconnect between the public discourse on archaeology and the understanding of it as held by participants in the field by examining the public discourse where it happens: in social media. Utilizing targeted data streams harvested from the Facebook API and applying TF-IDF textual analysis, I attempt to discern what differences can be found between the public discourse and that of professional archaeologists. By identifying the trends embedded in this divergence, I provide a framework within which the relationship between the academy and the outside world can be contextualized in order to more clearly assess the effectiveness of public outreach efforts and better target educational resources in the future.

Keywords: Archaeology, Analytics, Social Media, TF-IDF, Facebook

Introduction

When the word "archaeology" is introduced into a conversation, a myriad of images are brought to mind: Indiana Jones, Lara Croft, River Song, the Mayan Apocalypse, aliens in flying saucers, pyramids, booby traps, and of course treasure — always treasure: golden, jewel-encrusted, ancient and valuable beyond imagining. One image which is virtually never evoked, particularly among those who are only casually interested in archaeology, is the image of actual, practical archaeology: academic, scholarly, and heavily tied to anthropological theory. Of all the academic disciplines, archaeology has long held a particular romantic sway over the public imagination [FAGAN 1996]. Consumers of Western media have long associated archaeology and its practitioners with the archetypes of adventurers, mavericks, romantic and sexual ideals, and the relatable side of intellectualism [MCGEOUGH 2006; POHL 1996].

However, this preoccupation with the romanticized fantasy of archaeology and archaeologists must coexist with the realities of the discipline. Archaeologists, like the members any other academic field, need to secure

funding for their projects, publish their findings, and in between undergo the slow, laborious process of data collection and processing. In recent years, the prevailing misperceptions of archaeology in the public consciousness have shifted in a potentially destructive direction. Although this disconnect between popular representations of archaeology and archaeological practice has been characterized as “an identity crisis” [LOWENKOPF 1996] which can be traced back through multiple centuries [MOSER 2009], the current ability of technology to quickly disseminate false and harmful ideas has allowed these misperceptions to proliferate as never before. In the United States, several reality television shows purporting to exploit the public glorification of archaeology have aired and subsequently been chastised by archaeological professionals for their encouragement of looting and their irresponsible treatment of the archaeological record [LIMP 2012; WYCKOFF 2012]. A similar mismatch of priorities in televised representations of archaeology has likewise been developing in Europe [TAYLOR 2001]. In addition, following the U.S. government shutdown in the latter half of 2012 as a result of a legislative budgetary impasse, government critics cited funding from America’s National Science Foundation [NSF] for archaeological research projects as “Symptomatic examples of ill-conceived scientific research priorities” [MULLINGS 2013], and encouraged the NSF to cease funding these types of projects. It is clear from examples such as these that many members of the American public, and the general public at large, do not understand the scientific underpinnings of archaeological research, and do not appreciate the contributions of this research to society [POHL 1996]. What is less clear is the full extent by which the public discourse regarding archaeology diverges from the professional discourse, as well as the extent to which any significant misunderstandings interfere with or harm the ability of archaeologists to conduct their work [MULLINGS 2013].

Many of the major scholarly institutions and academic funding bodies in the United States have expressed significant interest in engaging in public outreach and advancing the public good (See references: AIA, ARCE, NEH, NSF, SAA), and some archaeologists have called on their colleagues to begin examining public representations of archaeology and archaeological work “As a matter of disciplinary concern” [MOSER 2009]. To this end, a variety of recent research has focused on non-academic representations of the past and knowledge construction in archaeology [DAY 1997; FAGAN 1996; FINN 2001; GERO and ROOT 1990; KUHN 2002; LOWENKOPF 1996; MCGEOUGH 2006; MCMANUS 1996; MOSER 2001; MOSER 2009; POHL 1996; STONE and MOLYNEAUX 1994; TAYLOR 2001; ZORPIDU 2004]. These efforts have largely been limited to examinations of the more readily accessible qualitative data on the subject, however, perhaps neglecting the quantitative data because they are much more difficult to locate, extract, and operationalize in a useful way. In this paper I attempt to define a methodology for operationalizing the quantitative aspects of public perceptions and misperceptions regarding the discipline of archaeology, more clearly define the disconnect between the public and professional discourses, and to provide analytical insight into the quantitative data in order to clarify potential strategies to reduce these discursive disparities.

Quantitative Analysis of Social Media

Methodology

The data available through the Facebook API (Application Programming Interface) served as the primary focus of this study. An Application Programming Interface is a code-based interface that many software

companies provide to the public in order to allow independent programming teams to have access to certain strategic elements of their software code in order to allow these teams to develop third-party applications—such as games—that can interact with and be integrated into the primary software's code. APIs do not allow users to alter the base code, but given the correct access permissions, they are able to retrieve information from within the software's database and in some cases output that information into an external file for analysis [FACEBOOK 2014a]. This data source was selected because of the immense popularity of the Facebook social media application, with an associated user base spanning a wide variety of demographic categories. Facebook's own publicly reported statistics cite 1.23 billion unique monthly active users of its social media application as of December 31, 2013 [FACEBOOK 2014b]. It was also selected under the assumption that the information shared by individuals using Facebook would be an accurate reflection of the ideas and concepts both of greatest interest to them, and which occupied the greatest amount of attention for both these users and their own intended audiences.

Using a short piece of JavaScript code and an online Term Frequency-Inverse Document Frequency (TF-IDF) analyzer called Tagul, I was able to retrieve the textual components of four specific data streams and analyze them for the themes and keywords that occurred with the greatest frequency. TF-IDF, or Term Frequency-Inverse Document Frequency, is a form of textual analysis which uses an automated digital process to parse the content of a block of text in order to discern which terms occur most frequently throughout the text, and weights these against a list of terms (either generated from the block of text itself or taken from a common list) which occur with extreme frequency in the course of normal language usage because they are common functional words. The resulting output of TF-IDF analysis is a list of key terms which reflect the thematic content of a given block of text, ideally without the distortion that might otherwise be caused by the presence of grammatically helpful but analytically messy words such as "the" or "when".

I chose to use the online word cloud tool Tagul to complete my TF-IDF analysis of the textual data pulled from my four data streams because it provided automated tools for completing the analysis, including granular control of words omitted and key terms, which allowed me to partially compensate for one of the largest challenges I faced during this stage of the project—cleaning the data to remove nonsense characters that resulted from the digital transcription of text written using scripts other than the Latin alphabet, as well as multilingual duplicates of key words. Another significant benefit of the use of Tagul as a TF-IDF tool is its ability to output a manipulable word cloud that summarizes the most common key terms in a visually appealing and easily parsable way. This kind of visualization was immensely helpful in the process of conducting my analysis and synthesis of the data.

The four data streams I analyzed were as follows: (1) a collection of professional archaeologists (Fig. 1); (2) a collection of public "pages" run by or associated with professional archaeologists (Fig. 2); (3) the generic feed provided by the Facebook API consisting of all public posts made to Facebook by individual users (Fig. 3); and lastly, (4) a collection of public "pages" run by archaeological amateurs and associated with general public interest in archaeology (Fig. 4). The purpose of separating and focusing on these four streams of data was to attempt to obtain a broad sample of text that would reflect both the subjects under discussion in personal conversational contexts (data streams 1 and 3), and the subjects under discussion in more publicly-oriented and topic-specific contexts (data streams 2 and 4).

The target group which comprised data stream 1 contained over 100 individuals who were professionally affiliated with archaeology in some way. For the purposes of this study, “professional archaeologist” was defined as including individuals either presently or previously employed as an archaeologist, individuals who had studied archaeology at a postgraduate level, or individuals who were either presently or previously employed in a closely associated profession which would enable them to participate directly in the

Fig. 1 – Tagul TF-IDF visualization of data stream 1, collected from individual professional archaeologists.

Fig. 2 – Tagul TF-IDF visualization of data stream 2, collected from professionally-oriented Facebook pages.

professional discourse (e.g., conservators, field photographers, epigraphers, etc). The individuals in the professional target group were of a wide variety of nationalities and held a variety of regional specializations. In order to sample this data stream I collected the qualifying individuals into a project-specific “FriendsList” using the Facebook List feature. I could then use JavaScript commands and the Facebook API to pull data only from the members of this list and output it into a text file for analysis. The data I was able to access through this method consisted of the information made available to me through these professionals’ own privacy settings, and the text file output contained only a continuous stream of text with no identifying information, date stamps, or indications of the divisions between separate posts. This automated sampling process ensured that all the data collected was fully anonymized, and that the privacy and anonymity of these individuals was completely preserved.

The target groups for each of the page-based data streams consisted of approximately 10 pages each, with all pages featuring a primary focus on material directly related to archaeology. The professionally-oriented pages were all administered by either individual archaeological professionals from one of several countries, or representatives of scholarly projects or institutions. The target audience of each page was not necessarily other professionals in every case, but the content was always generated by professional archaeologists, per the definition provided above. The publicly-oriented pages were all administered by individuals without clear ties to archaeology as a profession or an academic discipline. The target audience for the public pages was invariably other members of the public. No organizations or institutions were represented in this target group.

Unfortunately, due to the more complex querying process required to sample data from pages, it was not practical to use a large and fully representative sample set of pages for either category within the limited scope of this preliminary study. In future iterations of this research this shortcoming will be corrected.
The generic feed was queried for all posts containing the key term “archaeology”, as it was one of the most frequent and most relevant of the key terms from the text output of data stream 4. The sampling process used to anonymize user-generated content for data stream 1 was also applied to the content in this data stream.

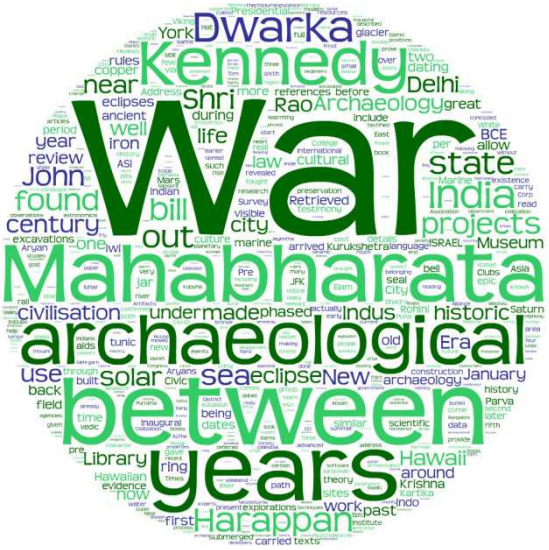

Fig. 3 – Tagul TF-IDF visualization of data stream 3, collected from all public Facebook posts.

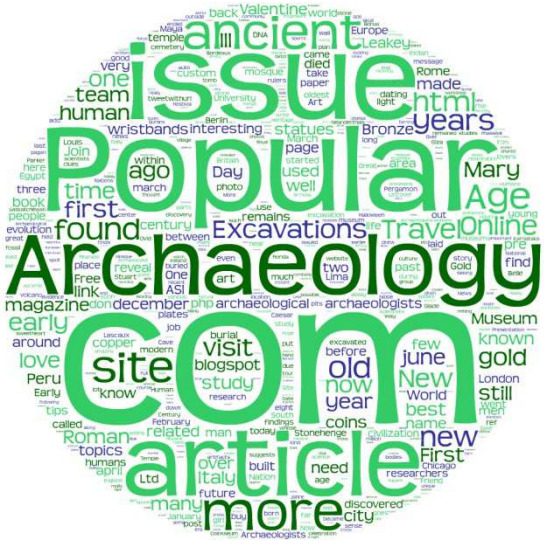

Fig. 4 – Tagul TF-IDF visualization of data stream 4, collected publicly-oriented Facebook pages.

These four streams were also selected in an attempt to compensate for the overall topically inconsistent and casual nature of conversation characteristic of Facebook as a medium. Data streams 2 and 4 were introduced after a preliminary investigation of the textual content of data stream 1 revealed an unexpectedly low volume of content directly relevant to archaeological themes.
During the examination of these four data streams, I pulled textual samples from two specific sets of content: “posts” and “statuses”. According to the code structure of Facebook’s data feed, “statuses” refers to the text-only “status updates” which individual users (or public page administrators) share directly onto their timeline feed. “Posts” refers to a much broader category of data which includes shared links, images and photos, along with any commentary either embedded with this shared information or added by the person broadcasting the content to their friends or followers. It was necessary to examine both of these content sets because although “status updates” might be considered historically more apt reflections of the thoughts and interests of Facebook users (to whatever extent the term “historically” might be applicable to Facebook, which was launched in 2004), since the addition of the “share” feature to Facebook’s repertoire of user interaction and information broadcasting, a significant amount of the content formation and opinion sharing the occurs on Facebook takes the form of annotation and commentary added to shared links and images broadcast by one user but originated by another. This added several levels of complexity to the collection and analysis of the data because shared posts contain many additional encoded fields of content than the

relatively straightforward status updates do. In the case of shared posts, not only must the appended "message" by the broadcasting user be collected and analyzed, but also the "name" and "description" of the shared object or link, as the commentary frequently takes the form of oblique reference to the content of the shared object or link, and rarely states explicitly what the content consists of. For the purposes of TF-IDF textual analysis, all keywords must be explicitly stated or will not be factored into the final analytical output. Therefore the additional content fields of the shared posts had to be included in order to allow for a more accurate representation of the topics under discussion.

Once the text for each data stream had been collected and the most common keywords ranked, I examined the TF-IDF output for notable patterns in thematic content, keyword similarity, and geographic emphasis. Through identification of areas of overlap or divergence between the output from each data stream, I was able to gauge the levels of discursive similarity between each source group and characterize these differences.

Analysis

When comparing the TF-IDF output of the four data streams, several patterns become evident. One of the most striking of these patterns is the prevalence of the word "archaeology" and its variants (i.e., "archaeological", "archaeologist", etc.) in data stream 4, the group of publicly-oriented pages. Data stream 4 featured by far the highest incidence of the term "archaeology" (including the alternate spelling "archeology") and its variants, whereas "archaeology" as a term appeared quite infrequently in the other data streams and its variants were virtually absent. This is particularly surprising given that the data pulled from data stream 3 was acquired specifically because it contained the key term "archaeology". This discrepancy might have several causes, though it seems most likely that its higher occurrence in one of the data streams associated with the general public and relative infrequency in the professional archaeologist streams is a result of the focus of the professional archaeologists on more specific topics within archaeology or a tacit understanding that their generated content falls under the topical heading of archaeology without having to state it explicitly. The general public data streams, by contrast, may have focused more on these terms specifically because without any additional formal affiliation with the subject matter, they must make their particular interest known by stating it. Another possibility is that the professionals are much more interested in sharing and commenting on topics relating to unusual research or more scientific interest, whereas the amateurs are more heavily focused on sharing updates about the basic activities of the professionals without much emphasis on the accompanying analysis (which might account for the incidence of such key terms as "excavations", "travel", and "study").

Another notable discrepancy between the professional and public data streams is the heavy focus of data stream 2 (generated from professionally-affiliated "pages") on professional and scholarly institutions, including the names of museums, universities, and other organizations. Data stream 4 (from publicly-affiliated "pages") instead focused much more on general terminology, such as place names. Data stream 2 also featured a comparable prevalence of the URL components ".org" and ".com" among the links shared to its audience, whereas data stream 4 featured an extremely high occurrence of the component ".com" and almost no occurrences of ".org". It seems likely that this is a result of the professional pages preferring to share links with correspondingly professional or scholarly sourcing and informational reliability, whereas the

amateur pages may concern themselves more with websites that address topics of greater general interest rather than the specificity or reliability of the content in the links shared.

A pattern also potentially of interest is the divergence of geographic focus among each of the four data streams. Each of the two professional streams focus most highly on Egypt, though this may be a result of the unstable political situation in Egypt as well as some sampling bias due to the focus of my own work on Egypt, which has resulted in my acquaintance with a higher proportion of Egyptologists. These professional streams also make reference to other regions of the world, though the discursive presence of these terms is significantly less. Data stream 4, by comparison, has a much more even distribution of geographic focus, with similar weight being given to terms related to Roman, Mayan, Egyptian, and Peruvian archaeology. The individual public data stream (data stream 3) has a significantly different focus from the others, with its geographic focus highlighting mostly terms related to archaeology on the Indian subcontinent, as well as some focus on Israel, Palestine, and Hawaii. There is also an extremely high and unexpected occurrence of terms related to deceased former U.S. President John F. Kennedy, which may indicate an interest in American historical archaeology, though it may also be a simple coincidence. Terms related to these regions do not appear at all in the other three data streams, a departure which may indicate that the general public has a much greater interest in archaeological ideas that carry prolonged contemporary political currency or appeal to issues of personal cultural affiliation, whereas professionals and dedicated amateurs are more interested in the “classic” ancient civilizations regardless of their ties to modern politics (although certainly the ongoing political upheaval in Egypt may play a part in the emphasis on Egypt-related terms, given its practical implications for archaeological study).

Among all four data streams there is a surprising, though not unwelcome, dearth of terminology related to most of the common conspiracy theories and pop cultural references associated with archaeology.

Discussion

Overall, the decision to focus on the content of digital social media as an access point into the public discourse seems to have been a successful strategy, as this avenue provided a significant volume of data directly related to public perceptions and information exchange. I would posit several caveats, however, which will need to be addressed in order to further develop and refine the effectiveness of the methodology outlined here prior to applying it to subsequent iterations of this research. One of the major concerns is the inconsistently representative nature of the accessible data. Although the total content posted by users to the Facebook platform may serve as a representative sample of a significant portion of the world’s population, not all of this content is available for study. In the data collection process utilized for this project, only two types of content were accessible: information consciously shared by users under a publicly accessible security setting, and information specifically shared with me but which may not have been made available to the public. Facebook has not published any statistics indicating what percentage of their total content is published under restricted privacy settings, so it is impossible to say whether the accessible content is as representative of the public as the demographic makeup of the whole user base may be. One potential solution to address this limitation would be to acquire comparative samples from alternative social media platforms which may not have as large a user base, but which encourage more public communication, such as Twitter or Reddit.

Another obstacle which bears addressing is the issue of consistent linguistic and cultural diversity among the users who comprised each data stream. The professional archaeologists (data stream 1), as previously stated, were of a variety of nationalities, and accordingly posted content in a plurality of languages. Where possible, all translations of a word were merged during the TF-IDF analysis so as to be more accurately represented in the resulting rankings and visualizations. Data streams 2 and 4, the group-oriented streams, were united by their topics of interest rather than specific terminology, and the posts likewise represented input from individuals of multiple nationalities and polyglot discussion. Within data stream 3 (the stream derived from individual public posts), however, this diversity was significantly harder to achieve, as the words themselves served as the means of locating and collecting data from individual posts. For the purposes of this preliminary study, only English search terms were used. For more accurate results, a list of representative terminology across several languages would need to be developed and applied across all data sources.

Finally, only a single round of data queries was utilized to populate each data stream. No further collection of data was undertaken for this study, as it was deemed necessary to develop an analytical workflow to process and interpret the data before the acquisition of additional material. In future, several rounds of data collection over a set period of time may be required in order to better account for fluctuations in the public discourse as well as larger discursive trends and reactions to media events.

Social media and other digital data is of great and increasing value for scholars of all fields interested in public reception of their work. While the methodology outlined here certainly requires further refinement, the initial results provide some insight into the public attitudes in question and offer some broad ideas into potential courses of action for engaging with these attitudes.

Conclusion

The two clearest contrasts between the data streams generated by professional archaeologists and the data streams generated by the public at large is the difference in regional focus and the difference in emphasis on very specific topics and organizations versus more general topics related to archaeology. Although the methodology developed here will require refinement and further study will need to be conducted in order to clarify the relevance of certain trends as well as to incorporate a wider variety of data input that more adequately represents the concerns of archaeological practitioners worldwide, the broader trends revealed by this preliminary study seem to show that the greatest areas where professional archaeologists will need to focus their public outreach efforts if they are to bring public perception of archaeology into better alignment with actual archaeological practice is on making the specifics of archaeology more accessible to the general public, and work on emphasizing the contemporary political salience of their work. It is clear that certain interested amateur groups have already achieved a good grasp of many of the basic and less-romanticized ideas about archaeology, but have not yet been able to parlay that into an interest or understanding of the more sophisticated and esoteric aspects of archaeological research: they emphasize travel and excavation over research results and challenges. By placing a heavier focus on drawing connections between the basics and the specifics, archaeologists can help guide interested amateurs to a better understanding and deeper appreciation of the academic and intellectual components of their field that they themselves find valuable. Likewise the data from the general public stream shows at least a rudimentary understanding of

topics associated with archaeology—as long as that archaeology is demonstrably relevant to contemporary topics valuable in the context of the lives of individual participants. By focusing on the connections between past and present, heritage, and political relevance, archaeologists may be better able to capture the interest of the general public and hopefully create a larger opening for education, outreach, and the improvement of public relations.

Acknowledgements

The research for this paper was financially supported by a grant from the University of California, Los Angeles (UCLA) Graduate Division. The author would also like to acknowledge the generous assistance of Justin Palumbo (Facebook) and David Shepard (UCLA), whose technical input proved invaluable.

Further Reading

- For additional information about Facebook's proprietary API refer to: https://developers.facebook.com/docs/graph-api/.
- For more information about the structure of Facebook's content fields refer to: https://developers.facebook.com/docs/graph-api/reference.
- TF-IDF Analyzer Tagul can be accessed at http://tagul.com.
- Twitter can be accessed at http://twitter.com.
- Reddit can be accessed at http://reddit.com.

References

AIA "About the AIA". Archaeological Institute of America. Retrieved 3-18-2013. http://www.archaeological.org/about>

ARCE "Mission and History". The American Research Center in Egypt. Retrieved 3-18-2013. <http://arce.org/main/about/historyandmission>

DAY, D.H. (1997). *A Treasure Hard to Attain: Images of Archaeology in Popular Film, with a Filmography*. Londo: Scarecrow Press.

FACEBOOK (2014a). "The Graph API". Retrieved 1-15-2014. <http://developers.facebook.com/docs/graph-api/>

FACEBOOK (2014b). "Statistics". Key Facts. Retrieved 2-1-2014. <http://newsroom.fb.com/Key-Facts>

FAGAN, B.M. (1996). "Portrayal of Archaeology in Popular Culture: Overview". *The Oxford Companion to Archaeology*, edited by B. Fagan. Oxford, Oxford University Press, 574.

FINN, C. (2001) "Mixed Messages: Archaeology and the Media". *Public Archaeology*, 1 (4): 261-268.

GERO, J. and Root, D. (1990). "Public Presentations and Private Concerns: Archaeology in the Pages of *National Geographic*", in: *The Politics of the Past*, edited by P. Gathercole and D. Lowenthal. London: Routledge, 19-37.

KUHN, R. (2002). "Archaeology Under a Microscope: CRM and the Press". *American Antiquity*, 67 (2): 195-212.

LIMP, W. F. (2012). Society for American Archaeology letter to John Fahey. SAA Press Releases. Retrieved 2-13-2013. <http://saa.org/Portals/0/SAA/Press/Diggers.pdf>

LOWENKOPF, S. (1996). "Archaeology in Fiction". *The Oxford Companion to Archaeology*, edited by B. Fagan. Oxford, Oxford University Press: 575-576.

MCGEOUGH, K. (2006). “Heroes, Mummies, and Treasure: Near Eastern Archaeology in the Movies”. *Near Eastern Archaeology*, 69: 174-185.

MCMANUS, P.M. ed. (1996) *Archaeological Displays and the Public*. London: Institute of Archaeology UCL.

MOSER, S. (2001). “Archaeological Representation: The Visual Conventions for Constructing Knowledge about the Past”, in: *Archaeological Theory Today*, edited by I. Hodder. Cambridge: Polity Press, 272-283.

MOSER, S. (2009). “Archaeological Representation: the Consumption and Creation of the Past”. *The Oxford Handbook of Archaeology*, edited by C. Gosden, B. Cunliffe, et al. Oxford, Oxford University Press.

MULLINGS, L. (2012). Letters Opposing Reality TV Shows on Unethical Archaeological Dig Practices. Retrieved 2-13-2013. <http://www.aaanet.org/issues/policy-advocacy/>

NEH “About NEH”. National Endowment for the Humanities. Retrieved 3-18-2013. <http://www.neh.gov/about>

NSF “NSF at a Glance”. National Science Foundation. Retrieved 3-18-2013. <http://nsf.gov/about/glance.jsp>

POHL, J. (1996) “Archaeology in Film and Television”. *The Oxford Companion to Archaeology*, edited by B. Fagan. Oxford, Oxford University Press: 574-575.

SAA “Welcome to the Society for American Archaeology”. Society for American Archaeology. Retrieved 3-18-2013. <http://saa.org/AbouttheSociety/tabid/54/Default.aspx>

STONE, P.G. and Molyneaux, B.L., eds. (1994). *The Presented Past: Heritage, Museums and Education*. London: Routledge.

TAYLOR, F. (2001). “Not Entirely What it Could Be: Historical Perspectives on Modern Archaeology TV Programmes”. *Antiquity*, 75: 468-478.

WYCKOFF, K. (2012). Stop Spike TV from Looting our Collective Past! Change.org Retrieved 2-14-2013. <http://www.change.org/petitions/stop-spike-tv-from-looting-our-collective-past>

ZORPIDU, S. (2004). “The Public Image of the Female Archaeologist: The Case of Lara Croft”, in: *The Interplay of Past and Present*, edited by H. Bolin. Huddinge: Södertörns högskola, 101-107.

Imprint:

Proceedings of the 18th International Conference on Cultural Heritage and New Technologies 2013 (CHNT 18, 2013)

Vienna 2014

http://www.chnt.at/proceedings-chnt-18/

ISBN 978-3-200-03676-5

Editor/Publisher**:** Museen der Stadt Wien – Stadtarchäologie

Editorial Team: Wolfgang Börner, Susanne Uhlirz