

# ARCHAEOLOGY 2.0

new approaches to communication & collaboration



edited by Eric C. Kansa, Sarah Whitcher Kansa, & Ethan Watrall



ARCHAEOLOGY 2.0:  
NEW APPROACHES TO  
COMMUNICATION AND COLLABORATION

## *Cotsen Digital Archaeology Series*

Volume 1. *Archaeology 2.0: New Approaches to Communication and Collaboration*, Eric C. Kansa, Sarah Witcher Kansa, and Ethan Watrall (Editors)

ARCHAEOLOGY 2.0:  
NEW APPROACHES TO  
COMMUNICATION AND COLLABORATION

EDITED BY  
ERIC C. KANSA,  
SARAH WHITCHER KANSA, AND  
ETHAN WATRALL

COTSEN DIGITAL ARCHAEOLOGY 1

THE COTSEN INSTITUTE OF ARCHAEOLOGY PRESS is the publishing unit of the Cotsen Institute of Archaeology at UCLA. The Cotsen Institute is a premier research organization dedicated to the creation, dissemination, and conservation of archaeological knowledge and heritage. It is home to both the Interdepartmental Archaeology Graduate Program and the UCLA/Getty Master's Program in the Conservation of Archaeological and Ethnographic Materials. The Cotsen Institute provides a forum for innovative faculty research, graduate education, and public programs at UCLA in an effort to positively impact the academic, local and global communities. Established in 1973, the Cotsen Institute is at the forefront of archaeological research, education, conservation and publication and is an active contributor to interdisciplinary research at UCLA.

The Cotsen Institute Press specializes in producing high-quality academic volumes in several different series, including Monographs, World Heritage and Monuments, Cotsen Advanced Seminars, and Ideas, Debates and Perspectives. The Press is committed to making the fruits of archaeological research accessible to professionals, scholars, students, and the general public. We are able to do this through the generosity of Lloyd E. Cotsen, longtime Institute volunteer and benefactor, who has provided an endowment that allows us to subsidize our publishing program and produce superb volumes at an affordable price. Publishing in nine different series, our award-winning archaeological publications receive critical acclaim in both the academic and popular communities.

#### THE COTSEN INSTITUTE OF ARCHAEOLOGY AT UCLA

Charles Stanish, Director

Gregory Areshian, Assistant Director

Willeke Wendrich, Editorial Director

Julie Nemer, Publications Manager

#### EDITORIAL ADVISORY BOARD OF THE COTSEN INSTITUTE OF ARCHAEOLOGY

Willeke Wendrich

Christopher Donnan

Jeanne E. Arnold

Jeffrey P. Brantingham

Aaron Burke

Lothar Von Falkenhausen

Sarah Morris

John Papadopoulos

Ex-Officio Members:

External Members:

Area Editor for Egypt, North, and East Africa

Area Editor for South and Central America

Area Editor for North America

Area Editor for the Paleolithic and Environmental Archaeology

Area Editor for Southwestern Asia

Area Editor for East and South Asia and Archaeological Theory

Area Editor for the Classical World

Area Editor for the Mediterranean Region

Charles Stanish, Gregory E. Areshian, and Julie Nemer

Kusimba Chapurukha, Joyce Marcus, A. Colin Renfrew, and John Yellen

Cover design by Ethan Watrall. Editorial and production services: Leyba Associates, Santa Fe, NM

#### Library of Congress Cataloging-in-Publication Data

Archaeology 2.0 : new approaches to communication and collaboration / edited by Eric C. Kansa, Sarah Witcher Kansa, and Ethan Watrall.

p. cm. — (Cotsen digital archaeology ; 1)

Outcome of a session held at the 2008 meeting of the Society for American Archaeology (SAA) in Vancouver, British Columbia. Includes bibliographical references and index.

ISBN-13: 978-1-931745-85-7 (pbk.) ISBN-10: 1-931745-85-4 (pbk.)

1. Archaeology—Computer network resources. 2. Web 2.0. 3. Archaeology—Data processing. 4. Archaeology—Methodology. 5. Archaeology—Information technology. 6. Digital electronics. 7. Internet. I. Kansa, Eric Christopher. II. Kansa, Sarah Witcher. III. Watrall, Ethan. IV. Society for American Archaeology. Meeting (73rd : 2008 : Vancouver, B.C.)

CC80.4.A66 2011

930.1028—dc23

2011030384

© Copyright 2011. Regents of the University of California.

This volume carries a Creative Commons BY-SA (By Attribution, Share Alike, <http://creativecommons.org/licenses/by-sa/3.0/>) license. In short, this means that others can freely distribute, remix, and build upon the contents of this volume, provided two very important conditions are met: the original author receives proper attribution (especially citation) and all subsequent works carry the same license. We chose a Creative Commons license primarily because of our deep concerns in the sustainability of sharply escalating costs in scholarly publishing. These costs make it increasingly difficult for educational institutions, our colleagues in commercial archaeology, students, and members of the interested public to (legally) obtain peer-review publications. Please note that the Creative Commons BY-SA license allows for commercial use, as well as free distribution both inside and outside of the Academy. Permissions for commercial reuse does not, however, mean commercial appropriation. The "copyleft" philosophy embodied by this license enables this work to move in many contexts, but any adaptation or enhancement of this work must be shared back, openly, with the community. Finally, because this license requires proper attribution in any subsequent duplication or adaptation, we hope this volume helps build exposure and recognition for our contributions, and that our colleagues follow in this example. With enough accessible and open data ("data" that includes content like this book), we open up more opportunities for text-mining, tagging, aggregating, linking, visualizing, and hopefully better understanding.



# CONTENTS

List of Figures	vii
List of Tables	viii
Volume Editors	ix
List of Contributors	xi
Preface	xiii
Acknowledgments	xiv
INTRODUCTION: New Directions for the Digital Past	
<i>Eric C. Kansa</i>	1
SECTION I: A Web of Archaeological Data: Infrastructure, Services, and Interoperability	
27	
CHAPTER 1: The Archaeology Data Service and the Archaeotools Project: Faceted Classification and Natural Language Processing <i>Julian Richards, Stuart Jeffrey, Stewart Waller, Fabio Ciravegna,     Sam Chapman, and Ziqi Zhang</i>	
31	
CHAPTER 2: Toward a Do-It-Yourself Cyberinfrastructure: Open Data, Incentives, and Reducing Costs and Complexities of Data Sharing <i>Eric C. Kansa and Sarah Whitcher Kansa</i>	
57	
SECTION II: The Technical and Theoretical Context of Archaeology on the Web	
93	
CHAPTER 3: Poor Relatives or Favorite Uncles? Cyberinfrastructure and Web 2.0: A Critical Comparison for Archaeological Research <i>Stuart Dunn</i>	
95	
CHAPTER 4: Archaeological Knowledge Production and Dissemination in the Digital Age <i>Robin Boast and Peter Biehl</i>	
119	
SECTION III: Archaeological Data Management and Collaboration	
157	
CHAPTER 5: Creating a Virtual Research Environment for Archaeology <i>Michael Rains</i>	
159	
CHAPTER 6: iAKS: A Web 2.0 Archaeological Knowledge Management System <i>Ethan Watrall</i>	
171	

CHAPTER 7: User-Generated Content in Zooarchaeology: Exploring the “Middle Space” of Scholarly Communication <i>Sarah Witcher Kansa and Francis Deblauwe</i> . . . . .	185
SECTION IV: Sustainability, Quality, and Access . . . . .	207
CHAPTER 8: UCLA Encyclopedia of Egyptology, Archaeological Data, and Web 2.0 <i>Willeke Wendrich</i> . . . . .	211
CHAPTER 9: Open Access for Archaeological Literature: A Manager’s Perspective <i>Jingfeng Xia</i> . . . . .	233
CHAPTER 10: What Are Our Critical Data-Preservation Needs? <i>Harrison Eiteljorg</i> . . . . .	251
CONCLUSION: <i>Web 2.0 and Beyond</i> , or On the Web, Nobody Knows You’re an Archaeologist <i>W. Fredrick Limp</i> . . . . .	265



# LIST OF FIGURES

Figure 1.1. A screen-shot of the prototype faceted classification browsing tree for the Archaeotools project .....	40
Figure 1.2. Archaeotools process architecture .....	43
Figure 1.3. Example annotations from a gray literature report showing title, author, and published date .....	46
Figure 1.4. A screen-shot showing a “tagstract” of indexing terms for a gray literature report .....	49
Figure 2.1. Open Context’s “faceted browse” tool, showing filtered results ..	62
Figure 2.2. Example of a project overview in Open Context .....	63
Figure 2.3. Overview of search strings that linked to Open Context content over a three-month period .....	66
Figure 2.4. Open Context uses Dublin Core metadata to generate bibliographic citation information and a stable URL for each item (or tagged set of items) .....	73
Figure 2.5. Illustrations of RESTful services in Open Context .....	78
Figure 2.6. Web services demonstration using Google Earth to visualize Open Context KML data, showing species ratios at different sites in the Near East .....	80
Figure 3.1. Thumbnail images on the Google Maps interface form a clear linear representation of Hadrian’s Wall .....	110
Figures 4.1–4.4. Multimedia applications in photo, video and excavation documentation and digital reconstructions and visualizations .....	130–133
Figures 4.5–4.6. Webcam and videostreaming and its practical application on site .....	135–136
Figure 4.7. Screen-shot of the experimental Blobjects interface (note the tag cloud and recent comments displayed on the right side) ..	139
Figure 4.8. Screen-shot of the control Blobjects interface (note the lack of the tag cloud and comments) .....	140
Figure 4.9. Comparison of Zuni comments (left) and MAA catalog records (right) for four sample objects, grouped according to metadata field or general category .....	146
Figure 5.1. A simple diagram showing stratigraphic units (contexts) and the stratigraphic links between them .....	162
Figure 5.2. Structure diagram with added image, plan, and document .....	162
Figure 5.3. Researching a coin assemblage within a virtual research domain (VRD) .....	163

Figure 5.4. Online collaborative document editing within a virtual research domain .....	164
Figure 6.1. The iAKS One Client/One Server model .....	176
Figure 6.2. The iAKS Many Clients/One Server model .....	177
Figure 6.3. The iAKS Many Clients/Many Projects/One Server model ...	178
Figure 7.1. Custom feeds of zooarchaeology-related data flowing from Open Context into BoneCommons .....	190
Figure 7.2. Diagram showing the flow of information between the ZOOARCH email list and BoneCommons. ....	191
Figure 7.3. Screen-shot of the Zooarchaeology Social Network .....	193
Figure 7.4. A search in Google Insights for Search shows that archaeology-related domains make up a smaller proportion of the Web over time ...	196
Figure 7.5. A Yippee (formerly Clusty) search for domain names shows an increase in .com at the expense of .edu .....	197
Figure 7.6. A Google Blogs search for “archaeology,” showing the relative strength of different domains .....	197
Figure 7.7. The “long tail” of 1,132 searches leading to BoneCommons (searches from April to June, 2010) .....	202
Figure 8.1. The interface of the UEE Open Version in eScholarship ( <a href="http://escholarship.org/uc/nelc_uee">http://escholarship.org/uc/nelc_uee</a> ) .....	216
Figure 8.2. The interface of the Uee Full Version .....	218
Figure 8.3. The UEE time map, overview .....	220
Figure 8.4. The UEE time map, zoomed in to the site of Tanis in the Eastern Nile Delta .....	223
Figure 9.1. A sample web interface from arXiv .....	237

## LIST OF TABLES

Table 1.1. Ontology mismatches between record entries and the appropriate thesauri after initial analysis .....	41
Table 1.2. “Actual” identifications for 906 gray literature reports .....	48
Table 1.3. “Reference” identifications for 906 gray literature reports .....	48
Table 1.4. “Actual” identifications for 3991 PSAS papers .....	51
Table 1.5. “Reference” identifications for 3991 PSAS papers .....	51
Table 2.1. Links to examples from Open Context discussed in the text ....	60
Table 8.1. Feature types for the UEE time map .....	221

## VOLUME EDITORS

### Eric Kansa

Eric Kansa leads development of Open Context (<http://opencontext.org>). His research interests explore Web architecture, service design, and how these issues relate to the social and professional context of the digital humanities. Eric also researches policy issues relating to intellectual property, including text-mining and cultural property concerns. He actively participates in a number of Open Science, Open Government, cyberinfrastructure, text-mining, and scholarly user needs initiatives. Over the past three years, he has taught and practiced project management and information service design in the UC Berkeley School of Information's Clinic program, and his ongoing work at UC Berkeley includes participation on Project Bamboo, a humanities and social science cyberinfrastructure initiative. Eric has been a principal investigator and co-investigator on projects funded by the William and Flora Hewlett Foundation, the NEH and IMLS, Hewlett-Packard, Sunlight Foundation, Google, and the Alfred P. Sloan Foundation.

### Sarah Whitcher Kansa

Sarah Whitcher Kansa is Executive Director of the Alexandria Archive Institute (<http://alexandriaarchive.org>), where she advocates for data sharing and publication in various archaeological and cultural heritage communities. In addition, she is Chief Editor of Open Context (<http://opencontext.org>), an open access data publication initiative for archaeology and related fields. Sarah has a Ph.D. in archaeology and has spent the past 15 years conducting zooarchaeological research at sites in the Near East. She also manages BoneCommons.org on behalf of the International Council for Archaeozoology (ICAZ), where she currently stands on the Executive and International Committees.

## Ethan Watrall

Ethan Watrall is an Assistant Professor in the Department of Anthropology and Associate Director of Matrix: The Center for Humane Arts, Letters & Social Sciences Online ([www.matrix.msu.edu](http://www.matrix.msu.edu)) at Michigan State University. In addition, Ethan is Director of the Cultural Heritage Informatics Initiative and the Cultural Heritage Informatics Fieldschool at Michigan State University ([chi.matrix.msu.edu](http://chi.matrix.msu.edu)). Ethan's research interests fall in the domain of cultural heritage informatics, with particular (though hardly exclusive) focus on digital archaeology and serious games for cultural heritage learning, outreach, and engagement.

## LIST OF CONTRIBUTORS\*

\* Email addresses are provided for lead authors only

Peter Biehl  
380 MFAC Ellicott Complex  
SUNY Buffalo  
Buffalo, NY 14261

Robin Boast  
Museum of Archaeology and  
Anthropology  
University of Cambridge  
Downing Street  
Cambridge  
CB2 3DZ, UK  
[robin.boast@maa.cam.ac.uk](mailto:robin.boast@maa.cam.ac.uk)

Sam Chapman  
Knowledge Now Limited  
Aizlewood's Mill  
Nursery Street  
Sheffield  
S3 8GG, UK

Fabio Ciravegna  
Web Intelligence Technologies Lab  
Organisations, Information and  
Knowledge (OAK) Group  
Department of Computer Science  
University of Sheffield  
S1 4DP, UK

Francis Deblauwe  
The Alexandria Archive Institute  
125 El Verano Way  
San Francisco, CA 94127

Stuart Dunn  
Centre for e-Research,  
King's College London  
26-29 Drury Lane  
London, WC2B 5RL, UK  
[stuart.dunn@kcl.ac.uk](mailto:stuart.dunn@kcl.ac.uk)

Harrison Eiteljorg  
P. O. Box 60  
Bryn Mawr, PA 19010  
[nickeiteljorg@mac.com](mailto:nickeiteljorg@mac.com)

Stuart Jeffrey  
Archaeology Data Service  
Department of Archaeology  
The King's Manor  
University of York  
Y01 7EP, UK

Eric Kansa  
School of Information  
University of California, Berkeley  
211 South Hall  
Berkeley, CA 94720-4600  
[ekansa@ischool.berkeley.edu](mailto:ekansa@ischool.berkeley.edu)

Sarah Witcher Kansa  
The Alexandria Archive Institute  
125 El Verano Way  
San Francisco, CA 94127  
[skansa@alexandriaarchive.org](mailto:skansa@alexandriaarchive.org)

Fredrick Limp  
University of Arkansas  
304 JBHT  
Fayetteville, AR 72701  
[fred@cast.uark.edu](mailto:fred@cast.uark.edu)

Michael Rains  
York Archaeological Trust  
[admin@yorkarchaeology.co.uk](mailto:admin@yorkarchaeology.co.uk)

Julian Richards  
Archaeology Data Service  
Department of Archaeology  
The King's Manor  
University of York  
Y01 7EP, UK  
[jdr1@york.ac.uk](mailto:jdr1@york.ac.uk)

Stewart Waller  
Archaeology Data Service  
Department of Archaeology  
The King's Manor  
University of York  
Y01 7EP, UK

Ethan Watrall  
Department of Anthropology  
MATRIX: The Center for the Hu-  
mane Arts, Letters, and Social Sci-  
ences Online  
418 Natural Sciences  
Michigan State University  
East Lansing, MI 48824  
[watrall@msu.edu](mailto:watrall@msu.edu)

Willeke Wendrich  
UCLA, Cotsen Institute of  
Archaeology  
308 Charles E Young Dr. North  
A210 Fowler Building/Box 951510  
Los Angeles, CA 90095-1510  
[wendrich@humnet.ucla.edu](mailto:wendrich@humnet.ucla.edu)

Jingfeng Xia  
755 W. Michigan Street, UL 3100N  
SLIS - IUPUI  
Indianapolis, IN 46202  
[xiaji@iupui.edu](mailto:xiaji@iupui.edu)

Ziqi Zhang  
Web Intelligence Technologies Lab  
Organisations, Information and  
Knowledge (OAK) Group  
Department of Computer Science  
University of Sheffield  
S1 4DP, UK

## PREFACE

This volume is the outcome of a session held at the 2008 meeting of the Society for American Archaeology (SAA) in Vancouver, British Columbia. The session carried the title “Web 2.0 and Beyond: New Tools for Archaeological Collaboration and Communication” and was organized by the editors. The majority of the chapters in this volume are based on presentations given in that session. However, in order to keep up with the sometimes fast pace of technological change, all contributors made significant updates to their chapters in mid-2010 to make the information as current as possible. In addition to the conference papers, the editors solicited two additional chapters (Chapters 7 and 10). These two chapters bring perspectives that we feel offer a more complete picture of the current uses and challenges of new Web-based technologies in archaeological communication.

This volume is somewhat experimental, being the first volume published by the Cotsen Institute of Archaeology Press in their new Cotsen Digital Archaeology (CDA) series, which uses the University of California’s eScholarship framework for digital publishing. Through eScholarship, the contents of this volume are available online for free, though to accommodate the expectations of traditional print publication, print-on-demand copies are also available for a fee. Publication of this volume, open access, in a primary electronic medium, seems well suited for its subject matter. It will be interesting to see if open access venues become more accepted and expected as scholarly communications continue to evolve.

Finally, readers should be aware that this volume does not represent a comprehensive overview of digital archaeology. Because this volume originated from a Society for American Archaeology conference, it does not explore digital initiatives outside of North America or the United Kingdom. Thus, impressive achievements of colleagues in Europe, Asia, and elsewhere are not represented. This book also does not cover the equally impressive work of museum researchers, classicists, and scholars of the ancient Near East and Pacific. While this volume is not comprehensive, the sampling of perspectives presented here can nevertheless inform more general debates over issues of accessibility, sustainability, information quality, and semantics that cross-cut disciplinary and geographic boundaries.

## ACKNOWLEDGMENTS

First and foremost, we would like to thank the contributors to this volume for sharing their valuable perspectives. The William and Flora Hewlett Foundation sponsored the session at the 2008 Society for American Archaeology (Vancouver, BC) where the contributions to this volume were first presented. A grant from the National Endowment for the Humanities and the Institute of Museum and Library Services (PK-50072-08) provided a portion of the support needed for the organization and production of this volume. Finally, we are grateful to Carol Leyba for her dedicated editorial work, Willeke Wendrich for her encouragement to publish in the Cotsen Digital Archaeology series, and two anonymous reviewers for their constructive comments on the original version of this volume.



## INTRODUCTION

# NEW DIRECTIONS FOR THE DIGITAL PAST

*Eric C. Kansa*

This volume looks at archaeology in the context of the World Wide Web, a communication system that has witnessed over two decades (and counting) of exponential growth. In many ways, the Web represents a revolution in communications and information sharing that rivals in significance the invention of the printing press or the origins of writing. In the past two decades, the Web has come to permeate virtually every aspect of our lives, transforming journalism, the arts, commerce, and the way we socialize.

## WHO SHOULD READ THIS BOOK AND WHY

As long as the centuries continue to unfold, the number of books will grow continually, and one can predict that a time will come when it will be almost as difficult to learn anything from books as from the direct study of the whole universe. It will be almost as convenient to search for some bit of truth concealed in nature as it will be to find it hidden away in an immense multitude of bound volumes.

—Denis Diderot, “Encyclopédie” (1755), quoted in Wikipedia

As part of this transformation, the Web’s growth means that we live in an information-saturated era. This volume adds minutely to this ever expanding body of data. While the Web increasingly enriches our work and social lives with new information, we face often overwhelming demands on our attention. Not only do we have new sources of news, information, and entertainment to jostle for our attention, we have many new sources of disinformation, spam, propaganda, and plain junk. It is no wonder that many scholars, overloaded with information, lament the passing of bygone days of quiet scholarly contemplation (Harley et al. 2010). No doubt many of the

complaints about today's Web reflect some romanticism about the past. Even with the development of telegraph networks in the nineteenth century, people complained about information overload (Standage 1998:165). While the phenomenon of information overload may not be a new symptom of modernity, even some pioneers of cyberspace worry about the current data deluge and its impact on creativity and deep thinking (Lanier 2010).

Given all these competing demands on our limited attention, this book about archaeology on the Web requires some justification. Instead of merely feeding information overload, we hope this volume will help archaeologists take stock and better understand how the Web is transforming the professional practice of archaeology, just as it transforms professional communications in other disciplines. Reflection on these changes can help us better understand the state of this discipline. Moreover, researchers who study scholarly communications more generally will find this book a useful source of case studies. Archaeology is an inherently multidisciplinary enterprise, with one foot in the humanities and interpretive social sciences and another in the natural sciences. As such, case studies in digital archaeology can help illuminate changing patterns in scholarly communications across a wide array of disciplinary contexts.

Archaeologists will find this book a useful guide in understanding how this revolution in communications technology reverberates across this discipline. Many of the contributions describe technologies, user interface designs, and organizational practices that attempt to mitigate some of the problems associated with the Web, especially information overload and disinformation. Contributions to this book also explore how the Web can be used to transform archaeological communications into forms that are more open, inclusive, and participatory. Some discussions focus on ways that Web-based systems can make archaeological knowledge production more open and transparent, while others focus on the challenges of archiving and preserving digital data. Finally, some chapters describe case study examples of digital projects. Sharing these experiences can provide useful guidance for other researchers wanting to create and apply technology to archaeology.

## LOOKING BACK AT WEB 2.0

The book is loosely themed on so-called Web 2.0 approaches to these issues. Most contributions presented here derive from presentations given at the

2008 Society for American Archaeology conference in Vancouver, British Columbia. However, as is typical of many scholarly publishing cycles, transitioning conference presentations to a book form has taken about three years. Thus, many of the new technologies and perspectives presented at the 2008 conference are no longer so new. Nevertheless, the slower pace of academic publishing does offer some advantages. The lengthier review and revision cycle gave many of our contributors some added perspective and a chance for more nuanced reflection. Thus, many discussions of Web 2.0 presented in this volume have an element of retrospection.

Originally coined in 2004 by Tim O'Reilly, founder and CEO of O'Reilly Media, the term "Web 2.0" describes social networking systems, blogs, and other Web-based platforms emphasizing collaboration and sharing, rather than the unidirectional flow of information in a traditional Web 1.0 architecture (O'Reilly 2005). Many archaeologists have embraced Web 2.0 tools and technologies, allowing them to integrate different bodies of content and develop new tools and interfaces for peer-to-peer communication and collaboration. Ultimately, these new tools and platforms allow opportunities for research and public participation in archaeology.

However, "Web 2.0" is fast becoming a clichéd and obsolete expression. It stems from the revival of investment and interest in the Web following the dot-com collapse at the turn of the century. Now, in the aftermath of another and far graver financial collapse, many of the startups branded as Web 2.0 will likely fail. The term "Web 2.0" soon may be consigned to history.

## THE PROMISE OF WEB 2.0

As archaeologists accustomed to dealing with "deep time," it makes sense to consider the Web's impact on the discipline with a longer time horizon than is typical of most discussions of Web 2.0. While much about Web 2.0 will be a passing trend, the term still points to perspectives and developments likely to have lasting value and significance. In general, the term evokes designs and services emphasizing user interaction and the user as a source of extra value. This value can take multiple forms, including (but not limited to):

- *User-generated content:* Many Web 2.0 systems provide users with platforms for sharing and publishing content. These range from images to videos, and from essays to short, 140-character messages in microblogging services like Twitter.

- ▶ *Crowd-sourced classification:* Many Web 2.0 systems provide mechanisms for users to create and share metadata (“information about information”) that can describe content to facilitate search and retrieval. Tagging and folksonomy (informal classification) systems provide important information retrieval services.
- ▶ *Remixable data:* Providing Web-based data in ways that can be easily manipulated by software represents another common theme of Web 2.0. Web services, often called “application program interfaces” (APIs), publish data on the Web in formats intended for use by third-party software. The intent behind an API is often to “crowd-source” interesting and useful software applications based on content. *Google Maps*<sup>1</sup> is currently an excellent example. Google provides mapping data and tools to third-party web developers. These developers install Google Maps on their sites, and in the process, Google (and its brand) becomes an increasingly ubiquitous feature of the Web.
- ▶ *Enhancing and evaluating information quality:* Web 2.0 often has connotations of an anarchic free-for-all, lacking traditional gate-keeping mechanisms to maintain quality. While information quality concerns rightly make researchers skeptical of some Web 2.0 platforms, in other contexts Web 2.0 systems try to promote quality. Sometimes Web 2.0 evaluations of quality are little more than popularity contests. In other cases, experimental scientific journals, such as *PlosOne*,<sup>2</sup> attempt to use Web 2.0-style rating and commenting systems as a new form of “enhanced peer-review.”

Leveraging user communities and user interactions in the ways described above will probably continue to feature in future technology developments, in both popular and scholarly media.

As is evident to many of us, the speed of implementation, rate of adoption, and impacts of drawing value from the efforts of users can be highly uneven. Developments in commercial web services, technologies, and tools are transforming the professional and personal lives of archaeologists, often by blurring the boundaries between these lives. The founders of *Google*,<sup>3</sup> one

---

<sup>1</sup> <http://maps.google.com/>

<sup>2</sup> <http://plosone.org>

<sup>3</sup> <http://www.google.com/>

of the first and most successful Web 2.0 companies, discovered a powerful method for finding value in the distributed work of millions, with the Page-Rank algorithm (Brin and Page 1998). Google searches, as well as more specialized services such as [Google Scholar](#),<sup>4</sup> have transformed the way researchers find information (Markey 2007; Yu and Young 2004). Google has also changed the way university libraries serve their patrons, influencing them to adopt single-box “full-text” search and stronger ranking algorithms.

Web 2.0 increases the diversity of content that people can find readily on the Web. Social networking sites, blogs, shared web bookmarks ([Delicious.com](#)<sup>5</sup> and the like), ratings systems ([Digg](#)<sup>6</sup>), and shared media ([Flickr](#)<sup>7</sup> and [YouTube](#)<sup>8</sup>) have expanded the range of media that people publish online. However, trends impact scholarly communications in slower and more tentative ways. While the peer-reviewed, journal-published paper is still the main currency of professional research communication, an expanding number of research-themed blogs make less formally published material available. More archaeologists now publish images to Flickr, share presentations on [SlideShare](#),<sup>9</sup> and participate in open access publication (usually through self-archiving).

While Web 2.0's impact is far reaching, it does seem to have limits. Web 2.0 platforms and services mainly facilitate informal communications among archaeologists. Web 2.0 systems are simple to use, fast, and geared to content that requires relatively minimal investment to create. Archaeologists tend not to use Web 2.0 platforms as the primary dissemination channel for forms of content that take a great deal of effort and expertise to create. In this light, data sets and sophisticated scholarly manuscripts see less circulation in Web 2.0 channels.

## FINDING WEB 2.0 SOLUTIONS FOR PRIMARY DATA?

If related data and documents can be linked together in a scholarly information infrastructure, creative new forms of data- and information-

---

<sup>4</sup> <http://scholar.google.com/schhp?hl=en&tab=ws>

<sup>5</sup> <http://www.delicious.com/>

<sup>6</sup> <http://digg.com/>

<sup>7</sup> <http://www.flickr.com/>

<sup>8</sup> <http://www.youtube.com/>

<sup>9</sup> <http://www.slideshare.net/>

intensive, distributed, collaborative, multidisciplinary research and learning become possible. Data are outputs of research, inputs to scholarly publications, and inputs to subsequent research and learning. Thus they are the foundation of scholarship (Borgman 2007: 115).

Borgman describes a desired goal for so-called cyberinfrastructure systems that support research through more efficient communication and data preservation. Unfortunately, most popular Web 2.0 tools and services cannot deal with the complexities required of such a system. Flickr, [Google Docs](#),<sup>10</sup> and other applications make certain aspects of archaeological fieldwork convenient to publish online, but they cannot describe key contextual information in a precise, consistent, and machine-readable way. For instance, contextual relationships are difficult to precisely describe in Flickr's annotation (tagging) system, limiting Flickr's usefulness in publishing images from an archaeological excavation. Similarly, [Google Docs](#), [ManyEyes](#),<sup>11</sup> [Swivel.com](#) (now defunct), and other online systems for sharing tabular (structured) data limit dissemination of structured data to data structures that can be represented on single tables. This is sufficient for archaeological data sets like individual zooarchaeological analyses, but it does not work well for publishing data on complex, multidisciplinary archaeological projects involving data sets generated by several different specialists and describing complex contextual relations.

Most Web 2.0 systems are simple to use and place minimal requirements on end users to prepare and describe content. Users generally find it easy to retrieve relevant pictures or videos by searching content indexed by user-generated tags. Contributing (publishing) to a Web 2.0 system can be quite easy, because the content published is generally simple and described with limited and informal metadata, such as user-generated tags. So, Web 2.0 systems sufficiently serve many popular needs and applications. However, they have not been widely used in academic communities for content central to research. To respond to that problem, this volume presents recent advances in archaeological data sharing and explores how Web 2.0 services affect communication and collaboration in archaeology.

---

<sup>10</sup> <http://docs.google.com/>

<sup>11</sup> <http://manyeyes.alphaworks.ibm.com/manyeyes/>

## OVERVIEW OF CONTRIBUTIONS

The chapters in this volume illustrate the possibilities and limitations of the Web in meeting the specialized needs and requirements of professional researchers. Because Web 2.0 is best understood loosely as a zeitgeist or even as a marketing term, contributors define and discuss Web 2.0 from their own perspectives. The chapters address various semantic, intellectual-property, technical, social, and professional challenges of networking archaeological information. They present different perspectives on conceptual, theoretical, and practical approaches to communicating archaeological knowledge with new technologies and platforms. Some have successfully implemented Web 2.0 tools and approaches, while others have rejected such approaches. Issues about information quality, audience, and authority also inform their discussion.

This volume shows how emerging digital forms of archaeological communication differ from traditional paper-based media, and how these differences require examination and rethinking of knowledge production processes. Many example projects in this volume are rich in structured data and multimedia content. Some of this content is generated “in real time” in active field programs and sees little editing or filtering before global dissemination. These projects hope to use the inherent capabilities of Web 2.0 technologies and platforms to make archaeology more collaborative and more transparent. However, they also raise difficult questions about information quality, information overload, intellectual property, and relationships between professional researchers, students, and different public communities.

This book is divided into themed sections to help highlight certain particularly salient points of discussion made by the various contributions. The section themes are summarized as follows:

- ▶ *Section I* focuses on information retrieval and information-access approaches, especially centered on gray literature and primary field data. These forms of content have traditionally seen little dissemination.
- ▶ *Section II* explores larger conceptual concerns regarding information access and management. The contributions in this section discuss practical as well as theoretical concerns inherent in various design choices for archaeology’s computing infrastructure.

- ▶ *Section III* presents projects that aim to enhance collaboration in archaeology through various approaches, such as the adaptation and development of certain technologies for mobile field-based collaboration, coordination and data management of field-based researchers and other specialists, and collaboration among the international researcher community.
- ▶ *Section IV* addresses scholarly communications issues, with a particular emphasis on concerns over information quality and access in light of sustainability and preservation imperatives.

## SECTION I

### Chapter 1

In “The Archaeology Data Service and the Archaeotools Project: Faceted Classification and Natural Language Processing,” Julian Richards and colleagues discuss design innovations in information retrieval and integration of different data services. A key problem for archaeological information sharing is information overload. Standard keyword search systems often retrieve too much irrelevant information or fail to deliver relevant information if keywords are not mapped to synonyms. These problems make keyword searches somewhat unreliable and prone to deliver different results depending on the keyword inputs. However, by applying faceted search mechanisms, researchers gain greater comprehension of an entire corpus of material and can progressively refine searches to obtain specific information relevant to their interests.

Beyond search methodologies, Richards et al. discuss emerging frontiers of archaeological information management. Techniques in natural language processing (NLP) promise to enhance the value of archaeological literature. Automated and semiautomated NLP techniques help address some limitations of Web 2.0 methods in narrow niche professional contexts. How do you crowd-source metadata creation via techniques like social tagging when there is no crowd? The archaeological research community is relatively small, typically has very specialized interests, and may not be especially interested in social collaborative tagging. Thus, NLP offers a viable strategy to generate metadata to improve information retrieval (through faceted search and other techniques) without requiring the action of a (non-existent) “crowd.”



Richards et al.'s discussion of faceted search and NLP techniques for metadata creation is on the cutting edge of archaeological informatics, indicating important features of the landscape of archaeological information retrieval for years to come. As these systems are developed and deployed, they will shape how professionals and members of the public encounter cultural heritage. In other words, our record of the past will be increasingly shaped and organized by algorithms. This trend will be a fascinating topic for future research and critique. Will these algorithms be part of unobserved background processes that rarely see scrutiny? Or will there be discussion and debate about who creates and deploys these algorithms, and for what agenda and purpose? How will different perspectives and agendas be accommodated? Many of the theoretical debates and concerns that shaped the content of archaeological literature may also emerge in the context of automated processes to organize and retrieve that literature.

## Chapter 2

In Chapter 2, “Toward a Do-It-Yourself Cyberinfrastructure: Open Data, Incentives, and Reducing Costs and Complexities of Data Sharing,” Eric Kansa and Sarah Whitcher Kansa discuss how techniques of Web 2.0 systems can be applied for research applications. Simple web services delivering machine-readable data can help make archaeological information open and reusable for research, instruction, and creativity. However, fitting new modes of communication and collaboration into traditional research practices poses potentially insurmountable problems with regard to time, recognition, technical challenges, and workflows. These concerns have guided new developments to **Open Context**,<sup>12</sup> an open source publishing system designed to facilitate sharing, collaboration, and integration of archaeological content.

The Kansas discuss some of the successes and failures of Web 2.0 in Open Context. They discuss how folksonomies provoked some initial curiosity in the system but failed to engage enough users to create useful metadata. In contrast, other aspects of Web 2.0 seem to have greater long-term traction and significance for their project, particularly exposure of machine-readable data through web services. Approaches to design and delivery of

---

<sup>12</sup> <http://opencontext.org/>

machine-readable data through web services represent one legacy of Web 2.0 likely to have long-term impact and application for research data sharing. They explore how web services enable data reuse across different applications and collections. When data are made machine-readable, content can be freed from individual silos and used with content from other sources. These may include other archaeological collections or systems supporting data sharing in other disciplines. Content is also freed from a single mode of presentation and visualization. Web services therefore encourage development of new user-interface paradigms and greater flexibility in user interactions with diverse content. Finally, certain technical design perspectives encourage architectures that better support important scholarly conventions, including citation and linking.

Discussing the reluctance of some researchers to share data, they emphasize a strategy that casts data sharing as a form of publication, where many of the conventions for citation and editorial oversight used in narrative publication can be applied. This perspective has increasing traction across many scientific fields, as indicated by recent editorial comments in the journal *Nature* (“Data’s Shameful Neglect,” 2009). Beyond serving as a useful publication model, narratives also provide context and meaning for archaeological data sharing. Kansa and Whitcher Kansa discuss “tacit knowledge” and the implicit understandings and background required to make sense of archaeological data. They discuss transmission of tacit knowledge via formal classification systems and ontologies, or through social scholarship enabled by Web 2.0 systems. They highlight the need to integrate data publication with narrative and interpretive publication to make shared primary data intelligible and usable by a wider community. This last point is made by many other contributors to this volume.

## SECTION II

### Chapter 3

Historically, archaeologists became interested in computing and databases to control huge quantities of excavation data. They looked to the computer as a tool to retrieve and analyze information across multiple data sets and excavations to create broad syntheses. Unfortunately, the promise of digitally based meta-analysis has not panned out. The mass of digital material generated by archaeological activity is geographically distributed, fuzzy, incom-

plete, inconsistent, and often hard to access. The resulting complexity deluge presents a whole new set of problems for archaeology. Stuart Dunn's contribution in Chapter 3, "Poor Relatives or Favorite Uncles? Cyberinfrastructure and Web 2.0: A Critical Comparison for Archaeological Research" critically reviews Web 2.0 methods and technologies that address this emerging problem. He explores cyberinfrastructure/e-science and how it relates to Web 2.0 technologies and techniques in archaeology. Dunn divides the process of using archaeological data into collection and harvesting; analysis, integration, and interpretation; and social research. In these three domains, he explores various hallmark technologies and methodologies commonly associated with both Web 2.0 and cyberinfrastructure. In particular, he sees folksonomy as a way to supplement and enhance traditional taxonomies. He advocates a "spade to screen" documentation process, to ensure that methods used to author and create digital objects are transparent and attributable. Dunn concludes that the top-down approach of cyberinfrastructure and the bottom-up approach of Web 2.0 are not two irreconcilable models, but different layers in the same structure.

## Chapter 4

In Chapter 4, "Archaeological Knowledge Production and Dissemination in the Digital Age" Robin Boast and Peter Biehl discuss the ways in which different cultural contexts shape information management, retrieval, and use. They explore the diversity of ontologies and classification systems among expert communities and others, including different indigenous communities. Their discussion begins with an exploration of archaeological approaches to knowledge creation, contrasting "classificatory" versus "interpretive" paths. Whatever interpretive approach is taken, tangible cultural heritage becomes embedded in intangible processes that shape understandings of that tangible heritage. They argue that online information systems are contact zones where different understandings collide and inform one another.

Boast and Biehl then discuss how different conceptual systems can inform one another through Web-mediated collaboration. They look at how museums and their educational programs try to bridge understandings between museum experts and various professional communities. However, as they note, museums typically concern themselves only with managing their own, expert-informed classifications and documenting their own collections.

They do little to document how different public communities understand these collections. The perspectives of outsiders, though present and expressed in museum educational performances, rarely end up being recorded or informing experts.

Alternatives to the model of “one-way” broadcasting of museum expert knowledge are emerging. Many of these use Web 2.0 ideas of two-way communication and participation in creating content and sharing ideas. Boast and Biehl’s exploration of classification is of interest to those wishing to foster reciprocal information sharing across different community settings. Essentially, they find that categorizing cultural heritage, even in loosely structured and constrained “folksonomy” approaches, is of limited appeal and interest to many people outside of museum professional circles. They find much more interest in digital representations of material culture to support narratives. In other words, opening up museum collections for social tagging resonates less than encouraging storytelling.

This observation has important implications. Concerns over classification and standards for classification dominate thinking about cyberinfrastructure, the Semantic Web (or “Linked Data”), and cultural heritage data sharing. Many discussions of Web 2.0 and folksonomies emphasize classification issues in information sharing. However, while classification is important, it is not the only concern. In some cultural contexts, construction of narratives has greater priority. By looking at how cultural heritage information is used in different contexts, Boast and Biehl highlight the need to move beyond classification to other social uses of information. Their insights help guide future attempts to bridge gaps between museums and other communities and also highlight the importance of narrative even within academic and museum professional circles.

## SECTION III

### Chapter 5

Archaeological projects are rarely blessed with full-time, permanent staff. In some cases, part-time specialists are employed full time at other museums or institutions, or work as freelance archaeological specialists involved in a wide range of additional projects. In other cases, specialists focus so intently on their particular research interest that they have no real sense of the totality of the project. The result is that specialists often feel isolated or semidetached.

In Chapter 5, “Creating a Virtual Research Environment for Archaeology,” Michael Rains discusses VERA (Virtual Environment for Research in Archaeology), a system attempting to address these issues. Funded by JISC (Joint Information Systems Committee), VERA is a Web-based virtual research environment (VRE) collaboratively developed by the University of Reading, University College London, and York Archaeological Trust. It is centered on the **Silchester Town Life Project**<sup>13</sup> at the University of Reading. This is a large-scale, ongoing excavation of part of the abandoned Roman town of Calleva Atrebatum at Silchester, approximately 80 km west of London. Silchester has used the Integrated Archaeological Database (IADB) as its data management system since the start of the project in 1996. Key aims of the VERA project include improving the flow of information from excavation, through analysis and research, to publication and dissemination, and developing a collaborative working environment involving all members of the project team. At the heart of the VERA interface is one or more interactive graphical representations or visualizations of part of the project database. For example, a particular phase in site development is displayed as a standard archaeological stratigraphy diagram. What is unique about VERA is that additional content from the project database, such as plans of stratigraphic units, photographs, and field notes, can be attached to the diagram. This process adds context to excavation materials and allows all Web-connected stakeholders, regardless of their location, to contribute to the project.

## Chapter 6

Ethan Watrall’s contribution in Chapter 6, “iAKS: A Web 2.0 Archaeological Knowledge Management System,” proposes a system to leverage web services and Web 2.0 technologies. iAKS aims to solve many data collection, storage, and visualization challenges currently faced by archaeologists. Database field tools are difficult to find and are often expensive, complex systems developed on a project-specific basis. iAKS, by contrast, is a flexible design that can be used by many different projects with very different research designs. Most importantly, iAKS can be used in the field and offers different types of service, depending on whether a project has Internet access, a local server, or simply a hard drive. Content created using iAKS is converted into XML,

---

<sup>13</sup> <http://www.silchester.rdg.ac.uk>

making it easy to share, integrate with other content, and preserve. This holistic approach, from field-based design and data collection to data sharing and integration via the Internet, makes the iAKS system particularly promising. Most importantly, the architecture approaches Watrall advocates explore new possibilities in mobile computing, where global information systems and infrastructure can come together with handheld, field-ready devices (phones, tablet computers, etc.). These capabilities can make archaeology increasingly “glocal” (simultaneously global and local), as particular finds and contexts observed locally can be related to other digital documentation found on global information networks.

## Chapter 7

In their contribution, “User-Generated Content in Zooarchaeology: Exploring the ‘Middle Space’ of Scholarly Communication,” Sarah Whitcher Kansa and Francis Deblauwe review the emerging role of user-generated content in archaeological communications. As discussed above, archaeologists mainly participate in Web 2.0 platforms for less formal types of communication and sharing. This chapter, based on the experiences of people who actively manage and participate in Web 2.0 systems, explores why archaeologists sometimes see a valuable role for Web 2.0 channels.

This chapter makes a number of interesting points with regard to incentives for professional researchers to participate in online social media. Blogging, as a social media platform, has been widely used now for several years. While there are several professionally oriented archaeological blogs, this chapter notes that blogging is still a somewhat niche activity in the discipline. Nevertheless, despite the fact that most blog posts will see little comment, the impacts of blogging may be greater than is immediately apparent: they can help spark interest in a paper, a website, or a grant announcement through the passive engagement of readers, and even through improving the search engine exposure of web resources referenced in a blog post.

Similarly, community portals are commonplace on the Web but are less discussed in calls for disciplinary “cyberinfrastructure.” One such portal, **BoneCommons**,<sup>14</sup> has seen various incarnations since its initial launch in 2006 and now enjoys continual use by the zooarchaeological community.

---

<sup>14</sup> <http://alexandriaarchive.org/bonecommons/>

The authors recount their experiences managing scholarly blogs and community portals, and note a rapid change in academic uptake and participation in social media. This last point is particularly important. Expectations and acceptance of technologies are not static, even in academia. In laying the foundation for digital infrastructure, we have to be mindful of trends and trajectories, and not just the current state of the research community in accepting a given technology or dissemination platform.

Finally, Chapter 7 touches on the expanding reach of digital preservation efforts to capture the ephemera of discussions on Twitter and email lists. This raises important questions about the scope and reach of data preservation efforts. When does data preservation go too far, and when does it start to appear invasive? This issue goes beyond social media and can include the primary field-documentation and notes of excavators. Such documentation can be full of irrelevancies that range from bickering to flirtations, and from complaints of confusion to inside jokes (sometimes very off-color). While this content could help contextualize archaeological data, should all of this sometimes embarrassing content go into the official archaeological record? What is the scope of privacy in data preservation?

## SECTION IV

### Chapter 8

Willeke Wendrich's chapter, "UCLA Encyclopedia of Egyptology, Archaeological Data, and Web 2.0," outlines the tensions between the various traditions, incentives, and quality concerns of professional scholarship, on the one hand, and the possibilities and environment of the Web, on the other. There is a widespread and often justified perception among academics that the Web is an unreliable foundation for scholarship. The Web is highly fluid, content can change at any time without notice, and resources may move or disappear entirely. At the same time, the Web has obvious advantages in reducing the cost and difficulty of disseminating scholarship. Wendrich's chapter describes the efforts of the UCLA Encyclopedia of Egyptology to take advantage of the best of aspects of the Web while avoiding the worst.

A clear implication of Wendrich's work relates to the concerns over what constitutes publication. Key attributes involve persistence, peer review, and editorial control. In some ways, these attributes run counter to the emphasis common to Web 2.0 systems: easy retrieval, immediacy, popularity, and

participation. Wendrich highlights the importance of reliable and credible citation. In contrast to the typical contributor to Web 2.0 systems, archaeologists participate in knowledge creation in very different ways, typically resulting in more complex, larger, and discrete works (the chapters in this volume are a good example). As Wendrich points out, giving credit to archaeological researchers as individuals is vital. Many do not want their authorial voice diluted or lost in a collective, as occurs in contributing to, for instance, Wikipedia. Many scholars also consider knowledge creation to be cumulative, where it is important to build upon works and contributions made across many decades. Quality and comprehensiveness are more important to scholars than they are to Web 2.0 users looking for easy dissemination and discussion.

In exploring these issues, Wendrich emphasizes the importance of considering Web-based publication as *publication*. However, she goes beyond a simple model that merely replicates traditional printed matter on the Web. The Encyclopedia of Egyptology's experiments with various forms of digital media beyond text illustrate how Web-based publication can support more depth and diversity in the content of scholarly communication. Nevertheless, while open to experiments with "new media," Wendrich makes a convincing case that digital dissemination must rest on a solid foundation of established scholarly traditions.

## Chapter 9

In Chapter 9, "Open Access for Archaeological Literature: A Manager's Perspective," Jingfeng Xia reviews open access archiving of content from the perspective of an experienced archival manager offering recommendations for the nascent field of archaeological publication archiving. Xia discusses the institutional archive approach and warns that, while it benefits from vast input by hired institutional managers, the content is often broad but shallow and not well informed. That is, people inputting content aim for breadth, while depth and accuracy in metadata suffer because managers may not understand the subject beyond abstracts or keywords. Learning from other disciplines, Xia encourages the archaeological community to adopt a subject repository approach, where the archive pools resources from many organizations and is managed by archaeological subject matter experts. Xia explains that subject repositories tend to offer deeper and more accurate metadata description of content, but may suffer from a lack of institutional infrastructure.



Without an organizational hub, who will manage the content? Who will ensure its longevity? Xia discusses possible next steps for data sharing in the archaeological community.

Xia's focus on the accessibility of archaeological publications has significance beyond impact and quality issues for human readers. The same sorts of text-mining and NLP approaches explored by Richards et al. in Chapter 1 can be applied to more mainstream archaeological publications. However, copyright restrictions, subscriptions, and login barriers now make it too difficult to obtain large corpora of published archaeological literature. Thus Xia's call for an open access repository in archaeology can pave the way for new research opportunities using advanced computational methods.

## Chapter 10

Another consensus among the contributors is that, despite its new possibilities, Web 2.0 by itself will not “crack the archaeological data-sharing nut.” In the penultimate chapter of this volume, “What Are Our Critical Data-Preservation Needs?,” Harrison Eiteljorg offers a “naysayer” position, enumerating the shortcomings of sharing data via a Web 2.0 repository. Eiteljorg distinguishes “data access” via a passive archive, where access involves “frozen” resources such as spreadsheets of data, versus “data organization,” employing Web 2.0 features such as data integration and user contribution. Beyond the specific issues, which range from controlled vocabularies to different file formats, data sharing via contributory systems faces an overarching challenge: how does one ensure that the content user fully understands (1) the project the data come from and (2) the data collection process itself? Furthermore, how can we logically compare “resources that are inherently dissimilar because they are derived from data collected in different ways by different people at different times and with different purposes”? Eiteljorg reviews disincentives to contributing data to a Web 2.0 repository, including the limited professional rewards for doing so and the lack of momentum to archive data once a project has been published (and the “big push” is over). In contrast, many of the other chapters in this volume discuss efforts at eliciting professional rewards for data sharing. However, data sharing is a new concept for most researchers, who are still getting accustomed to the idea of archiving print publications. Eiteljorg recognizes the promise of Web 2.0 and suggests that Web 2.0-style approaches continue to be explored as a

means of data sharing, but in conjunction with the static archiving of data sets.

## CONCLUSION

Fred Limp's concluding chapter, "*Web 2.0 and Beyond*, or On the Web, Nobody Knows You're an Archaeologist," recapitulates concepts discussed in preceding chapters and paints an optimistic picture of the future of archaeological data sharing. In reviewing the contributions to this volume, Limp notes the importance of differentiating between *goals* and *techniques* to accomplish those goals. As technologies rapidly evolve, specific implementations will vary, but strategic needs will be more stable and should guide the professional community's efforts more than fixations on the latest technological fashions. Limp explores strategic concerns affecting the viability of attempts at archaeological data sharing. In comparison with commercial uses of Web 2.0, archaeological data sharing, Limp notes, has some unique requirements. One key element is the need for sustainability. Commercial Web 2.0 initiatives need not bear the burden of maintaining the irreplaceable record of humanity's cultural heritage on volatile media and technology platforms long into the future. Sustained and credible institutional support—such as the support of the California Digital Library, the new organization Digital Antiquity (a welcome new development since 2008 when these chapters first came together), or the Archaeology Data Service—is a requisite for the discipline. In addition, Limp argues that Web 2.0 services often emerged in situations where there was a large amount of valuable content readily available to “prime the pump” and attract sustained interest and use in their platforms. For archaeology, this is more difficult, because of a limited supply of content ready for digital dissemination. Beyond these difficulties, Limp sees challenges in motivating greater data sharing and in agreeing upon technical and semantic standards.

Sustainability and reaching a critical scale for content sharing remain important strategic questions. Limp's point about separating implementation specifics from strategic goals helps to highlight options available to the archaeological community. Development of Web 1.0 (and also Web 2.0) was very much a distributed effort, with many failures and some successes. Similarly, prospects for archaeological data dissemination must be considered more broadly than the successes or failures of any given project. What are the

overall trends, and is there evidence for increasing and more cumulative archaeological data sharing? Given the growing list of initiatives discussed and referenced in these chapters, we suspect that archaeological data sharing and Web engagement, as a distributed phenomenon, will likely continue to grow. While individual projects may come to an end, technical expertise, data, standards, and experience will continue to grow.

Sustainability requires long-term institutional credibility and resources typically found only in organizations like libraries, universities, and government agencies. Ultimately, sustainability may also require sustained public financing to maintain “public goods.” To put this issue in perspective, archaeology as a discipline is manifestly *not sustainable* without continued public financing. Without public support, a sustainable business model for archaeology would probably look much like the antiquities trade! Since the entire enterprise of archaeology cannot be sustained without public support, archaeological knowledge preservation and dissemination will also likely require continued public financing.

Sustainability strategies must also accommodate the reality that archaeological data-sharing efforts are scattered among several diverse initiatives and projects. This experimentation fosters innovation and builds technical capacity and expertise throughout the discipline. It also reduces the danger that there will be one and only one preferred approach to managing and making sense of archaeological data. As discussed below, digital approaches to archaeology, like any other methodology, should be considered contestable. Keeping the playing field open for multiple technical, semantic, and even ethical perspectives is therefore in the interest of the discipline as a whole. However, many archaeological data-sharing projects exist only on limited grant-funded support. Nevertheless, these may be innovative and may publish valuable content while they explore important questions in interface design and technology. An important goal should be to ensure continued experimentation and innovation of these distributed initiatives while safeguarding and preserving data. Standards efforts and archaeological cyberinfrastructure should focus on supporting widely distributed digital efforts to help ensure that their contributions will outlast their grant funding. We hope future efforts will find feasible and cost-effective strategies to enable “data preservation as a service” so that content can be preserved by the organizations most capable of doing so, while reducing the costs and risks of innovation and experimentation in different digital methods. The California

Digital Library's model for preservation micro-services represents a very encouraging step in this direction. Establishing a reliable preservation infrastructure open for the widely distributed community to build upon would encourage greater dynamism in this field.

## INFORMATION OVERLOAD AND ITS DISCONTENTS

In Chapter 10, Harrison Eiteljorg highlights a great challenge in sharing data with contributory systems: How does one ensure that the content user fully understands the project the data come from and the data collection process itself? In other words, how can a user understand a data set if that user was not involved in its creation?

The same question can be asked of synthetic publications for field projects: How can readers understand the project if they were not involved in it? As discussed above, knowledge has tacit components that often go unrecognized. Thus, even researchers who strive to communicate as comprehensively and transparently as possible will probably not be able to provide enough explicit metadata and explanation to reveal all the assumptions, motivations, and decision-making behind their data. While often an admirable goal, total transparency in archaeological research will probably always be unattainable.

Whether explicit or not, various contributors to this volume offer approaches to the problem raised by Eiteljorg. Some contributors, such as Boast and Biehl (Chapter 4), favor greater attention to linking archaeological data sharing to narratives and interpretations. They argue that digital representations of cultural heritage often find the greatest meaning embedded within narratives. In that sense, they question the universal utility of metadata and the structural formalisms of disciplinary semantic standards. The Kansas (Chapter 2) also argue that data sharing needs to be linked with narratives, both for the sake of intelligibility and to better fit with familiar patterns of scholarly communications.

While narratives can offer more depth to guide interpretation, unfortunately, deep reading of contextual and narrative nuance "does not scale." Archaeologists, like many other twenty-first-century knowledge workers, face increasing demands on their attention. While we work to produce more and more documentation, analyses, and interpretations about the past, we seemingly have less time and attention to devote to understanding this wealth of data. Web 2.0 systems both help and hinder in that regard. On the positive

side, user-generated tags, ratings, and recommendations from social networks may help one rapidly find useful information. Whitcher Kansa and Deblauwe (Chapter 7) argue this point in the context of archaeological blogging and in the context of social media use among the zooarchaeology community. On the negative side, participation in these social networks requires precious attention. One may find too much useful information to adequately process and understand. Unfortunately, information overload problems are not limited to the world of Web 2.0, and many scholars lament the glut of literature published by their colleagues (see Harley et al. 2010: 37–38).

Thus, information overload is one of the most critical problems archaeologists face today. To help mitigate information overload, some emphasize common standards, including formal domain ontologies to explicitly define the meaning of archaeological data according to widely held community understandings. This approach has the advantage of being automation-friendly. Human effort and attention are in short supply, and the more computer systems can automate documentation, retrieval, and aggregation of archaeological content, the more content researchers can hope to use. NLP, text mining, the Semantic Web, and other automation techniques offer useful strategies to help archaeologists understand and utilize their colleagues' research findings, and overcome information overload (see Crane's 2006 insightful discussion). In this regard, web services and other approaches to integrate different collections also relate to this discussion. Such services help pool data from multiple sources, making search, retrieval, and use of the data easier and more efficient.

While promising for some applications and research perspectives, understanding the past through algorithmic processes will probably not be universally welcome. Efficiency has tradeoffs, especially if your theoretical perspective is more "reflexive." Relying upon semantic standards or machine-produced metadata will be somewhat "lossy,"<sup>15</sup> in the sense that local nuance and context may be lost to imperfect and partial mappings to a global standard. Moreover, any standard or algorithm privileges a certain set of expectations and goals. Who will set the agenda in determining the semantic standards behind automation? What perspectives will become enshrined and

---

<sup>15</sup> "Lossy" is a term referring to how some compression algorithms degrade quality and fidelity of images and other digital media in order to reduce storage and transmission costs.

codified as required standards by funding bodies and professional societies, and what perspectives will be left on the margins?

It does not take much imagination to see emerging theoretical tensions between archaeological knowledge production driven from algorithms and formalized ontologies versus archaeological knowledge constructed from different threads of narrative. In some ways, the tensions between advocates of “deep reading” and advocates for “interoperability” continue long-standing theoretical disputes in archaeology. Some researchers emphasize contextual nuance and particularistic interpretations, while others seek more generalized patterns in more or less interchangeable empirical data. Each different theoretical orientation fits better with a different type of technical style and systems implementation.

One would hope that the discipline will benefit from the best ideas of both the “deep reading” and the “interoperability” perspectives. Transparency and openness in analytic methods as well as in data sources should be a key requirement for technologically enabled archaeological research. Data sources, services, and software open to “deep reading” can earn greater trust. For example, it would be much better if the corpora and algorithms used in a text-mining project were open for others to use and adapt to serve other agendas. Without such openness, it is impossible to go beyond the perspectives, assumptions, and limitations of the initial text-mining or semantic data project. Openness to critique and outside improvement can lead to greater trust and legitimacy in archaeological information systems, even if few will have the time and inclination to actually bother with inspecting their inner workings.

The intersection of archaeological theory and digital technologies needs far more exploration. While we should avoid being “techno-determinists,” we would be foolish to ignore the role of technology in shaping scholarly life, including theoretical outlooks. It will be interesting to see how new technology opportunities and challenges co-evolve with theoretical trends in archaeology. How will ready access to structured data text-mined from over a thousand publications change archaeological interpretation? How will the professional community evaluate the significance of a sprawling multi-threaded conversation taking place between museums and distributed social media outlets? Who will have the time and attention to devote to deep reading, and where will they focus their attention? What sorts of information will

be taken for granted and what will attract great scrutiny? How much will *information convenience* drive future research agendas? All of these are important topics for continued research.

## FUTURE DIRECTIONS AND TENSIONS: A BRAVE NEW WORLD FOR THE PAST

Many contributions in this volume note that Web 2.0 technologies and data sharing need to work in the context of scholarly communications. This volume offers many cautionary tales about applying emerging technologies in professional settings where such technologies may clash with incentives and perceptions of risks and rewards. Specific technologies change rapidly, but many of the issues explored in this volume will have long-lasting significance. The evolution of scholarly communication and how researchers recognize and communicate expertise and authority will remain important topics long into the future. Similarly, no matter what specific technology we deploy, we must grapple with how interfaces, data structures, and architectures are guided by, and also guide, interpretive priorities. Thus, many of the concerns explored in this volume will foreshadow areas of future research and debate, even after the term “Web 2.0” loses currency.

In looking at the longer-term impacts of these discussions, it is impossible to ignore more general trends shaping the public Web. Archaeologists, even those working on cyberinfrastructure initiatives, may not be the primary agents shaping the future of archaeology’s digital communications. Already, Google has reshaped how students and researchers search and retrieve scholarly content, an issue touched on by many chapters in the volume. While some of Google’s search and ranking algorithms are known (especially PageRank), other algorithms behind search results are trade secrets. Moreover, Google and other search engines continually change their methods and often offer personalized recommendations to individual users with little transparency. Potentially, every user of Google gets different search results, algorithmically personalized to their interests and search history. What does this mean to researchers searching through archaeological literature? How does this challenge or reinforce personal biases? How do personalized recommendations help shape archaeological discourse? These are important issues for further discussion.

As we look ahead, the ways in which archaeological content is aggregated, ranked, and presented will, in large part, be driven by the needs and interests of commercial Web giants. An increasingly large part of archaeological information retrieval is being shaped by “black-box” processes invisible to the research community. Other disciplinary communities face similar issues. But because archaeology has important relationships to tourism and marketing of cultural heritage, it is likely to feel disproportionately greater impact from emerging commercial information services than other disciplines. These extend beyond search and include various Web-based collaboration, visualization, and mapping applications. For better or worse, Google Maps, Google Earth, and social media services are now the windows through which many students and the public will encounter archaeological data. With the tremendous growth of mobile computing and location-based services, Google and other commercial web giants will likely play an increasing role in shaping how cultural heritage is delivered and presented on the Mobile Web (Kansa and Wilde 2008). The ways that commercial aggregation and ranking of cultural heritage will affect public perception and experience of the past will deserve increasing scrutiny (see also Vaidhyathan 2011).

As technologies for disseminating, organizing, and retrieving information increasingly shape archaeological communications, debates about the theoretical implications and assumptions behind those technologies will receive greater attention. In that sense, this volume is an early sample of conversations to come.

## ACKNOWLEDGMENTS

I thank my co-editors Sarah Witcher Kansa and Ethan Watrall for providing synopses of many of the contributions to this volume and advice on the structure of this chapter.

## REFERENCES CITED

- Borgman, C. L.  
 2007 *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*. Cambridge, MA: MIT Press.
- Brin, S., and L. Page  
 1998 The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems* 30: 107–117.



Crane, G.

- 2006 What Do You Do with a Million Books? *D-Lib Magazine* 12. Retrieved from <http://www.dlib.org/dlib/march06/crane/03crane.html> (accessed October 16, 2008).

“Data’s shameful neglect”

- 2009 Editorial comment. *Nature* 461: 145 (September 10, 2009).

Harley, D., S. K. Acord, S. Earl-Novell, S. Lawrence, and C. Judson King

- 2010 Assessing the Future Landscape of Scholarly Communication: An Exploration of Faculty Values and Needs in Seven Disciplines. UC Berkeley, Center for Studies in Higher Education. Retrieved from [http://escholarship.org/uc/cshe\\_fsc](http://escholarship.org/uc/cshe_fsc) (accessed June 22, 2010).

Kansa, E. C., and E. Wilde

- 2008 Tourism, Peer Production, and Location-Based Service Design. In *IEEE International Conference on Services Computing*, vol. 2: 629–636. Los Alamitos, CA: IEEE Computer Society.

Lanier, J.

- 2010 *You Are Not a Gadget: A Manifesto*. 1st ed. New York: Knopf.

Markey, K.

- 2007 The Online Library Catalog. *D-Lib Magazine* 13. Retrieved from <http://www.dlib.org/dlib/january07/markey/01markey.html> (accessed October 15, 2010).

O’Reilly, T.

- 2005 What is Web 2.0?. Retrieved from <http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html> (accessed May 28, 2009).

Vaidhyanathan, S.

- 2011 *The Googlization of Everything*. Berkeley: University of California Press. <http://www.ucpress.edu/excerpt.php?isbn=9780520258822#readchapter1> (Chapter 1 excerpt accessed September 10, 2010).

Standage, T.

- 1998 *The Victorian Internet*. New York: Berkley Publishing Group.

Yu, H., and M. Young

- 2004 The Impact of Web Search Engines on Subject Searching in OPAC. *Information Technology & Libraries* 23/4 (December): 168–180.



## A WEB OF ARCHAEOLOGICAL DATA: INFRASTRUCTURE, SERVICES, AND INTEROPERABILITY

This section explores the range of technical approaches, from mash-ups to more formalized Semantic Web systems, to putting archaeology on the Web. The technical choices illustrated here reflect a variety of differences in theoretical perspectives and institutional settings.

One of the key developments of the Web 2.0 era has been the proliferation of content aimed primarily at consumption and further processing by software, and only secondarily at people. Such content, often called “machine-readable” data, is expressed in structured formats that are easily processed by software. Web services, or APIs (application program interfaces,) make data available for use by remote, third-party-developed software applications. In more informal settings, machine-readable data make “mash-ups” possible. Mash-ups take data from different sources and combine and aggregate them in new ways. Relatively little effort and basic programming skills are needed to create mash-ups. Open Context, described in Chapter 2, represents an example of this approach to disseminating machine-readable data.

As mash-ups grew in popularity during the first decade of the twenty-first century, researchers in academic and commercial labs devoted great effort to developing the concepts, standards, and technologies of the so-called Semantic Web. Sir Tim Berners-Lee, the key initial architect of the World Wide Web, helped to spearhead and promote Semantic Web research and development. According to Berners-Lee and his fellow travelers, the Semantic Web (sometimes called “Web 3.0,” or now “Linked Data”; see below) represents the next stage in the Web’s evolution.

Like “mash-up” approaches, the Semantic Web emphasizes machine-readable data. However, the Semantic Web is more ambitious and usually much more formalized in its approach. As illustrated in the discussion of Open Context, the web services typical of the mash-up world have a great

deal of ambiguity. With Open Context, as in most mash-up scenarios, data may be easy to parse and process by software, but the meaning of those data is not necessarily clear. A human programmer usually decides whether and how to relate data elements from different sources. That requires the programmer to have some specific understanding of the different data sources used in the mash-up. In contrast, Semantic Web / Linked Data is not only easy for software to process, but that same machine-readable data will also clearly link to specific and formally defined concepts that denote meaning. Semantic Web approaches, potentially, do not require a programmer to have intimate knowledge of each and every data source being aggregated, since in Semantic Web scenarios, these different data sources “self-declare” the meaning of their content, at least to some extent.

While Semantic Web technologies can help clarify the meaning of data, the difficulties inherent in this approach have thus far slowed adoption. Semantic Web approaches are often much more formalized and conceptually difficult. They involve still largely unfamiliar technologies and data formats. Semantic Web approaches also often use conceptually difficult ontologies, expressed in difficult formats and formalisms, such as RDF-OWL. Finally, using a shared ontology requires participants in the Semantic Web to agree on a common standard of meaning. Determining semantic standards has obvious political dimensions. Who sets the agenda and for what purpose? Which sorts of meanings are important and allowed in a given semantic standard, and which are excluded? Can the meaning of “cultural heritage” (including the archaeological record) be reduced to a single set of universally shared concepts?

Because they require more specialized forms of expertise and technologies, Semantic Web approaches tend to see greater adoption in more centralized and better-funded settings. The chapter by Richards et al. (Chapter 1) represents a case in point. The Archaeology Data Service (ADS) is the leading data repository in the United Kingdom. It works with various government heritage agencies to manage large-scale collections that see great use because of their size, comprehensiveness, and authority. Semantic Web standards and approaches see implementation in similarly centrally managed heritage collections in other parts of the European Union where state-sponsored heritage organizations manage large data resources. In these more centralized settings, the costs of implementing formal semantic standards and

Semantic Web technologies are easier to justify. In more centralized settings, it is also easier to formally define and enforce a common ontology, such as the CIDOC-CRM for describing data.

In contrast, the United States conspicuously lacks large, centrally managed archaeological data resources. Use of Semantic Web/Linked Data technologies is therefore harder to justify in the U.S., since the effort and expense in implementing Semantic Web technologies will typically benefit data collections that see limited use. The Semantic Web in general has also generated somewhat less enthusiasm in U.S. academic and commercial settings than it has in the European Union. After over ten years of promotion by Tim Berners-Lee, Google is only now taking its first limited steps in supporting services based on Semantic Web data.

However, emerging technologies, including natural language processing (NLP) and text mining, may help pave the way for more rapid growth of Semantic Web data in archaeology. Richards et al. describe how the Archaeotools Project uses NLP software to extract important metadata from large corpora of gray literature. This removes a great deal of the human effort previously required to classify and describe documents. Metadata extracted in this way can be harnessed for Semantic Web applications.

At the same time, some recent developments in Semantic Web technologies and standards are reducing barriers to entry. This change is reflected in the rebranding of the Semantic Web as “Linked Data,” a less conceptually loaded and grandiose term for Semantic Web approaches. Linked Data has connotations for more incremental adoption, reduced complexity, and less emphasis on universal semantic harmonization. As a result, Linked Data approaches are gaining traction among the community of “mash-up” developers on the Web. Key Linked Data architectural principles, especially retrieval of web resources and referencing of shared concepts using stable URL/URIs (hyperlinks), are increasingly common on the open Web. Thus, even a project like Open Context, which began with more of a “mash-up” orientation, recently began to adopt many Linked Data standards and approaches. We should expect continued co-evolution of Linked Data and “mash-up” approaches on the Web in general, and in digital archaeology in particular. No doubt, different projects will continue to explore the dynamic tensions between semantic formalism and informalism and find different solutions in this continuum.



# The Archaeology Data Service and the Archaeotools Project: Faceted Classification and Natural Language Processing

*Julian Richards, Stuart Jeffrey, Stewart Waller,  
Fabio Ciravegna, Sam Chapman, and Ziqi Zhang*

This chapter describes the development of **Archaeotools**,<sup>1</sup> a major cyber-infrastructure project in archaeology. The goal of the Archaeotools Project was to use faceted classification and natural language processing to create an advanced infrastructure for archaeological research. The project aimed to integrate over one million structured database records referring to archaeological sites and monuments in the United Kingdom with information extracted from semi-structured gray literature reports and unstructured antiquarian journal accounts, in a single-faceted browse interface. The project has highlighted the variable level of vocabulary control and standardization that currently exists within national and local monument inventories. Nonetheless, it has demonstrated that the relatively well-defined ontologies and thesauri that exist in U.K. archaeology mean that a high level of success can be achieved using information extraction techniques. Further refinement of the machine-generated indexing might be achieved through a Web 2.0-style application that would allow users to make corrections or to tag additional terms according to their own “folksonomies.” The resulting indexes could also be made available via a variety of web services, including Web 2.0-type “mash-ups” for incorporation within other interfaces.

---

<sup>1</sup> <http://archaeologydataservice.ac.uk/research/archaeotools>

## INTRODUCTION

The **Archaeology Data Service (ADS)**<sup>2</sup> has been providing free online access to digital resources for teaching, learning, and research in the United Kingdom since 1998. The **ArchSearch catalogue**<sup>3</sup> now provides access to over one million metadata records which reference the archaeology of the British Isles. These are mainly derived from the National Monuments Records (NMRs) of England, Scotland, and Wales, as well as from county-based records, now known in the United Kingdom as Historic Environment Records (HERs). The ADS acts as an information broker and aggregator for these resources, providing a “shop window” that is used especially within the higher education community, and also within the commercial sector.

In addition, since 2005 ADS has held a steadily growing library of unpublished fieldwork reports, or “**gray literature**.”<sup>4</sup> In the United Kingdom, about £125 million is spent per annum on developer-funded archaeology required under Planning Policy Guidance, with an average of 6,000 interventions per annum. In the United States, between \$650 million and \$1 billion is spent annually on cultural resource management, a large proportion of which is devoted to archaeology. Nearly all of this work is performed to comply with laws that require government agencies to take into account the effect of their actions on archaeological and historical resources. On the order of 50,000 field projects a year are carried out by federal agencies under these mandates, with another 50,000 federal undertakings requiring record searches or other inquiries that do not result in fieldwork. However, there is no legal requirement to publish the outcome of this activity, either in the United States or the United Kingdom.

On both sides of the Atlantic, therefore, this activity generates vast numbers of reports that together constitute the unpublished “gray literature” whose inaccessibility has long been an issue of major concern. With so much work being performed and so much data being generated, it is not surprising that archaeologists working in the same region—let alone those working in different continents—do not know of one another’s work. Decisions about whether to preserve particular sites, how many sites of specific types to excavate, and how much more work needs to be done are being made in an infor-

---

<sup>2</sup> <http://archaeologydataservice.ac.uk/>

<sup>3</sup> <http://archaeologydataservice.ac.uk/archsearch/>

<sup>4</sup> <http://archaeologydataservice.ac.uk/archives/view/greylit/>



mational vacuum. Furthermore, new data are not fed into the research cycle, and academic researchers may be dealing with information that is at least ten years out-of-date. Nonetheless, the fact that such reports are not fully published should not be taken to suggest that the value of the archaeological data or interpretation is not significant enough for publication (Falkingham 2005).

In recent years, the academic community has begun to recognize the detrimental effect of having such large amounts of archaeological information tucked away and inaccessible. Researchers such as Bradley (2006) and Lock (2008) have questioned why it is not more widely available. Digital collection and online delivery of both newly created (i.e., “born digital”) and legacy material could provide a solution to these access issues. However, good access is predicated on good discovery mechanisms, and these rely on, among other things, good data about data, or metadata.

As part of the **OASIS Project**,<sup>5</sup> the ADS actively gathers digital versions of gray literature fieldwork reports (Hardman and Richards 2003; Richards and Hardman 2008). The ADS gray literature library currently (as of June 2011) comprises over 10,000 reports, although it is increasing at the rate of 50–100 per month. All reports can be downloaded free of charge, and there is a high level of demand; from May to July 2010, there were 44,483 downloads. Since the library was launched in August 2005, there have been 400,000 downloads. Each of the reports has manually generated resource discovery metadata covering such attributes as author, publisher, and temporal and geospatial coverage, adhering to the **Dublin Core metadata standard**.<sup>6</sup> Generating metadata this way may be feasible where it is created simultaneously with the report’s deposit with the ADS. It would not be a feasible means of dealing with the tens of thousands of legacy reports known to exist. Any attempt to digitize these disparate and distributed sets of records to facilitate broader access would require automated metadata generation.

Within the ADS digital library, there are also electronic versions of more conventional journals and reports, including all *Council for British Archaeology Research Reports* 1–100<sup>7</sup> and a complete back-run of the *Proceedings of the Society of Antiquaries of Scotland* (PSAS)<sup>8</sup> going back to 1851. Many of the same

---

<sup>5</sup> Acronym for Online Access to the Index of Archaeological Investigations; <http://www.oasis.ac.uk/>.

<sup>6</sup> <http://www.dublincore.org/>

<sup>7</sup> [http://archaeologydataservice.ac.uk/archives/view/cba\\_rr/](http://archaeologydataservice.ac.uk/archives/view/cba_rr/); doi:10.5284/1000332

<sup>8</sup> <http://archaeologydataservice.ac.uk/archives/view/psas/>; doi:10.5284/1000184

indexing issues arise with reference to digitized versions of such early or short-run published material. As an increasing number of journal back-runs are digitized and held within large online libraries such as JSTOR, or by smaller, discipline-specific repositories such as ADS, providing deeper and richer access to these resources becomes an increasing priority. Furthermore, there is the potential to provide joint access to published and unpublished literature within a single interface, allowing users to cross-search both types of resource. However, indexing of journal back-runs rarely goes beyond author and title. This is generally inadequate for the scholar wishing to investigate previous research on a particular site or artifact class. Furthermore, while modern fieldwork reports generally provide Ordnance Survey grid references for site locations, antiquarian reports use a variety of nonstandard and historic place-names, making it impossible to integrate this sort of information into modern geospatial interfaces. Ideally, a methodology to automatically generate metadata for gray literature should be flexible enough to be applicable to this additional data set with the minimum of reworking.

However, the main role of ADS is as a long-term preservation and access facility for primary data, and for over 300 individual sites ArchSearch now provides a way into much richer online archives, containing site context and finds databases, CAD or GIS plans, as well as text reports and digital photographs. While some of these archives also have their own individualized “special collection” query interface, users are currently unable to search across these heterogeneous data sets, but must interrogate each archive in a serial manner. In the longer term, ADS wishes to facilitate cross-searching at a deeper and richer level across all of its collections. A faceted query system again appears to offer the most powerful solution. Such a system would allow users to search simultaneously within published journal articles, unpublished fieldwork reports, and data archives, and to draw their own links between these different types of resource.

There are a number of ways that users of the ADS website could previously locate the resources they were interested in. The ArchSearch interface provided three ways to search: a standard “search box” basic keyword search; a more advanced search in which users could develop queries from any combination of “What,” “When,” and “Where”; and a crude, clickable map that could also be used to define a geographic area of interest. In addition, if users already knew that the ADS held an archive for a site of interest, there was a simple page providing a link to every project archive, sorted by region or

theme. However, as the number of ADS collections has grown, traditional search methods have become increasingly inadequate. The classic empty search box has always relied on users having some prior knowledge of the index terms held in the ArchSearch database. While this is generally adequate for specific site types, such as “moated manor house” (assuming the user knows the proper terminology), it is less satisfactory for users interested in more general themes, such as “Bronze Age landscapes.” As the size of the information resource grows, it can also become increasingly difficult to isolate the results of interest from the hundreds or even thousands of peripherally relevant results. Most search engines deal with this by ranking the results by some measure of relevance, with the net effect that only those hits returned within the first page or two are investigated, with the potential risk of missing important resources. In the days when archaeological records could only be consulted in person in the offices of the agency responsible for that region, there was usually a human expert on hand who had an intimate knowledge of the peculiarities of their own system and who could help guide the researcher to the resources of interest. However, as more and more research is conducted remotely and online, there is a need to replace the human information specialist with more sophisticated and powerful automated search tools (Robinson 2007). Such tools should not only allow users to find specific items of interest according to prerecorded index terms, but they should also allow them to create user-defined trails according to new themes of interest.

During 2004–05, the ADS and Adiuri Systems developed the **Archaeobrowser**,<sup>9</sup> a proof-of-concept archaeological faceted classification demonstrator on behalf of the **Common Information Environment (CIE)** group.<sup>10</sup> The Archaeobrowser used the Waypoint application developed by Adiuri Systems. Waypoint and other faceted browsers seek to overcome the disadvantages of a Google-like “type and hope” approach by replacing it with a “point and browse” search strategy. The Archaeobrowser has demonstrated the power of a faceted query approach in allowing users to narrow down a search from over one million records to a realistic handful of relevant results with just two or three mouse clicks. However, as can often be the case with

---

<sup>9</sup> <http://archaeologydataservice.ac.uk/research/cie/>

<sup>10</sup> <http://www.jisc.ac.uk/aboutus/committees/workinggroups/disbanded/commoninfoenvironment.aspx/>

many faceted classification applications, the level of computing power required to produce the faceted index proved unsustainable for a working system. The lessons learned in this project led the ADS into a partnership with the Natural Language Processing Research Group at the University of Sheffield, which had developed software for several working systems in industrial applications. The Archaeotools Project was funded under the United Kingdom's Arts and Humanities e-Science Initiative, a program designed to bring researchers from the arts and humanities together with computer scientists, and itself a collaborative scheme between three major U.K. funding bodies: the Arts and Humanities Research Council (AHRC), the Engineering and Physical Sciences Research Council (EPSRC), and the Joint Information Systems Committee (JISC).

The Archaeotools Project therefore addressed two practical issues in applying advanced information extraction techniques to data sets generated in the arts and humanities: first, the creation of search mechanisms that go beyond the naive text-string searching approach of the classic search engine search box; and second, the automated creation of the resource discovery metadata required to underpin these more sophisticated searches. The solutions to these two broad issues—automatic metadata extraction, and browsing by facet—are, in fact, extremely complementary. It is the Archaeotools implementation of these solutions together that offers such potential. Not only is it intended that the faceted classification browser work as an interface to the aggregated data sets hosted by the ADS, but it is also intended that the gray literature holdings, and even historic literature holdings, will be integrated into these data sets, making them discoverable and searchable via the same faceted browsing interface. In short, the objective of the project is to allow archaeologists to discover, share, and analyze data sets and legacy publications that, despite their importance, have hitherto been either impossible or very difficult to integrate into existing digital frameworks. Furthermore, although outside the scope of the Archaeotools Project, it is anticipated that the same resources will also be made available as web services to allow users to consume them within other search interfaces.

The Archaeotools Project has followed a trajectory that has allowed it to achieve its aims in three discrete stages:

1. The creation of an advanced faceted classification and geospatial browser. The underlying data set comprises over 1,000,000 records (held in Or-

acle RDBMS) aggregated from the National Monuments Records of Scotland, Wales, and England as well as Historic Environment Records from numerous local authorities and the ADS's own archive holdings. The facets selected are the standard hierarchical "What," "Where," and "When" facets, plus a "Media" facet to allow the selection of particular subsets of resources. The facets are populated from existing thesauri (such as the *Thesaurus of Monument Types*) in XML format and extended and integrated to allow for geographical differences, such as terminological differences in monument and period types between Scotland and England. The Archaeotools Project also integrates thesauri served in XML by web services based on **Simple Knowledge Organisation Systems** (SKOS)<sup>11</sup> and developed by the Arts and Humanities Research Council-funded **Semantic Tools for Archaeology (STAR) Project**<sup>12</sup> based at the University of Glamorgan. SKOS provides a simple means of expressing thesauri and can help widen the scope of their usability.

2. The creation of a reusable natural language processing (NLP) system to automatically extract resource discovery metadata (and other facet types) from 1,000 unpublished archaeological gray literature reports.
3. The extension of the NLP systems to capture metadata from legacy historical documents, using the *Proceedings of the Society of Antiquaries of Scotland* as an exemplar corpus and utilizing the University of Edinburgh's EDINA **geoXwalk**<sup>13</sup> service to recast place-names and locations extracted from text as national grid references (NGRs), allowing enhanced geospatial searching of the data.

### THE FACETED CLASSIFICATION BROWSER: ARCHSEARCH III

By 2009 the search mechanism and interface to the ADS aggregated data sets was called ArchSearch II, having evolved from the ADS's original ArchSearch mechanism developed in the late 1990s. The Archaeotools Project was designed to further develop this search mechanism into a faceted classification

---

<sup>11</sup> <http://www.w3.org/2004/02/skos/>

<sup>12</sup> <http://hypermedia.research.glam.ac.uk/kos/star/>

<sup>13</sup> <http://www.geoxwalk.ac.uk/>

browser and associated interactive geospatial search. The faceted classification approach to presenting structured data sets is increasingly common in the commercial Web, but clearly lends itself to the discovery of any structured data set (Denton 2003). A faceted query engine has been employed by a team at Columbia University to provide an interface to archaeological finds data sets (Ross et al. 2005, 2007) and to the **Open Context** system<sup>14</sup> at the Alexandria Archive Institute (see Kansa and Whitcher Kansa, Chapter 2, this volume); applications of faceted classification are becoming increasingly common in the cultural heritage sector.

Previous work on user needs and faceted classification carried out by the ADS in the Archaeobrowser Project (Jeffrey et al. 2008) demonstrated that the most appropriate search facets for archaeological data sets are:

- ▶ What—what subject(s) does the record refer to?
- ▶ When—what is the archaeological date range of interest and exact singular temporal point?
- ▶ Where—what is the location(s) or region(s) of interest?
- ▶ Media—what is the form of the record you are ultimately interested in?

These are far from being the only possible facets, and some others can be seen as highly desirable (e.g., Who—to whom does the record relate?), but as a matter of practicality, these four facets are expected to offer the greatest utility for the archaeological researcher. As mentioned above, the Archaeo-tools Project has also investigated how additional facets might be specified and whether user-generated facets are either desirable or feasible.

To facilitate browsing, each facet needs to have an associated ontology, expressed as a hierarchy of terms. Fortunately, in the historic environment sector there are hierarchical thesauri deployed or under development that allow a browsing structure to be populated for each facet, apart from “Media.” These thesauri, or controlled word lists, have been generated from a number of sources, but it is key to their usefulness and sustainability that each has a controlling body, each is recognized as a *de jure* or *de facto* standard, and each is either already being broadly used or is in the process of being adopted. For the above facets, three thesauri were selected:

---

<sup>14</sup> <http://www.opencontext.org/>

What—**The Thesaurus of Monument Types** (TMT, English Heritage, 2008)<sup>15</sup>

When—**MIDAS Period list** (MIDAS, 2008)<sup>16</sup>; **FISH**, 2008<sup>17</sup>

Where—County, District, Parish (CDP) (U.K. Government list of administrative areas)

The hierarchical structure for a (hypothetical) detailed record of the monument type “Tower Keep” might look like this:

What→

Defense→

Castle→

Keep→

Tower Keep

This example shows that the hierarchical structure lends itself to a “point-and-click” browsing approach such that each level of the hierarchy can be expanded or collapsed by a mouse click. Each record in the target data set is assigned a “What,” “When,” and “Where” value from the selected thesauri. The power of this approach for a normalized data set is demonstrated by a user’s ability to drill down to a specific (and complete) set of records with the minimum of clicks. In tests on the Archaeobrowser system, it was possible to go from the maximum number of 1,000,000 or so records to a selected set of 16 records representing Bronze Age funerary monuments within 5 km of a specific location in North Yorkshire with just three or four clicks of the mouse (depending on the route taken). Not only does this compare very favorably with traditional search-box-based techniques, but the fact that the data have been mapped to the terms of the thesaurus means that the user can have a much higher level of confidence in the completeness of the returned results and is much less troubled by the return of false positive results. The indexing mechanism adopted by the Archaeotools Project was built on top of Solr, an open source enterprise search server based on the **Lucene** Java search library.<sup>18</sup>

---

<sup>15</sup> [http://thesaurus.english-heritage.org.uk/thesaurus.asp?thes\\_no=1](http://thesaurus.english-heritage.org.uk/thesaurus.asp?thes_no=1)

<sup>16</sup> <http://www.midas-heritage.info>

<sup>17</sup> <http://www.fish-forum.info/index.htm>

<sup>18</sup> <http://lucene.apache.org/solr/>

The screenshot shows the Archaeology Data Service (ADS) website. The header includes the ADS logo and navigation links: HOME, ARCHSEARCH, ARCHIVES, LEARNING, ADVICE, OUR RESEARCH, ABOUT US, myADS, and LOGOUT. A search bar is located in the top right corner.

The left sidebar contains a faceted browsing tree:

- WHAT**
  - Aspire Event 2
  - Monument Types **42**
  - Oasis Event 2
  - Object Types **42**
- WHERE**
  - England 7
  - Scotland 42**
    - Borders **14**
    - Central **14**
    - Dumfries And Galloway 13
    - Fife 13
    - Grampian 24
    - Highland 42**
    - Lothian 14
    - Orkney Islands Area 10
    - Shetland Islands Area 10
    - Strathclyde 15
    - Tayside 13
    - Western Isles Area 26
- WHEN**
  - Early Prehistoric 1
  - Later Prehistoric 1
  - Roman 35
  - Early Medieval 42**
  - Medieval 39
  - Post Medieval 7
  - Modern 4

The main search results area shows 42 total results for the query: **WHAT** Cross, **WHERE** Scotland, Highland, **WHEN** Early Medieval. The results list includes:

- DYCE SAINT FERGUS CHURCH PICTISH CROSSSLAB** (National Monuments Record of Scotland). Alternative name(s): CHAPEL OF ST FERGUS, OLD PARISH CHURCH OF DYCE; DYCE, OLD CHURCH. ABERDEEN, CITY OF.
- ULBSTER THE ULBSTER STONE** (National Monuments Record of Scotland). No description. HIGHLAND.
- ALTYE HOUSE** (National Monuments Record of Scotland). No description. MORAY.
- TARBAT** (National Monuments Record of Scotland). No description. HIGHLAND.
- ELGIN CATHEDRAL PICTISH CROSSSLAB** (National Monuments Record of Scotland). No description. MORAY.

The right sidebar shows a map of Scotland with a vertical label 'EXTERNAL MAP'.

Figure 1.1. A screen-shot of the prototype faceted classification browsing tree for the Archaeotools Project. The window on the left shows the “What,” “When,” and “Where” ontologies, with the “Where” and “When” trees expanded to the first level. Early Medieval Crosses in the Highland region have been selected. The numbers in bold after each term indicate the number of records classified according to that facet. The window on the right displays the first 5 records of the 42 returned by this query (on test data).

## RECORD TO ONTOLOGY MISMATCHES

Any large monument inventory, indeed any large data set, especially one that has developed over a number of years, is unlikely to conform perfectly to any rigid terminology standards, especially if these were created subsequent to the inception of the data set. The Archaeotools Project is the first instance of



any archaeological project in the United Kingdom that has both generated metrics for these mismatches and mitigated the problem via a combined automated and manual approach. This mitigation generated interesting statistics, which are summarized in Table 1.1. It is true to say that all data sets contributed to these mismatches more or less equally and that there was no obvious data set where the terminology used diverged more radically from the thesauri than all the others.

The numbers given in Table 1.1 are derived from a total aggregated record set of 1,001,107 records, and all percentages represent a percentage of this number.

Contrary to our original expectations, it has proved possible to completely map these record sets to the thesauri, and therefore to the facets, by a combination of automatic rule-based expressions and manual techniques. The “When” facet provides an example of the success of this combined approach. There is a large number of ways in which archaeological dates and date ranges can be written—for example, 1066, 1001–1100, 11th century

Table 1.1. Ontology mismatches between record entries and the appropriate thesauri after initial analysis

FACET / RECORD DESCRIPTION	QUANTITY
What	
Records that have no subject information	19,269 records (2%)
Records that use terms not found in TMT, so these records cannot be indexed (6,442 unique terms)	101,507 records (10.1%)
When	
Records that have no temporal information	292,793 records (29.2%)
Records that use period terms not found in MIDAS, so these records cannot be indexed (457 types of irresolvable dates)	114,505 records (11.4%)
Where	
Records that have no spatial information	11,126 records (1.1%)
Records that use terms not found in CDP, so these records cannot be indexed.	245,601 records (24.5%)

Source: Jeffrey et al. 2009 and forthcoming.

(*sic*), C11, 11C, eleventh century. Most of these were mapped directly to MIDAS-defined date ranges. Analysis initially recovered 457 instances of irresolvable dates, equating to 114,505 records that could not be classified. After automated processing using regular expressions, however, this was reduced to 148 concepts and only 7,528 records. This is a perfectly manageable number to expect to be corrected by manual intervention. Similarly, the initial figures for “Where” with terms not found in the County/District/Parish (CDP) list (24.5%) can be safely ignored, as these figures were generated prior to the integration of the Scottish CDP list into the thesauri set; this comfortably accounts for the majority of these missing terms. The variety of uncontrolled terminology used for the “What” facet, combined with a significant number of records with no subject information, proved more intractable but was not a serious problem, as most records still appeared under either the “When” or “Where” facet. In total, of 1,001,595 records submitted for classification, 995,907 appeared in at least one facet, leaving only 5,688 records totally unclassified.

Having successfully constructed a facet tree from the structured data derived from database records, the Archaeotools Project then turned to the application of information extraction techniques to allow the incorporation of free text-based resources.

## NATURAL LANGUAGE PROCESSING AND INFORMATION EXTRACTION

Natural language processing (NLP) is the branch of artificial intelligence concerned with extracting meaning from human speech and text. Ontology-based text annotation is seen as making a key contribution to the development of the Semantic Web (Uren et al. 2006). While the present-day Web has largely been built for human consumption, the Semantic Web incorporates annotated texts, thereby allowing machines to make linkages between items of structured information. It is therefore seen as having tremendous potential for research. However, work on the Semantic Web has focused on commercial applications; there have been few research-driven projects, and fewer still in the arts and humanities. Archaeology has some potential as a test-bed in this field because, despite its humanities-based focus, it has a relatively well-controlled vocabulary. Amrani et al. (2008) have reported on a pilot application in a relatively specialized area of archaeology; the Open-

Boek Project experimented with memory-based learning in extracting chronological and geographical terms from Dutch archaeological texts (Paijmans and Wubben 2008), while the STAR Project has focused on mapping data sets to the **CIDOC-Conceptual Reference Model** (CIDOC-CRM)<sup>19</sup> (Binding et al. 2008). Byrne has also explored the application of NLP to extract event information from archaeological texts (Byrne and Klein 2010). In the United States, Giles and his colleagues (1998) have developed Archseer, an adaptation of their successful **CiteSeer system**<sup>20</sup> to archaeology. Archseer provides the ability to search archived literature by author, title, abstract, text, or citation, as well as to cross-reference citations with other literature and extract tables and figures based on captions and table text. The Archaeotools Project has employed NLP across a range of archaeological texts.

Figure 1.2 shows the process architecture adopted for the Archaeotools Project. In brief, selected fields are extracted from the ADS Oracle database in MIDAS XML format data and converted to a Resource Description Framework (RDF) format. XML (OWL) versions of the thesaurus are

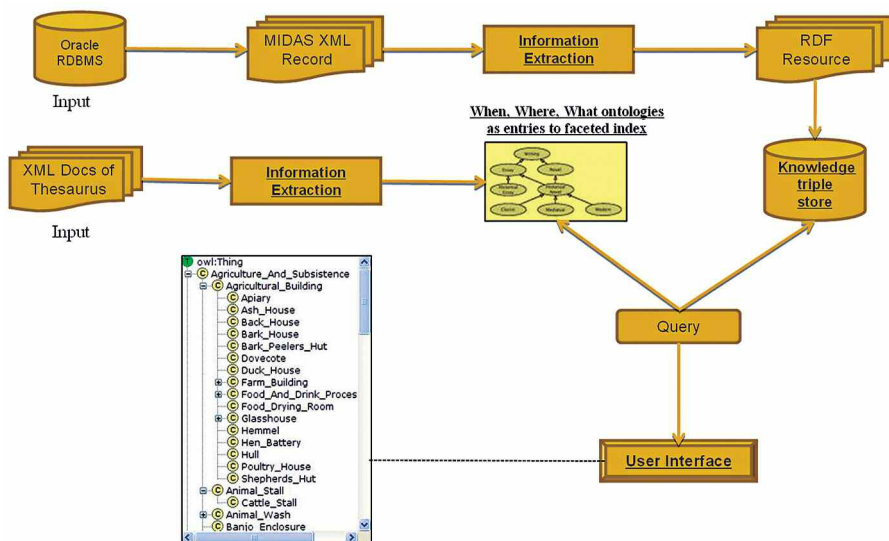


Figure 1.2. Archaeotools process architecture.

<sup>19</sup> <http://cidoc.ics.forth.gr/>

<sup>20</sup> <http://en.wikipedia.org/wiki/CiteSeer>

extracted to create workable ontologies, and these in tandem with the **RDF knowledge triple store** are queried to classify the records.<sup>21</sup>

Information extraction (IE) is the process of automatically extracting structured information from unstructured natural language texts (Cowie and Wilks 2000). One of the processes key to IE is the application of NLP technologies which enable computers to discern semantic meaning in natural languages. The outputs of NLP are linguistic data that are crucial to IE tasks, including sentence boundaries, part-of-speech tags, and grammar parsing. Conversely, IE usually requires human input to define (via templates) the general form of the information to be extracted; these templates then guide the extraction process.

Typical IE tasks include:

- ▶ *Terminology extraction*—identification of relevant terms for a given corpus—for example, identifying the most relevant terms for an archaeological corpus or data set, such as the ADS gray literature holdings.
- ▶ *Named-entity recognition (NER)*—identification of entities in a document, such as archaeological period terms, parish names, district names, archaeological findings, and so on.
- ▶ *Fact extraction*—identification of facts, which could be statements of relationships between entities—for example, these might link each identified archaeological findspot to identified parish names, thus constructing a relationship of the form “artifact-found-at-place.”

The Archaeotools IE tasks fall under NER and fact extraction. The first objective was to extract the following types of information units from a corpus of over 1,000 unstructured archaeological gray literature reports, such that this corpus could be indexed and searched by a number of attributes, including the predefined facets:

- ▶ Subject (topics covered, findings mentioned)—mapped to the “What” facet
- ▶ Location (place-names related to events and findings)—mapped to the “Where” facet
- ▶ Temporal (temporal information related to findings)—mapped to the “When” facet

---

<sup>21</sup> <http://www.w3.org/TR/owl-features/>

- ▶ Grid reference—mapped to the “Where” facet
- ▶ Report title, creator, publisher, publisher contact, publication date
- ▶ Event dates
- ▶ Bibliography and references

## APPROACHES TO BUILDING IE SYSTEMS

There are two basic approaches to the design of IE systems, the knowledge engineering approach (KE) and the automatic training (AT) approach (Appelt and Israel 1999). In the KE approach, an IE expert and a domain expert manually read through a moderate-size domain corpus. The domain expert identifies information units to be extracted, and the IE expert generalizes the textual patterns and translates those patterns into formal rules usable in programming. Next, they apply the rules to several corpora, examine extracted information to see where the rules undergenerate and overgenerate results, and revise the rules accordingly. It is obvious that the IE expert’s skills play a critical role in building working systems. An example from Archaeotools might be an IE system for extracting information about the publishers of archaeological reports, where a sample rule can be as simple as “the first organization that appears following the report title, and is a registered name on the Institute for Archaeologists list.” A disadvantage of this approach is that, in this example, the rule will not work for any unregistered organizations.

In the AT approach, it is not necessary to have detailed knowledge of IE systems and rule formalism. On the contrary, the most difficult rule-induction process is handled by the machine. Typically, domain experts are required to produce adequate volumes of sample annotations—usually a subset of the entire corpus—which are tagged to mark expected information units to “train” the IE system; and then they specify features that are likely to distinguish these sample annotations from unannotated sections of documents. Examples of features could be text units, generic entity types (person, organization, location, etc.), existence in gazetteers or dictionaries, position in the document, and so on. Next, an IE algorithm is run on the training corpus, consuming the selected features and producing a model that stores generalized rules to be applied to novel texts.

For example, Figure 1.3 displays a snapshot of a sample training document with three types of annotations: topic (red), creator (blue), and publication date (green). We can see that to extract creators using a KE approach, the key indicators (or “features”) are the first preceding word (“by”), its location relative to the title and publication date, and whether the word is a person’s name. However, it is difficult to manually build a rule to take all possible occurrences into account because the creator may not always be preceded by the word “by,” and its relative location may vary from document to document. Therefore, the better solution is to use the AT approach by providing the system with these example annotations, specifying possible features, and leaving the feature consumption and rule induction to machines. To this end, the ADS staff, all of whom are archaeologically trained, carried out extensive annotation exercises on a subset (about 150 reports) of the gray literature corpus.

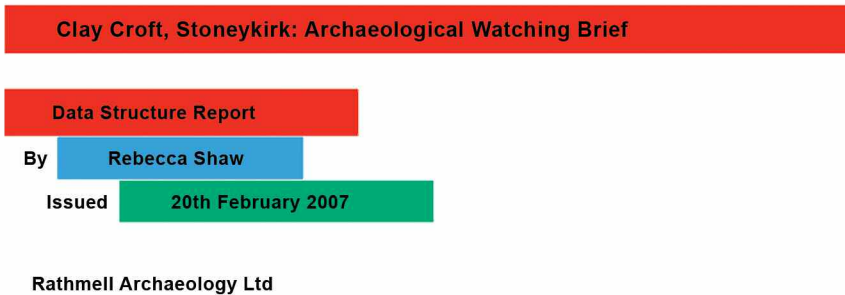


Figure 1.3. Example annotations from a gray literature report showing title, author, and published date.

## COMPARISON OF THE TWO APPROACHES

The advantage of the KE over the AT approach is that there is no need to prepare training data, which, in the case of Archaeotools, has proved to be quite time-consuming and is generally a laborious and tedious task. In situations where information occurs in regular and limited patterns and contexts, it is easy to develop systems that perform well. However, the KE approach itself requires an extensive amount of manual input. Exporting systems to different domains (e.g., from engineering to archaeology) is difficult, as rules are often context- and domain-specific and thus exporting usually requires a

system rebuild. Furthermore, when information to be extracted is very diverse (as are artifact finds in archaeology) and may occur anywhere and within any context in a document, the task can become extremely difficult. In contrast, the AT approach has better domain portability. Exporting IE systems to a different domain is comparatively straightforward, only requiring the rebuilding of a training corpus and feature tuning, and it costs far less than rebuilding an IE system. Also, the AT system handles diversity well and can be applied to large-scale data sets. The main drawback of the AT approach is that sometimes training data can be expensive to build (being time-consuming). On the other hand, feature selection is at least as important as the learning algorithm for a system that performs well, but in many cases, feature tuning can be difficult and time-consuming.

Taking into account the above comparison, both approaches have been employed in Archaeotools IE tasks, depending on the types of information being worked on. The KE approach is applied to information that matches simple patterns, or occurs in regular contexts, such as NGRs and bibliographies; the AT approach is applied to information that occurs in irregular contexts and cannot be captured by simple rules, such as place-names, temporal information, event dates, and subjects. Both approaches have been combined to identify report title, creator, publisher, publication dates, and publisher contacts.

## INFORMATION EXTRACTION APPLIED TO GRAY LITERATURE

Relatively high levels of success were achieved when the above techniques were applied to the sample of 1,000 semi-structured gray literature reports. After removing files that could not be converted to machine-readable documents due to file formatting issues, a working sample of 906 reports remained.

The greatest problem encountered was that of distinguishing between “actual” and “reference” terms. In addition to the “actual” place-name referring to the location of the archaeological intervention, most gray literature reports also refer to comparative information from other sites, here called “reference” terms. The information extraction software returned all place-names in the document, masking the place-name for the actual site among large numbers of other names. However, this was solved by adopting the simple rule that the primary place-name would appear within the “summary”

section of the report. If it was not possible to identify a summary, then the first 10 percent of the document was used instead.

Out of 960 reports in the working sample, there were 162 documents where it was not possible to identify a place-name in the summary or first 10 percent of the report (Table 1.2). However, there were only 17 left unidentified according to the “What” facet, 20 with no “Where” information, and 40 where it was impossible to identify a “When” term (Table 1.3).

Table 1.2 “Actual” identifications for 906 gray literature reports

FACET	NO DATA	PERCENT
What	159	17.5%
Where	162	17.9%
When	263	29.0%

Table 1.3 “Reference” identifications for 906 gray literature reports

FACET	NO DATA	PERCENT
What	17	1.9%
Where	20	2.2%
When	40	4.4%

Although these figures do not guarantee that the terms identified were meaningful, as long as users are shown why a document has been classified according to those terms, they represent acceptable levels of classification. Each document returned in response to a query is presented alongside the “tag cloud” reflecting the terms used to index it, with the font size of each term representing its frequency of occurrence within the document. We have coined the term “tagstract” to describe these visual representations of the content of a report, to reflect that they are hybrids of conventional abstracts combined with Web 2.0 tag clouds. They allow users to assess the relevance of the search returns to their query and to readily dismiss false positives. Figure 1.4 illustrates the “tagstract” for a gray literature report for an archaeological evaluation of a site in the English county of Northamptonshire.



# INFORMATION EXTRACTION APPLIED TO HISTORIC LITERATURE

A sample of original *PSAS* text is relatively unstructured and uses non-standard terms:

49

The bronze ring inscribed with Runic characters, presented to the Society, was found in the year 1849, in the Abbey Park, in the immediate neighborhood of St Andrews. It is a large Bronze Finger-Ring inscribed on the two faces in Anglo-Saxon Runes, and is of peculiar interest, as being, it is believed, only example of the Palaeography (sic) of our Anglo-Saxon forefathers hitherto found in Scotland, with the single, but most important exception of the noble monument at Ruthwell, Dumfriesshire (Wilson 1851: 23).

Using NLP, the following data can potentially be extracted from it.

- ▶ What—Bronze Ring, Runic Inscription
- ▶ Where—Abbey Park, St Andrews (not Ruthwell)
- ▶ When—Anglo-Saxon (found 1849)
- ▶ Who—Wilson, D.
- ▶ Media—PSAS (PDF)

We were surprised to find that, despite the highly unstructured nature of the text and the antiquated use of language, once trained on the gray literature reports, the IE software achieved comparable levels of success with the antiquarian literature. Problems were encountered with more synthetic papers and other types of document, but where the primary subject of the article was a fieldwork report, it was possible to identify the key “What” “When,” and “Where” index terms using the same approach adopted for the gray literature.

After discounting prefatory papers, such as financial accounts or election reports, the *PSAS* corpus was reduced to 3,991 papers referring to archaeological discoveries. By applying the rule that the actual What, Where, and When would appear in the first 10 percent of the paper, it was possible to identify a subject term for all but 277 of the papers (Table 1.4), although there was less success with a geospatial location (627 papers with no location), and the least success with period terms (2,056 papers with no “When” term). However, these results could be improved somewhat by looking at the “Reference” terms. Although less certain to provide the primary identification of the key “What,” “Where,” and “When” for each paper, these left far fewer papers unclassified (Table 1.5).

Determining place-names within the CDP thesaurus proved a challenge, particularly given the number of historic names used in older accounts, but

Table 1.4. “Actual” identifications for 3,991 *PSAS* papers

FACET	NO DATA	PERCENT
What	277	6.9%
Where	627	15.7%
When	2,056	51.5%

Table 1.5 “Reference” identifications for 3991 *PSAS* papers

FACET	NO DATA	PERCENT
What	123	3.1%
Where	238	6.0%
When	1049	26.3%

the geo-gazetteer web service hosted by EDINA at the University of Edinburgh was used to resolve many of the outstanding names. Extracted place-names were sent directly to this service, and the GeoXwalk automatically returned NGRs for the place-name (centered) or, in the case of some urban areas, an actual polygon definition. This allowed the relevant place-names from *PSAS* to be mapped in the Archaeotools geospatial interface and therefore made them as discoverable and searchable as standard monument inventory data sets.

Of the total of 3,991 *PSAS* papers, it was initially impossible to find an Ordnance Survey grid reference for 3,388 (85 percent), compared with a figure of just 185 (20 percent) for the gray literature. This reflects the fact that older reports did not tend to use precise geospatial references to refer to site or find locations. However, by using the GeoXwalk service, it was possible to resolve a place-name into a grid reference for all but 268 reports (6.7 percent); of these, there was no “Where” term for 238 reports, leaving just 30 for which a place-name had been identified that could not be georeferenced by the EDINA web service. Manual checking revealed that the majority of these were instances where a county name was the most precise spatial location used in the published paper. This result demonstrates the power of the GeoXwalk service to assign geospatial coordinates to antiquarian accounts of archaeological discoveries.

The analysis of the *PSAS* also provided some tantalizing glimpses into how IE tools might allow us to chart the development of more controlled and standardized vocabulary, as well as to investigate changes in archaeological research interests over time (Bateman and Jeffrey 2011).

Clearly, this type of extracted data meshes perfectly with the faceted browsing interface discussed earlier. Therefore, it is possible to aggregate resource discovery metadata relating to the antiquarian accounts of fieldwork published in the *PSAS* directly with the other data sets that have been made searchable in this way.

## CONCLUSION

The Archaeotools Project has achieved its objective of successfully implementing a faceted classification browsing system in the context of aggregated archaeological records. This service has been released for public access as a replacement for the existing ArchSearch II interface. It is now embedded within the search interface available to **ADS users**,<sup>23</sup> and its long-term success will be judged by how effectively users are able to undertake their research. It has built on previous projects such as the Archaeobrowser and, in turn, will provide the foundation for further developments facilitating effective cross-searching within and between ADS data archives.

During the process of preparing the data sets for classification, useful insights have been gained into the level of vocabulary control within archaeological monument inventories. Although work has had to be done regarding the apparent mismatch between the seemingly loose terminology of the historical data sets and the rigorous word lists, thesauri, and ontologies, in practice a combination of automated and manual approaches allowed for the classification process to be both comprehensive and meaningful. The classification and data-cleansing process itself can be seen essentially as a single operation, as existing records are rarely changed. There may be over one million records in the data sets, but these are added to at a fairly slow rate (about 5,000 per annum), meaning that future mismatches or missing facets are much more likely to be in small and manageable volumes. This has also given the ADS the unexpected benefit of being able to report back to donor organizations, not just on the level of the data cleansing required, but also on the specific records and problematic fields encountered.

---

<sup>23</sup> <http://archaeologydataservice.ac.uk/archsearch/>

It is within the sphere of data cleansing that Web 2.0 technologies may have some potential value for the Archaeotools Project. While machine-based tagging may achieve high levels of accuracy, inevitably there will be some errors, and here it may be possible to harness special-interest groups to carry out additional indexing. While the faceted browser described here operates according to three predefined facets, the original Archaeobrowser system also demonstrated the potential for user-defined trails within the data sets, and four exemplars were created, covering “War and remembrance,” “Landscapes of salvation,” “Learning and labor,” and “Ages of migration.” The Open Context system also incorporates the facility for user-generated facets (see Chapter 2), and although this might require the distinction to be made between “official” validated tags and user-defined tags, this approach has some merit in allowing the development of new research themes. When the Archaeotools Project began, the technologies associated with Web 2.0 were growing in popularity and project developers envisaged using special-interest groups to amend errors and clean data. On further reflection, the implications for validating and archiving such changes became worrisome, particularly since the data sets were often “owned” by other bodies. In addition, it became clear that user-tagging can create interesting questions about performance, since re-indexing an item consumes a lot of computational resources. A balance needs to be struck between allowing users to add their own tags, and re-indexing all resources according to those tags to generate a new facet tree.

Web 2.0 technologies may also bring benefits to the ADS catalog via web services that allow others to integrate the results of distributed searches within their own interfaces. Within the **ARENA2 Project**,<sup>24</sup> the ADS is developing a distributed web service architecture with a number of European partners. However, this raises the issue of how far state heritage agencies will be willing to allow unauthenticated users to create “mash-ups” that incorporate their data. The issue here is not so much the potential load on the server, but the incorporation of archaeological data in applications that take the data out of their institutional context and into websites where the primary audience might be treasure hunters, for instance.

Nonetheless, the Archaeotools Project has demonstrated the potential for applying automated data and metadata extraction to both gray literature and legacy literature. The combined attack on the data using KE and

---

<sup>24</sup> <http://archaeologydataservice.ac.uk/research/arena2>

AT information extraction has already proved successful to the extent that automated resource discovery metadata extraction can confidently be achieved for gray literature as well as for published antiquarian accounts of fieldwork. This removes a major obstacle to digitizing the huge backlog of this material, which, we hope, will help to unlock its potential to significantly influence the development of archaeological theories in the future. Archaeological researchers will always depend on a range of historic journal articles, contemporary textual reports, and structured data. The Archaeotools Project fits into the larger picture of scholarly communication by providing an effective means of indexing, combining, and browsing a heterogeneous range of resources within a single cyberinfrastructure. Like a more conventional library, it provides a dynamic resource which will continue to grow as more documents are added. And, as is the case with a library, users need to be able to trust the completeness, accuracy, and integrity of the classification and information retrieval systems. This is partly based on the reputation and branding of ADS as the host (the “publisher”), but it also relies on the authority of the suppliers of texts and data (the “authors”) and on the clear documentation of the source of all data. It is also essential that the information retrieval systems are transparent and not prone to giving higher rankings to specific resources—for commercial gain, for example. The “tagstract” reveals precisely why a particular result has been returned.

The success of the Archaeotools Project and the ability to index records within a hierarchical facet tree have depended on the prior existence of standardized hierarchical thesauri. The United Kingdom is relatively well provided with such word lists, particularly because of the well-established work of the English Heritage Data Standards Unit, the former Museum Documentation Association, and other bodies. Other parts of the world, including the United States, do not have such well-developed standards initiatives. Nonetheless, in these cases, automated techniques of information extraction might usefully be explored in order to generate ontologies, allowing similar benefits to be attained.

## REFERENCES CITED

- Amrani, A., V. Abajian, Y. Kodratoff, and O. Matte-Tailliez  
 2008 A Chain of Text-Mining to Extract Information in Archaeology. *Information and Communication Technologies: From Theory to Applications, 2008. ICTTA 2008, 3rd International Conference*, pp. 1–5.

- Appelt, D. E., and D. Israel  
1999 *Introduction to Information Extraction Technology*. IJCAI-99 Tutorial, Stockholm. Retrived from <http://www.ai.sri.com/~appelt/ie-tutorial/IJCAI99.pdf>
- Bateman, J., and S. Jeffey  
2011 What Matters about the Monument: Reconstructing Historical Classification. *Internet Archaeology* 29. Retrieved from [http://intarch.ac.uk/journal/issue29/bateman\\_index.html](http://intarch.ac.uk/journal/issue29/bateman_index.html).
- Binding, C., D. Tudhop, and K. May  
2008 Semantic Interoperability in Archaeological Datasets: Data Mapping and Extraction via the CIDOC CRM. *Proceedings (ECDL 2008), 12th European Conference on Research and Advanced Technology for Digital Libraries, Aarhus*, pp. 280–290.
- Bradley, R.  
2006 Bridging the Two Cultures. Commercial Archaeology and the Study of Pre-historic Britain. *Antiquaries Journal* 86: 1–13.
- Byrne, K. F., and E. Klein  
2010 Automatic Extraction of Archaeological Events from Text. In *Making History Interactive. Proceedings of the 37th Computer Application in Archaeology Conference, Williamsburg 2009*, ed. B. Frischer, J. W. Crawford, and D. Koller, pp. 48–56. Oxford: Archaeopress.
- Cowie, J., and Y. Wilks  
2000 Information Extraction. In *Handbook of Natural Language Processing*, ed. R. Dale, H. Moisl, and H. L. Somers, pp. 241–260. Boca Raton: CRC Press.
- Denton, W.  
2003 How to Make a Faceted Classification and Put It on the Web. Retrieved from <http://www.miskatonic.org/library/facet-web-howto.html>
- Falkingham, G.  
2005 A Whiter Shade of Grey: A New Approach to Archaeological Grey Literature Using the XML Version of the TEI Guidelines. *Internet Archaeology* 17. Retrieved from <http://intarch.ac.uk/journal/issue17/5/index.html>
- Giles, C. L., K. D. Bollacker, and S. Lawrence  
1998 CiteSeer: An Automatic Citation Indexing System. In *Proceedings of the Third ACM Conference on Digital Libraries (DL '98)*, ed. I. Witten, R. Akscyn, and F. M. Shipman III. New York: ACM. 89–98. DOI=10.1145/276675.276685.
- Hardman, C., and J. D. Richards  
2003 OASIS: Dealing with the Digital Revolution. In *CAA2002: The Digital Heritage of Archaeology. Computer Applications and Quantitative Methods in Archaeology 2002*, ed. M. Doerr and A. Sarris, pp. 325–328. Athens: Archive of Monuments and Publication,s Hellenic Ministry of Culture.
- Jeffrey, S., W. Kilbride, S. Waller, and J. D. Richards  
2008 Thinking Outside the Search Box: The Common Information Environment and Archaeobrowser. In *Layers of Perception. Proceedings of the 35th*

- International Conference on Computer Applications and Quantitative Methods in Archaeology (CAA) Berlin, Germany, April 2–6, 2007*, ed. A. Posluschny, K. Lambers, and I. Herzog, pp. 206–211. Kolloquien zur Vor- und Frühgeschichte Band 10. Bonn: Habelt.
- Jeffrey, S., J. D. Richards, F. Ciravegna, S. Waller, S. Chapman, and Z. Zhang  
2009 The Archaeotools Project: Faceted Classification and Natural Language Processing in an Archaeological Context. In *Crossing Boundaries: Computational Science, E-Science and Global E-Infrastructures*, ed. P. Coveney, pp. 2507–2519. Special themed issue of the *Philosophical Transactions of the Royal Society A*, 367.
- Forthcoming When Ontology and Reality Collide: The Archaeotools Project, Faceted Classification and Natural Language Processing in an Archaeological Context. In *CAA 2008, Proceedings of the CAA Conference, Budapest*.
- Lock, G.  
2008 A Professional Mockery. *British Archaeology* 101: 36–37.
- Paijmans, H., and S. Wubben  
2008 Preparing Archaeological Reports for Intelligent Retrieval. In *Layers of Perception. Proceedings of the 35th International Conference on Computer Applications and Quantitative Methods in Archaeology (CAA) Berlin, Germany, April 2–6, 2007*, ed. A. Posluschny, K. Lambers, and I. Herzog, pp. 212–217. Kolloquien zur Vor- und Frühgeschichte Band 10. Bonn: Habelt.
- Richards, J. D., and C. S. Hardman  
2008 Stepping Back from the Trench Edge. An Archaeological Perspective on the Development of Standards for Recording and Publication. In *The Virtual Representation of the Past*, ed. M. Greengrass and L. Hughes, pp. 101–112. London: Ashgate.
- Robinson, B.  
2007 From Sites and Monuments Records to Historic Environment Records: From Planning to Research. Unpublished Ph.D. thesis, University of York. [http://libcatalogue.york.ac.uk/F/?func=direct&doc\\_number=001628305](http://libcatalogue.york.ac.uk/F/?func=direct&doc_number=001628305)
- Ross, K. A., A. Janevski, and J. Stoyanovich  
2005 A Faceted Query Engine Applied to Archaeology. *Proceedings of the 31st International Conference on Very Large Data Bases*, pp. 1334–1337. Trondheim: ACM.
- 2007 A Faceted Query Engine Applied to Archaeology. *Internet Archaeology* 21. Retrieved from <http://intarch.ac.uk/journal/issue21/3/index.html>
- Uren, V., P. Cimiano, J. Iria, S. Handschuh, M. Vargas-Vera, E. Motta, and F. Ciravegna  
2006 Semantic Annotation for Knowledge Management: Requirements and a Survey of the State of the Art. *Journal of Web Semantics* 4: 14–28.
- Wilson, D.  
1851 Inscribed Runic Ring. *Proceedings of the Society of Antiquaries of Scotland* 1: 23–25.



## CHAPTER 2

# TOWARD A DO-IT-YOURSELF CYBERINFRASTRUCTURE: OPEN DATA, INCENTIVES, AND REDUCING COSTS AND COMPLEXITIES OF DATA SHARING

*Eric C. Kansa and Sarah Whitcher Kansa*

## THE CHALLENGE OF DISPARATE DATA

Field research generates a dizzying amount of information, painstakingly gathered by teams of people over many years of a project. Technological advances in the past two decades have escalated the quantity and range of information gathered in field studies. Now, in theory (and sometimes in practice), every moment of fieldwork can be documented, using a combination of media, including cameras, video recorders, laser scanners, geographic information systems (GIS), and global positioning systems (GPS). Back in the lab, specialists add further documentation with scanned drawings, digitized maps, spreadsheets of analytical notations, and measurements. The continuing decline in storage costs and the growing sophistication of database systems help fuel the drive for more complete and thorough recording. In most cases, these new approaches produce richer and more comprehensive documentation than was previously feasible with traditional paper and photographic recording techniques. But actually using that content requires sophisticated analytic tools and software to facilitate information retrieval and to summarize mountains of data.

This is especially the case with structured data, the content of databases and spreadsheets. Although researchers sometimes seek individual items of

structured data, more often they want to interrogate structured data for overall patterns. For success at this level of interrogation, they need access to disparate data from a project or multiple projects, along with generalized information about the project to understand the fundamentals of methodologies and meaning.

This desire to precisely interrogate large sets of content motivates researchers to develop and work with structured data. Systems publishing this kind of content should keep these motivations in mind. However, the analytic capabilities that people demand from structured data impose challenging and expensive requirements for Web-based dissemination systems. Researchers need to judge how, and to what extent, they should standardize and codify their documentation. This is no easy task, because the institutional setting of archaeology works against greater formalism in documentation. Researchers in the humanities and social sciences typically work in decentralized institutional settings within different research traditions. Time and budgetary constraints further inhibit the development of widely adopted recording and data management standards. Therefore, scholars generally lack consensus on standards of recording and tend to make their own customized databases to suit their individual research agendas and theoretical perspectives (see also Denning 2003; Hodder 1999; Boast et al. 2007).

Furthermore, the size and complexity of archaeological databases challenge even expert metadata (“information about information”) documentation. Archaeological databases may include hundreds of thousands of individual records created by multidisciplinary teams, in complex interrelationships. If data must be downloaded and deployed on appropriate software, they may still be difficult to use even with adequate documentation. Once deployed, the data are so complex that users must familiarize themselves with a project’s database organization and interface to make use of the information. Downloading and deploying such databases requires too much effort for casual browsing and searching. Thus, making data sets available for download (even with adequate metadata) is not an ideal solution for archaeological communication. Without some generalized search and analysis tools, downloadable data is hard to “digest” by others.

A better solution is to serve archaeological databases on dynamic, online websites, to make content easy to browse and explore. Because this typically requires complex and expensive custom web development, only a handful of well-funded projects offer dynamic access to databases of primary results via

the Web. The enormous [Çatalhöyük database](http://www.catalhoyuk.com/database/catal/)<sup>1</sup> is a good example of project-specific data sharing. Its extensive catalog of excavated contexts and finds facilitates analysis and collaboration among the project's large team of specialists. While this is a fundamental contribution to scholarship, Çatalhöyük's system is not readily scalable; if other projects seek to adopt Çatalhöyük's online database to share content, they would have to conform to its recording system.

Most archaeological projects take place in smaller research programs with less funding and technical support than Çatalhöyük. These smaller projects have little capacity to develop customized, Web-accessible database solutions. They may develop rich bodies of documentation, but without Web dissemination much material will never see publication because the vast amount of content cannot be accommodated by print. Therefore, thousands of bones, seeds, potsherds, lithics, and other artifacts and ecofacts that are analyzed and recorded, as well as the maps, photos, and log entries associated with a typical project, almost never see publication.

In the end, field researchers who complete a project, no matter how they organized their data, find it challenging to make their raw data available in an integrated and intelligible fashion.<sup>2</sup> Other priorities usually win out, so that few researchers invest the effort and expense required for Web dissemination. The harsh reality of time and money constraints means Web publishing must return clear and tangible value to researchers. Finding the right "carrots" and "sticks" (to paraphrase Julian Richards) will be required for more general participation in data sharing.

## OPEN CONTEXT: SIMPLE TOOLS FOR COMPLEX CONTENT

With these challenges in mind, we developed Open Context, a free, Web-based data publishing tool providing access to primary data from multiple projects. The system can manage structured data, narrative texts, and other media (images, video, GIS, etc.). It employs a "bottom-up" approach, with simple, easily adaptable services and tools to suit diverse users. User flexibility in visualizing the content allows tuning presentation and visualization styles to meet different needs. Openness in sharing machine-readable data

---

<sup>1</sup> <http://www.catalhoyuk.com/database/catal/>

<sup>2</sup> See the review of the International Dunhuang Project in the September 2008 *CSA Newsletter* (<http://csanet.org/newsletter/fall08/nlf0802.html>) for a discussion of the costs associated with a customized access system.

helps avoid the “one-size-fits-all” trap of locking content into one mode of presentation, visualization, and interpretation. Instead, a high-quality user experience *requires openness* and the ability for users to draw on the capabilities of data sources and visualization and analysis services now present on the Web.

Open Context is a common portal for browsing, simple “Google-like” searches, complex Boolean queries, data summary, data export, and tagging of several pooled data sets (Table 2.1). Content ranges from archaeological field projects to geological science, public health, and zoological data sets. The largest data set served by Open Context includes some fifteen years of field investigations conducted by Brown University in Jordan. Open Context’s flexibility stems from 20 years of database design development and field-testing by David Schloen, lead of the University of Chicago OCHRE system (Schloen 2001). While OCHRE provides sophisticated data management tools targeted for active research projects, Open Context uses a subset of the OCHRE data structure described by the Archaeological Markup Language (ArchaeoML) to streamline Web-based access and organization of di-

Table 2.1. Links to examples from Open Context discussed in the text

EXAMPLE WEB RESOURCE	LINK
Open Context home page	<a href="http://opencontext.org">http://opencontext.org</a>
Faceted search	<a href="http://opencontext.org/sets">http://opencontext.org/sets</a>
Item details	<a href="http://opencontext.org/subjects/30C3F340-5D14-497A-B9D0-7A0DA2C019F1">http://opencontext.org/subjects/30C3F340-5D14-497A-B9D0-7A0DA2C019F1</a>
Table output	<a href="http://opencontext.org/tables/39fd14fe7196aea0821ce8c7e08251f8">http://opencontext.org/tables/39fd14fe7196aea0821ce8c7e08251f8</a>
Project overview	<a href="http://opencontext.org/projects/CBB6B9F7-500C-4DDD-71AA-4D5E5B96CDBB">http://opencontext.org/projects/CBB6B9F7-500C-4DDD-71AA-4D5E5B96CDBB</a>
Media	<a href="http://opencontext.org/media/009AB9BB-EB46-427E-26B7-A9C29BFAA859">http://opencontext.org/media/009AB9BB-EB46-427E-26B7-A9C29BFAA859</a>
Web services documentation	<a href="http://opencontext.org/about/services">http://opencontext.org/about/services</a>
Publishing documentation	<a href="http://opencontext.org/about/publishing">http://opencontext.org/about/publishing</a>

verse field research content. The global data schemas in Open Context accommodate field data sets without predetermined standard vocabularies or recording systems. This flexibility enables a high degree of interoperability and efficient retrieval without requiring adherence to highly specified standards (Kansa 2005).

During the system's development, we addressed some challenges researchers face regarding the concept and practice of data sharing by focusing on the following:

- ▶ System development works under an iterative design process, taking lessons from users (and potential users) to address their needs, workflows, data collection processes, and professional incentives. As a result, we continually adapt Open Context. User feedback led to enhancements in search and discovery features (especially faceted search; see Figure 2.1 and discussion below) and greater emphasis on editorial processes and metadata quality.
- ▶ Open Context content is structured in a way that allows “mash-ups,” or recombinations and visualization of multiple media mixed with highly structured data. Again, the development of these capabilities is partially driven by user demand for specific features, especially features to improve Open Context's use as an instructional aid.
- ▶ Open Context's data structure is generalized to allow for organizing diverse data sets without predetermined recording systems. Contributors need not conform to overly specified standards or change their research design or language. Open Context removes the need for projects to develop specialized software to display their research results (see example in Figure 2.2). This significantly reduces the cost of data dissemination, thus increasing the likelihood that content will be shared.

## MAKING SENSE OF DISPARATE DATA

Because Open Context sees iterative development, the project experiments with different models for user interaction. Some of these experiments fail, and understanding such failures leads to changes in direction and development priorities. Open Context's early experiments with Web 2.0 techniques represent illustrative examples of reevaluation and reprioritization.



Web-based research data publication

BETA

Home About Projects Browse Lightbox Tables Details My Account New Search GO

Click on one of the links below to further limit your view of items in Open Context.

Or, type a search term in the box below (works like Google):

Search this Set

- Context
  - Select Multiple
    - [Hazor](#) 4134
    - [Domuztepe](#) 3913
    - [Pinarbasi](#) 1226
    - [Petra Great Temple](#) 1108
    - [Khirbat al-Mudayna al-Ahva](#) 217
- Project
  - Select Multiple
    - New Projects
      - [Khirbat al-Mudayna al-Ahva](#) 217
    - Featured Projects
      - [Domuztepe Excavations](#) 3913
      - [Petra Great Temple Excavations](#) 1108
    - Other Projects
      - [Hazor](#)
      - [Zooarchaeology](#) 4134
      - [Pinarbasi 1994: Animal Bones](#) 1226
- Category
  - [Animal Bone](#) 10598
- User Tag
  - [goat metapodials](#) 45
- Tag Creator
  - [Eric C. Kansa](#) 45
- Descriptive Variable
  - [Taxon](#) 10598
  - [Element](#) 10596
  - [Side](#) 5961
  - [Sex](#) 4227
  - [Stratum](#) 4122
  - [Element Certainty](#) 3913
  - [Pathology Noted](#) 3913
  - [Taxon Certainty](#) 3913
  - [Fragment Present](#) 3566
  - [Symmetry](#) 3029
  - [More...](#)
- Related People
  - Select Multiple
    - [Justin Lev-Toy](#) 4351
    - [Sarah Whitche Kansa](#) 3913
    - [Denise Carruthers](#) 1226

Open Context currently has 10598 items, filtered by the following criteria:

- (1) Contained in: [Jordan OR Turkey OR Israel OR Palestinian Authority](#)
- (2) Category: [Animal Bone](#)
- (3) Variable 'Taxon': [With value: sheep/goat OR Ovis aries / Capra hircus OR Ovis sp. OR Ovis aries/Capra hircus OR Ovis aries OR Capra hircus OR Ovis/Capra OR goat OR sheep](#)

Last Updated: August 26, 2010, 1:22 am

Results: items 1 to 10 out of 10598 items

	<b>Petra Great Temple Excavations</b>	<b>Animal Bone</b> Bone GT-0625 Context: Jordan / Petra Great Temple / Upper Temenos / Special Project 85 / Locus 19 / Seq. SP85128	
	<b>Petra Great Temple Excavations</b>	<b>Animal Bone</b> Bone GT-2142 Context: Jordan / Petra Great Temple / Temple / Trench 15 Part II / Locus 106 / Seq. 15273	
	<b>Petra Great Temple Excavations</b>	<b>Animal Bone</b> Bone GT-2223 Context: Jordan / Petra Great Temple / Upper Temenos / Trench 84 and Special Project 91 / Locus 18 / Seq. 84157	
	<b>Petra Great Temple Excavations</b>	<b>Animal Bone</b> Bone GT-2172 Context: Jordan / Petra Great Temple / Upper Temenos / Trench 84 and Special Project 91 / Locus 18 / Seq. 84174	
	<b>Petra Great Temple Excavations</b>	<b>Animal Bone</b> Bone GT-1943 Context: Jordan / Petra Great Temple / Upper Temenos / Trench 77 Part II / Locus 9 / Seq. 77160	
	<b>Petra Great Temple Excavations</b>	<b>Animal Bone</b> Bone GT-2003 Context: Jordan / Petra Great Temple / Lower Temenos / Trench 71 / Locus 19 / Seq. 71389	

Figure 2.1. Open Context's "faceted browse" tool, showing filtered results.



Web-based research data publication



[Home](#)
[About](#)
[Projects](#)
[Browse](#)
[Lightbox](#)
[Tables](#)
[Details](#)
[My Account](#)

[GO](#)



**Project / Collection:** Dove Mountain Groundstone

**Description:** Analysis of groundstone finds from the Dove Mountain Project in the Tucson Basin

Number of Views: 408

### Project / Collection Overview

The Dove Mountain project was conducted by [Desert Archaeology, Inc.](#) during two seasons in 2001 and 2005. Deborah Swartz was Project Director and the editor of the published volume (Swartz 2008). The project area covers 87m2 in the upper bajada of the Tortolita Mountains, which form part of the northern boundary of the Tucson Basin. Drainages cross the area flowing from northeast to southwest, the largest of which are Ruelas, Wild Burro, and Cochise canyons. These canyons are narrow near the source and widen into broad, sandy washes as they leave the mountains and drain into the Santa Cruz River. All three canyons contain semipermanent water. Recorded springs are upstream from the sites in these three canyons.

Most of the sites in the Dove Mountain parcel are situated along these three drainages, including all the current project sites. These sites lie near the juncture of the steep foothills with the more gradual bajada slopes. The foothills are covered with granite boulders associated with bedrock outcrops, and it is in these locations that petroglyphs were found. The remaining sites lie in the more gently sloping bajada areas on Pleistocene alluvial fans with well-entrenched washes below the fan surfaces. Elevations of the project sites range from 2,780 ft to 2,900 ft (847-884 m) above sea level (asl) in Wild Burro Canyon and from 3,420 ft to 3490 ft (1,042-1,064 m) asl in Ruelas Canyon. This assemblage from the Dove Mountain sites is useful for assessing various theories about how bajada settlements fit into the regional settlement system of the Tucson Basin, as proposed in the project research design. One theory that has been proposed is that the bajada settlements were locations of limited activity related to seasonal resource procurement for more permanently inhabited settlements in the floodplain of the major Tucson Basin drainages (Roth 1995). A second theory is that these were permanent settlements, complete with all the activities that commonly occurred year-round. Ground stone assemblages recovered from limited-activity sites are expected to be much different from assemblages of continuous occupations. Differences in the assemblages are reflected in artifact design strategy, wear amounts, use strategies, and activity diversity.

Ground stone artifacts analyzed for this chapter were collected from six excavated sites: Atlatl Ridge, AZ AA:12.84 (ASM); Desert Tortoise, AZ AA:12.83 (ASM); Wild Burro Canyon, AZ AA:12.170 (ASM); Ruelas Canyon, AZ AA:12.785 (ASM); AZ AA:12.783 (ASM); and AZ AA:12.787 (ASM) (Appendix Table D.1). Atlatl Ridge was partially excavated in 1985 and 1987, with results published in 1995 (Roth 1995). During a testing phase of the Dove Mountain project, Atlatl Ridge was again partially excavated, in addition to Desert Tortoise, Wild Burro Canyon, and 13 other sites (Vint 2000). Ground stone items recovered during the testing phase are described elsewhere (Adams and Silva 2000); however, those from the testing phase at Atlatl Ridge, Desert Tortoise, and Wild Burro Canyon are included here.

As a result of the Desert Archaeology excavations at Dove Mountain, there is now a clearer picture of ground stone technology development in the Tortolita bajadas, which contributes to a better understanding of the development of grinding technology in the greater Tucson Basin. The use of the bajadas was probably significantly more complex than previously realized, and the ground stone assemblage reflects this complexity.

**Publications:**

Adams, Jenny L.

2008 "Grinding Technology in the Tortolita Mountain Bajadas," pp. 307-342 in D. L. Swartz, ed., *Life in the Foothills: Archaeological Investigations in the Tortolita Mountains of Southern Arizona*, Anthropological Papers No. 46. Center for Desert Archaeology, Tucson.

### Browse this Project

	<a href="#">Desert Tortoise</a>	173 items contained in this context
	<a href="#">Wild Burro Canyon</a>	63 items contained in this context
	<a href="#">Ruelas Canyon</a>	39 items contained in this context
	<a href="#">Atlatl Ridge</a>	13 items contained in this context
	<a href="#">AA:12.783</a>	2 items contained in this context
	<a href="#">AA:12.787</a>	2 items contained in this context

### Keywords for this Project

*Southwest, Pioneer Period, Sedentary Period, Hohokam, Early Agricultural, Groundstone, Arizona, Archaeology*

### Tags Used in this Project (0)

Items from this project/collection have been tagged by: 0 people.

### Linked Media (1 files)

	<a href="#">Dove Mountain Groundstone</a>
---	---

Figure 2.2. Example of a project overview in Open Context.

## User-Generated Tags

Each item in Open Context contains contextual and descriptive information and can be linked to other items by the contributor or through user-generated “tags.” Users can tag single items or groups of items for their own use; or, they may share their tags publicly, permitting colleagues to build on their

research. This facilitates application of new terms and relationships, even across items from different projects. This tagging, done under editorial oversight, helps organize content according to any desired conceptual framework.

Recent studies have shown that folksonomies, or collaborative categorization of content such as tagging, can develop metadata for collections with content of sufficient quality to meet many professional needs (Trant 2006). Used initially for museum collections, folksonomies are now featured in professional knowledge management applications (McAfee 2006). A common problem in small-scale, field-based sciences is “meta-analysis,” where researchers compare disparate data sets from different projects. However, different research designs and their resulting data sets (even if accessible) are often incompatible (see the ecology example in Steward et al. 2007). The ability to add user-defined metadata to content can be an important research tool, to identify reports where field observations were collected with compatible methodologies. Such determinations are essential to valid comparisons and syntheses across multiple studies. Finally, since determining whether methodologies and data sets are comparable is an interpretive judgment, such decisions should be contestable (Kansa 2005). User-generated tags have the advantage of being suggestions, as opposed to a definitive, final word. However, tagging has not been a widely used feature of Open Context for reasons discussed in more detail below.

## Uptake Issues with User-Generated Content

Not many users have participated in Open Context’s tagging, most likely because the system is already well documented in metadata, making tagging less necessary. Besides Open Context, there are few other archaeological data sets available on the Web and few other relevant resources can be tagged. In addition, the design architectures of many systems, including online collections of many major museums, do not support tagging well. Many of these systems lack stable URIs<sup>3</sup> to specific items. These design flaws break tagging systems.

---

<sup>3</sup> A URI is a uniform resource identifier, which identifies resources on the Internet. URLs (uniform resource locators) are subsets of URIs that specify a resource’s location and method of retrieval. On the Web, URLs begin with *http://* and are more colloquially referred to as “hyperlinks.” We use the term URI in this chapter because it is the preferred term in discussions of “Linked Data.”



The 200 users now registered with Open Context represent a user community that is too small to sustain sufficient tagging activity. In many Web 2.0 collaborative systems, only a small fraction of users actively participate and add value to the collective resource. This is also true for professional applications of Web 2.0 approaches, where, according to a report by the consulting firm McKinsey, only 3 to 6 percent of users provide 75 percent of all user-generated content in enterprise settings (Bughin 2007). The five (!) users who have tagged in Open Context neatly fit this uptake pattern. Interestingly, searches of “open context” and “opencontext” in the social bookmarking site Delicious show 90+ tagged pages relating to Open Context. Current development of Open Context will do away with the present user account system for tagging and replace it with more widely used systems, including [OpenID](http://openid.net/)<sup>4</sup> and Google logins. In addition, the [Delicious](http://www.delicious.com/)<sup>5</sup> tagging API will enable anyone with a Delicious account to use that account in tagging material in Open Context.

While the tagging features have low usage at the present moment, there is very little need for users to sign up for an account and log in to Open Context. Open Context offers all data and web services free of charge with no login required. Thus, though Open Context experiences low participation in tagging, it sees a high amount of passive use. According to server statistics, Open Context has approximately 4,500 unique visitors every month, with an average of five pages per visit. The majority of visits come from direct lookups or bookmarks (50 percent), followed by search engines (30 percent). Ten percent of visitors follow links from other websites to Open Context, especially from Wikipedia. The frequency of search engines bringing visitors to Open Context illustrates their central role in information retrieval. These data are corroborated by an informal survey of archaeological user needs. Google ranked highly as an information retrieval tool used by professionals.

To understand search engine use, in early 2008 we conducted a three-month study of search strings linked to Open Context content (Figure 2.3). The results were a classic “long tail” graph, where 47 percent of searches were for a person, place, or object; and the remaining 53 percent comprised diverse requests for other kinds of information. This highlights the importance of having Google-type search capabilities and also of being able to track searches to see what users want.

---

<sup>4</sup> <http://openid.net/>

<sup>5</sup> <http://www.delicious.com/>

**Followed links to Open Context content  
based on 3,600 searches, January–March 2008**

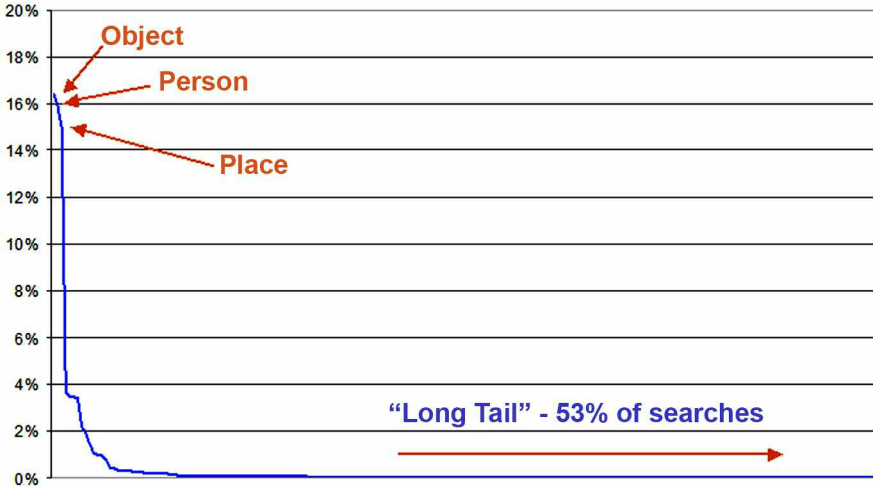


Figure 2.3. Overview of search strings that linked to Open Context content over a three-month period.

Finally, the limited uptake of Open Context’s tagging system should be regarded as preliminary. In most peer production environments, participation follows a “power law,” which means that a few very active users dominate, supplying the most contributions (Wilkinson 2008), and so expanding the user base to attract committed “super contributors” is critical to success. This need takes us back to the relevancy issue. Without enough relevant content to satisfy the majority of users, social and collaborative tools will likely fail. Since the job of building a critical mass of relevant content is so large, it must be distributed among many projects sharing common frameworks for interoperability. Thus, Open Context design priorities have shifted away from features typical of “destination websites” that need active participation of many users. Instead, current developments enhance Open Context’s capabilities to work with distributed systems on the Web. Instead of expecting users to visit Open Context, we are doing more to bring Open Context data to places where users already work. These measures to enhance data portability are explored below.

## UNDERSTANDING USERS OF DIGITAL DATA

The above discussion of user tagging in Open Context helps illustrate the value of learning from users and adapting development strategies to better meet user needs. Understanding what users want should be the key concern driving investments in data-sharing systems. By focusing on the users, developers gain an understanding of the suite of issues that impact uptake—not only technological, but also the more ambiguous social, political, environmental, and workflow parameters in which users work (see also Harley et al. 2010). Open Context demonstrates and validates a *potentially* interesting and useful approach toward data integration. We need a better understanding of users, their workflows, and data needs to make the right investments in future development.

For a data resource to succeed, users must find it appealing, credible, easy to use, and relevant for meeting specific goals. These relate to the “user experience,” a person’s overall impression when using a product or system. The concept of user experience “places the end-user at the focal point of design and development efforts, as opposed to the system, its applications or its aesthetic value alone” (Rubinoff 2004). How a user perceives a system will depend on the user’s specific needs and experience with multiple facets of that system, such as content, branding, functionality, the user community, and the context of using the system. Online businesses and computer games often invest heavily in elaborate (and sometimes expensive) programs to improve user experience. In contrast, academically focused systems are typically grant-funded and noncommercial and do not benefit from the large user design investments common in commercial services. Thus, user experiences with humanities computing systems tend to be “poor,” with clunky websites, tools that are difficult to use, and systems that integrate poorly with familiar workflow patterns (Juola 2008; Warwick et al. 2008).

Open Context takes an iterative approach to development: build the tool, release it, see how it is used, revise it, release it, and so on. This iterative approach, focusing on user needs and constant revision, may address some challenges in sharing archaeological research content. Systems need constant adapting because serving small, highly specialized communities that work with structured data represents a moving target. Expectations and incentives change as user communities become more comfortable with online data sharing (see Chapter 7) and as granting foundations make increasing demands for

information transparency. While iterative, user-centered design is an important strategy, it cannot be divorced from overall systems architecture. Some fundamental technical design decisions must be made early in development in order to facilitate later iterative development. Systems based on sound architectural models, especially with an eye for extensibility, make for easier adaptability to meet changing user needs. In the case of Open Context, the use of a generalized data model (ArchaeoML), and a design emphasis that makes data portable and easy to repurpose, facilitate more responsive approaches to meeting user needs.

## Faceted Navigation

As discussed above, Open Context has gone through several iterations and revisions. Instead of emphasizing social tagging, it now features enhanced information retrieval and data portability features. The first beta version of Open Context was launched in April 2006. Even with little promotion and community building in the first year, the site has seen fairly active use. Server logs and Google Analytics help us identify how people enter and use Open Context, and we receive email feedback from users requesting feature changes. It is clear that few users employ the Advanced Search option, and, as discussed, most want simple search or browse tools to find people, places, and objects. This research led us to develop user-friendly search and navigation tools as well as search functions to find records associated with specific individuals.

Searching and retrieving information from large bodies of complex data is a challenge for many digital libraries and information services. Keyword searches are a common solution, but they can yield incomplete and ambiguous results. This uncertainty is problematic for professional research, since the “hit-or-miss” nature of keyword searches adds uncertainty to information retrieval. That uncertainty limits the effectiveness of information repositories for regulatory and scientific work. For example, the spottiness of gray literature and primary data retrieval limits the scientific reliability of syntheses attempting to review multiple reports (Roberts et al. 2006).

To avoid some of the difficulties associated with keyword searches, Open Context uses a faceted navigation system, which allows users to explore hierarchically structured metadata with “point and click” selections and progressively home in on more specific information from a larger collection (Figure

2.1). Because filters are applied across an entire collection, users can be certain that their results will be more comprehensive than with keyword searches. Feedback, in the form of subtotals of the numbers of items in each facet, helps guide user selection of additional filters. This helps users understand the size and composition of the collection they are searching (Hearst 2006). The Archaeotools Project (Richards et al., this volume, Chapter 1; Jeffrey et al. 2009) represents another high-profile application of faceted navigation in archaeology.

Faceted navigation is based on organizing collections using a common data structure. Types of facets available for navigation depend on the data structure. The Open Context data structure offers fine-grained control and flexibility in information retrieval, allowing faceted filtering of content based on project- or collection-level metadata. In addition, Open Context represents each project and collection data set in the same way, allowing facets based on the contextual and descriptive properties of items within each project and collection (Figure 2.1). In other words, Open Context's facet navigation enables users to discover and filter records of individual items (sites, contexts, finds, media, documents, etc.) contained within diverse projects. Thus, Open Context's faceted navigation allows seamless and simultaneous navigation not only *within* individual projects and collections, but *between* them as well.

## Data Sharing as Publication

Most researchers who publish data with Open Context express great interest in linking data dissemination with print publications. This fits with the traditional approach to publication in the archaeological community and offers a reason for scholars to share primary data, by having it repeatedly referenced from the synthesized, printed publication. With unique URIs for each item, citations to Open Context can be made from the printed text, making Open Context an extensive digital appendix with unlimited page numbers. This is not the most exciting or interactive use of the system, but it is one that serves researchers well. Open Context works in conjunction with print publication by enhancing printed syntheses and providing a place where primary field data can be housed, accessed, and commented on. This gives a new level of transparency to research and could help raise the bar for how scholarship is judged and valued by the community. Such transparency,

rather than threatening scholars, can provide testimony to the quality of their research. Sharing research openly demonstrates that a scholar's contributions are supported by a solid foundation of high-quality evidence and documentation. Over time, we believe that publication without full disclosure of underlying evidence will no longer be considered sufficient.

Until then, the greatest barrier to digital dissemination of primary data may be the current lack of reference implementations—that is, projects with research outcomes demonstrating the value of sharing primary content. Scholars only invest time and resources in organizing and sharing primary content if there is a clear and significant reward. That reward may be different for different scholars: some seek increased visibility of their research offered by digital dissemination; others seek access to data from other projects; and still others see making primary data available as a responsibility of the discipline.

The benefits of these motivations have yet to be clearly demonstrated. Nevertheless, recent policy changes at the National Science Foundation (NSF) suggest that data sharing is becoming a more mainstream expectation of research. The NSF now requires grant-seekers to provide a Data Management Plan. While these plans do not necessarily require open access data, they do require researchers to explain and justify dissemination and archiving plans. To help guide researchers in how they may meet this new requirement, the NSF archaeology program now links to Open Context and **Digital Antiquity's tDAR** system<sup>6</sup> (see McManamon and Kintigh 2010). It will be interesting to observe how this new requirement affects the uptake of systems like Open Context. Conversely, will the example of Open Context help set expectations for open access, open licensing, and data portability? To help educate the research community on the importance of legal and technical openness as well as data quality and documentation standards, Open Context provides NSF grant-seekers a Web-based form. This form generates language tailored for a researcher's project that can be used for inserting into NSF Data Access Plans. Over time, we hope that NSF's new requirement and Open Context's guidance will help improve the quality and comprehensiveness of archaeological data, its dissemination and longevity.

---

<sup>6</sup> <http://www.tdar.org/>

## Measuring Impact and Use

While server statistics can be useful, the best indicator of success for data-sharing projects like Open Context will come from measuring the impact of publications that cite data retrieved through the system. However, such subsequent analyses and publications will take time and will occur only years after a project's data are published online. To attempt to gauge the usefulness of published digital resources more rapidly, other less direct metrics must be established. Currently, Open Context offers the following metrics for estimating impact:

- ▶ *View counts:* Each item in Open Context is related to at least one specific individual. Each item also belongs to a project or collection. Open Context counts the number of times visitors retrieve content associated with each individual data contributor and project.
- ▶ *Social use:* Similarly, Open Context counts the numbers of tags attached to the content associated with each individual and each project. The assumption here is that tagging frequency is a proxy measure for interest. The more frequently content is tagged, the more interesting it may be to the broader community.
- ▶ *Widgets:* Contributors to Open Context can install special widgets to their online curriculum vitae (perhaps hosted on their departmental website) that display current web metrics for the content they published in Open Context.

These web metrics are proxy measures of research impact. Web metrics from other disciplines with more mature Web-based publishing platforms show fairly strong positive correlations between web metrics and citation (Vaughan and Shaw 2005; Brody et al. 2006; Piwowar et al. 2007). Thus, while we cannot yet examine citation impacts of data publications, web metrics may offer useful interim data to understand researcher uptake and impact.

## Citation and Licensing

Open Context users and contributors strongly demand clear citation features. Open Context uses Dublin Core metadata to generate bibliographic citation information, along with a stable URI, for each item (or a tagged set

of items) in its database (Figure 2.4). Using *Zotero*,<sup>7</sup> a citation and bibliographic management tool, investigators can capture bibliographic information associated with Open Context materials. This clear and easy citation makes Open Context very useful for scholarly applications.

The citation framework also integrates *Creative Commons* licensing information.<sup>8</sup> Creative Commons licenses allow users to freely and legally use material if they properly attribute the original creator. They include machine-readable RDF metadata from commercial search engines such as Yahoo and Google (Kansa 2005, 2007), making Open Context's openly licensed content available to web searches. Such openness ensures that Open Context content is of maximum value for reuse in instructional and research applications. While openness to reuse is arguably, in many cases, an ethical imperative, such openness is not universally appropriate. Intellectual property issues take on added importance and complexity when one looks beyond professional research circles. Various segments of the public, especially indigenous and descendant communities, may have strong claims about the past (Hollowell and Nicholas 2008). Here the application and perceived benefit of Creative Commons licenses are more problematic (Christen 2005; Kansa et al. 2005; Kansa 2009). As discussed below, standards, including standard Creative Commons copyright licenses, are not politically and culturally neutral. Such standards express and help reinforce particular world views and agendas (Bowker and Star 2000; Boast et al. 2007). Diverse ideas and concerns over knowledge privacy and custodianship shared among some indigenous peoples may map poorly to these standard licenses. Thus, Open Context's approach of open access and open licensing is clearly *not* a universally appropriate model for all archaeological data.

These ethical challenges make clear it that one-size-fits-all solutions will not work for archaeology. Open Context has a large but limited scope, especially in terms of ethics and cultural property considerations. Other systems need to be developed and deployed to meet needs that cannot be met by Open Context. Diverse solutions, technologies, intellectual property frameworks, and organizational participants must be encouraged in digital archaeology.

---

<sup>7</sup> <http://www.zotero.org>

<sup>8</sup> <http://creativecommons.org>



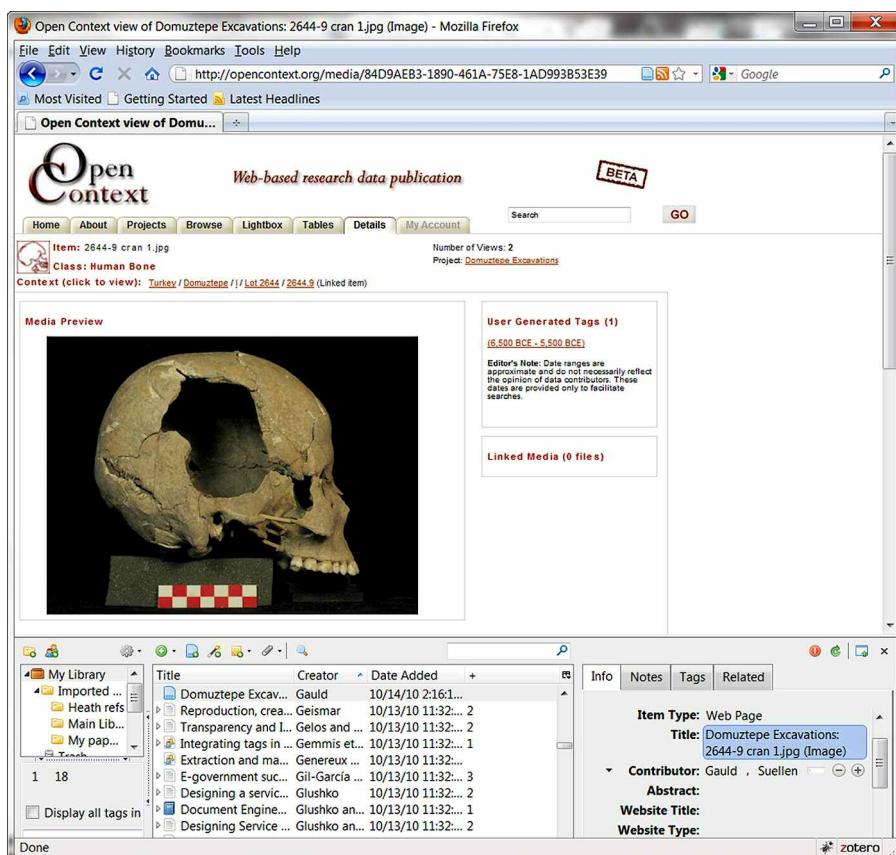


Figure 2.4. Open Context uses Dublin Core metadata to generate bibliographic citation information and a stable URI for each item (or tagged set of items).

## MAKING DATA PORTABLE

Some important user needs go beyond interactions through web browsers. Many of the “users” most interested in working with Open Context like to develop their own information systems for their own audiences. Supporting the needs of distributed systems thus marks an important development priority for Open Context. In the emerging ecosystem of disciplinary data-

sharing systems, Open Context will coexist among a growing array of both specialized and more general tools and resources. Technologies, data-sharing needs, and applications are all rapidly evolving. It is unlikely that any individual system can provide for all users and every application. Furthermore, the speed of change within this domain implies that systems, tools, and technologies will be superseded and discarded. Indeed, even successful web companies can quickly fade away when the next “killer app” emerges. It would be foolhardy to assume that Open Context (or any one system, for that matter) represents *the* future of archaeological data. However, we must continue to steward irreplaceable archaeological data, even as such systems evolve or disappear.

Given the risks and volatility of the Web, many researchers worry about using the medium as a stable foundation for scholarly inquiry. How can a researcher study online material and know the hyperlinked references will be available in two months, much less two years or decades hence? There are no such guarantees with online resources. Of course, no scholarly resource is necessarily permanent; libraries *do* discard materials. However, the volatility of the Web is especially acute.

To mitigate the risks and volatility of the Web, Open Context’s content is as portable as possible and is integrated with as many data longevity programs as possible. Portability improves the content’s chance of discovery and preservation. However, to achieve portability, the content must be open. Openness, in terms of technological standards (file formats, service interfaces) and intellectual property (licensing terms), is the key to Open Context’s strategy for longevity.

## Data Portability for Preservation

The content from Open Context can be moved and incorporated into another system with no technical or legal barriers. So, users and contributors to Open Context are free to use the best and most trusted tools and resources available, whether or not Open Context meets their needs. Each item of Open Context content has its own unique identifier in a common XML standard (the Archaeological Markup Language), so data can be retrieved from static backup copies if URIs fail to resolve.

Digital curation and archiving requires institutional commitments far beyond what a small project like Open Context can manage. Institutional

continuity as well as dedicated professional expertise must underlie data-preservation processes. In other words, data preservation is as much a social and institutional concern as it is a technical matter. While small projects like Open Context can take certain measures—such as use of open and widely used file-formats, adequate metadata documentation, and open licensing of content to facilitate data preservation—data preservation requires the sustained commitment of institutions that have long-term experience with archiving digital resources. Typically, this means that data archiving requires the support of major universities, libraries, or government agencies.

While Open Context is not itself an archive, it does benefit from data archiving and curation services provided by the University of California's California Digital Library (CDL). As in many areas, emerging distributed services and infrastructure can help data-sharing efforts like Open Context meet their challenges. The CDL offers a host of data-preservation and curation services as part of the library's participation in the National Science Foundation's DataNet initiative. These include:

- ▶ *Minting and binding of ARKs (archival resource keys):* ARKs are special identifiers managed by an institutional repository. The CDL will help ensure that the objects associated with these identifiers can be retrieved in the future, even if access protocols such as "HTTP" change.
- ▶ *Data archiving:* The CDL also provides data curation and stewardship to maintain integrity of digital data and to migrate data into new computing environments as required.

The CDL provides the services outlined above as Web-based "micro-services." The same sorts of Web 2.0 "mash-up" technologies for sharing machine-readable data now help support data archiving and curation processes. For example, the CDL now uses Open Context Atom feeds as a manifest of web resources needing archiving. Thus, the same web services that Open Context provides to facilitate access now facilitate data longevity. These services give Open Context a strong institutional foundation for citation and data archiving.

While the CDL is now Open Context's primary provider of data archiving services, all Open Context content can be copied and archived in its entirety into other organizational settings. For example, Zotero is currently developing Zotero Commons as a stable, institutionally backed repository for web content cited by researchers using the Zotero tool. Initially, Zotero

Commons will be supported by the Internet Archive, but it can also receive support from other institutional digital archive programs, including **DSpace**<sup>9</sup> and **Fedora**.<sup>10</sup> Automatic copying and safekeeping of content helps “crowd-source” the work by identifying material worth digital preservation. In the long term, Zotero Commons shows that openness and usefulness are important to digital longevity. Material that is accessible, open, and useful to researchers is more likely to get digital longevity support.

Citation features, backed by permanent identity services provided by the CDL, help situate Open Context in the larger context of scholarly communications. Citation can be a powerful motivation for organizations and individuals to openly publish data. In some domains, publication of data can enhance the impact of associated papers (Piwowar et al. 2007). Each record in Open Context is associated with the names and even organizational affiliations of responsible analysts. Such links to related publications will likely be one of the most important reasons future researchers will have for publishing data in Open Context. We anticipate that linked publications will also provide one of the richest and best sources of “metadata” to document content in Open Context.

Beyond making Open Context content easier to understand and more meaningful, citations and links to scholarly content elsewhere on the Web will make Open Context data easier to discover. As search engines expand their coverage, they will become increasingly good at discovering links between Open Context and other scholarly publishing venues. As more people reference Open Context, the more search-engine “Page rankings” (see Brin and Page 1998) will improve, further elevating the visibility of research in the system. Open Context therefore represents a good solution for people and organizations, now struggling in some obscurity, to gain more exposure for their professional output. Thus, through Open Context, researchers can publish in a way that builds their public visibility. These “carrots” of positive incentives have helped to motivate several organizations and researchers to publish with Open Context.

---

<sup>9</sup> <http://www.dspace.org/>

<sup>10</sup> <http://fedoraproject.org/>

## Data Portability for Reuse

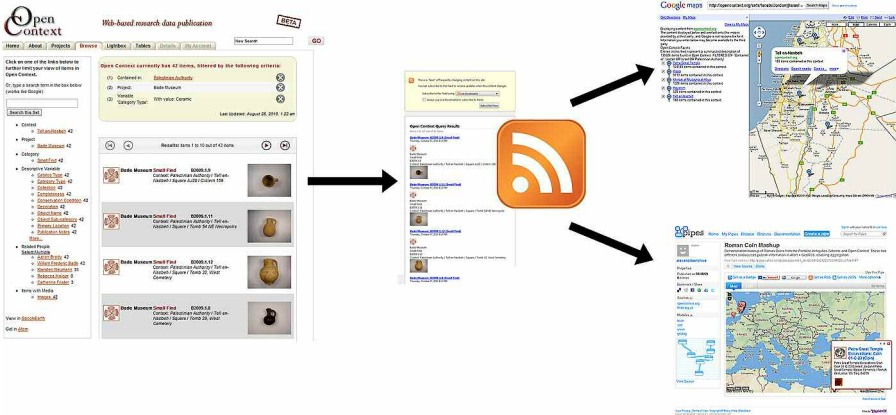
Data portability helps ensure that content can be discovered and is preserved, and is also a key aspect of making data *useful*. As discussed above, Open Context cannot meet every user need. Users have different requirements for content, analyses, and data presentation. Users may want to compare a collection in Open Context with results from another project. They may want to visualize Open Context data in a geospatial viewer like Google Earth, or incorporate Open Context data in a 3D virtual world. Just as data portability facilitates use of emerging data archiving and preservation infrastructure, data portability also encourages innovative approaches to visualization and analysis beyond the capabilities of Open Context itself.

Advances in standards and practices of data portability may be lasting and important contributions of the Web 2.0 era. Many Web 2.0 services include machine interfaces called “web services” or “application program interfaces” (APIs). A web service is a system offering information in a format easy to parse and process by machines (typically XML or JSON). This information is available through requests using HTTP, the Web’s communication protocol. These services allow mash-ups, or combinations of content, processing, and visualization capabilities from different sources. They also let other websites and applications use Open Context content and analytic capabilities. For example, the Badè Museum website uses Open Context’s API to show Badè content hosted by Open Context (Figure 2.5b).

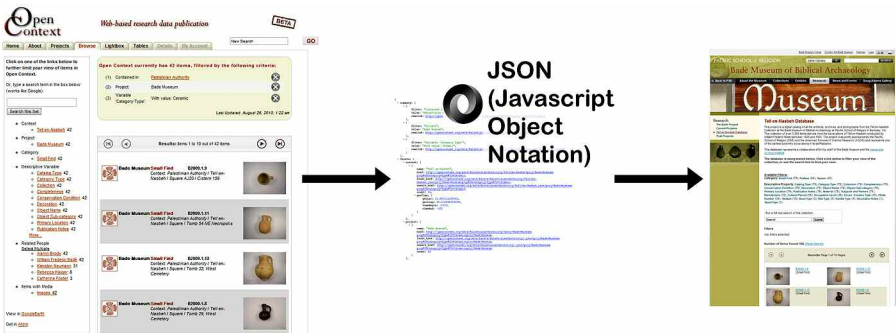
Recent developments to Open Context focus on machine interfaces, aimed at reducing development costs and improving capability for “mash-ups” (see examples in Figure 2.5a, b, c). An important web development involves REST architecture. REST (Representational State Transfer) means data, including query results or machine-readable expressions of data, are retrievable as resources at specific Web addresses. The World Wide Web uses this idea of resources bound to URIs, removing the need for elaborate messages to get information. You get the data needed in a format convenient for reuse and processing just by following a hyperlink. The hyperlink shown for Query A is a simple query to retrieve a specific item:

Query A:

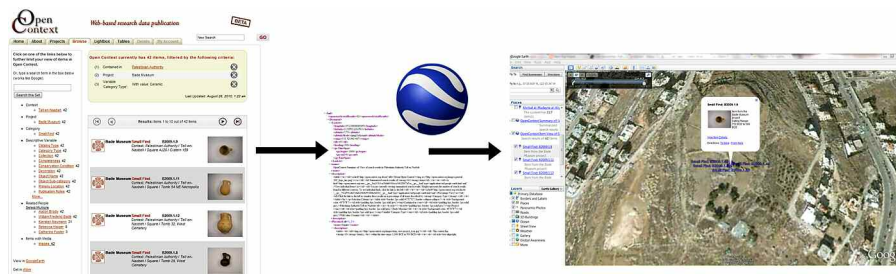
<http://opencontext.org/subjects/DC8F1345-5AEC-4455-FDFF-F335CEB2987D>



(a) Open Context Atom feed used in a Google Maps visualization and aggregation of ancient coins data with the Portable Antiquities Scheme.



(b) JSON (JavaScript Object Notation) service used to show Badè Museum collections hosted in Open Context on the Badè Museum website.



(c) Open Context KML service to visualize Badè Museum collections in Google Earth.

Figure 2.5. Illustrations of RESTful services in Open Context.

This link queries Open Context and creates a resource (web page) providing a record of an artifact from the Petra Great Temple. Sets of items can also be retrieved. REST architecture allows data requests in desired formats just by specifying a hyperlink.<sup>11</sup> For example:

Query B:

<http://opencontext.org/sets/?bBox=30,30,90,90&cat=Small+Find>

This link returns a resource (web page) generated by a query in Open Context for “small finds” between 30E, 30N to 90E, 90N. At the time this paper was written, 86,543 items were identified. The link below returns a different resource:

Query C:

<http://opencontext.org/sets/?cat=Small+Find&bBox=30,30,90,90&t-start=200&t-end=360>

This resource is a web page generated by a query similar to Query B, but limited to items dated between 200 BCE and 360 CE. So, REST architecture enables linking individual items and groups of items. In all of these cases, queries were executed by specifying a web address. In Open Context, the default return is a human-readable web page. However, Open Context can also return other formats easier to process by different software applications. For example:

Query B-2:

<http://opencontext.org/sets/.kml?bBox=30,30,90,90&cat=Small+Find>

This link returns the same results as Query B, but in KML format, the form of XML used by Google Earth. Data from Query B-2 can be used by Google Earth or other such tools, providing a different visualization tool than a standard web browser (see examples in Figures 2.5c, 2.6).

REST architecture is often used to retrieve data in formats convenient for software processing, such as Atom Syndication Format, a widely used XML vocabulary. Atom supersedes various RSS (“Really Simple Syndication”) formats and is one of the most widely used, extensible, and extended standards on the Web. Atom and REST are used in Open Context in several ways:

---

<sup>11</sup> Of course, there is no need for a person to specify the links. Software often automatically specifies hyperlinks to request required data.



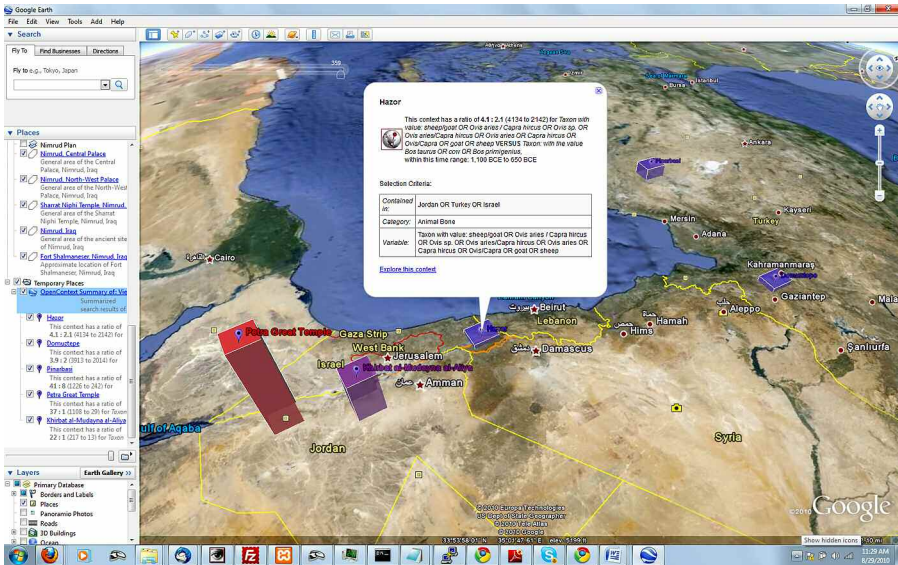


Figure 2.6. Web services demonstration using Google Earth to visualize Open Context KML data, showing species ratios at different sites in the Near East.

1. *Atom search result feeds*: Open Context uses Atom to broadcast results of queries in a widely used, convenient format. Results of searches in Open Context, expressed as Atom feeds, can be used by other web applications (Figure 2.5a). They can be aggregated with other data sources using feed-processing tools like **Yahoo Pipes**.<sup>12</sup> Since Atom support is ubiquitous in code libraries, an Atom feed of query results can be imported into other applications, such as a feed reader.
2. *Atom-based “facets feeds”*: Open Context offers faceted browsing/faceted search underpinning a user interface and a web service to express content in Atom format (Figure 2.1). In addition to syndicating individual search results in an Atom feed, as described above, the faceted search tool produces another Atom feed syndicating a summary of the faceted search results. This “facets-feed” syndicates data structures that can be queried along with the number of results matching criteria defined by each structure. The facets-feed machine-readable syndication provides

<sup>12</sup> <http://pipes.yahoo.com/pipes/>



an overall summary of a query result set, allowing Open Context to describe how it can be queried in a simple, machine-readable format.

Atom can be used for everything from syndicating news stories published by Reuters to sharing archaeological data. It is a standard container allowing complex information to be shared and managed with widely available feed-management infrastructure. It can carry custom “payloads” of XML content. The XML content can be ArchaeoML (the XML vocabulary used by Open Context) or other vocabularies used by other applications. For instance, Google publishes Atom feeds for many Google applications using its own XML vocabulary, GData. The additional XML payloads do not invalidate an Atom feed. Using Atom, resources can be shared to enable communication of complex data-structures (such as an XML representation of archaeological artifacts), while ensuring the recipient will gain some basic utility from the content even if the content cannot be fully processed by that recipient. Atom gives recipients the ability to use XML data in increasing levels of sophistication, based on their need and capacity. By enabling different levels of sophistication in adoption, Atom is a good solution for archaeology, a field always struggling with deficiencies in financial and technical support. If researchers routinely share data in Atom feeds, interdisciplinary data exchanges are easier, since outside systems supporting other research communities can obtain some value even from content they cannot fully process. Using formats like Atom, Open Context data can be aggregated with other relevant data sources, such as *Pleiades*,<sup>13</sup> a major gazetteer project for the Classical Mediterranean world. *Pleiades* has also pioneered similar RESTful architectural and service design patterns (Elliot and Gillies 2009). Even though Open Context and *Pleiades* do not share common discipline-specific standards like ArchaeoML, through Atom they gain at least some level of interoperability (see also Kansa and Bissell 2010).

## LEVERAGING THE POWER OF THE HYPERLINK

The above discussion may be fairly technical for this venue; however, these architecture issues are important for developing web systems to meet the needs of archaeological researchers. REST architecture is an important element of interoperability and usability. Specific web addresses enable linking

---

<sup>13</sup> <http://pleiades.stoa.org>

and referencing of resources (records of individual artifacts, contexts, or sites; or expressions of sets of items defined and limited by query parameters). The addresses are hyperlinks that can be bookmarked, emailed to colleagues, or used in software to run mash-ups. Search-engine spiders can follow such links and index content in systems using REST (Heath 2010; Isaksen 2008). Search-engine discovery is important to Open Context use and to the impact of research more generally. Thus, very sophisticated applications can be powered by an inherently very simple and elegant concept: the hyperlink.

REST architecture is not (yet) common in many scientific cyberinfrastructure and digital library systems that often rely on more complex messaging protocols for exchanging information. The architectural style of these earlier systems originated in the closed and controlled world of mainframe computing typical of large institutional settings, including commercial enterprises. The popular acronym SOA (“service oriented architecture”) usually implies enterprise-style architectures relying on complex messaging using elaborate (WS\*/SOAP-based) web service interfaces. They require custom software development to use, are difficult to maintain, and impose high development costs (Pierce et al. 2007). These approaches, with high barriers to entry, are not scalable. They do not lend themselves to adaptive and responsive design of new information services. Such adaptability is needed in research designs using data from multiple and often unanticipated sources.

When linkable data and retrieval services are combined with expressions of data in commonly used, machine-readable formats, the possibilities for data use greatly expand. Open Context expresses all data in human-readable web pages and in Atom feeds to maximize the reusability and customization potential of Open Context. With everything in convenient feed formats, different interfaces for different communities—for groups or for projects—can be developed quickly and cheaply. Each different user community, with its own custom-determined subset of Open Context content, could have its own portal with its own presentation style. Thus, customization need not be sacrificed in the name of interoperability. On the contrary, interoperability can make customization easier. In this scenario, communities can create their own portals, with their own identities and websites, drawing on Open Context data and data from other sources of information published in similar portable formats. In this way, REST architectures coupled with Atom (and

similar widely used formats) allow customization, reuse, and aggregation across sites.

This point about REST architecture and machine-readable data opening up web systems is important for long-term success of archaeological data sharing. It is unlikely that one system can meet all user needs, and archaeologists, like most research professionals, often have specialized niche interests. Open Context has several projects and large data sets, including the comprehensive results of excavations at Petra, a World Heritage site in Jordan, but it is far from housing content relevant to *most* archaeologists. Unless you are specifically interested in Nabataean archaeology (and a few other niche areas), the content in Open Context is mainly a curiosity. It will take tremendous effort to accumulate enough content so that most users will find information relevant to their own specialized interests most of the time. Building this “critical mass” of content will be much easier if we distribute the job and share resources, querying capabilities, and the like with other projects aiming to publish archaeological data. Sadly, this key issue of interoperability, an essential requirement for delivering relevant content to users, is in its infancy for archaeology.

## THE PLACE OF DIGITAL DISSEMINATION IN THE FUTURE OF ARCHAEOLOGICAL RESEARCH

Knowledge production, dissemination, learning, and research involve far more than accumulation of data and the transmission of bytes over the Internet. They are embedded in social practices and interactions among colleagues, researchers, and students and involve much more than exchange of data.

The social embeddedness of archaeology cannot easily be replicated or replaced by Web-based data dissemination. Much of the social interaction where people make sense of primary data involves tacit knowledge that is never fully expressed. Tacit knowledge, including unarticulated habits, assumptions, practices, and modes of behavior acquired by individuals (and even teams), are a challenge to “knowledge management systems” in general (Haas and Hansen 2007). This issue of tacit knowledge must be recognized by systems aiming to improve archaeological knowledge production. Tacit knowledge helps make archaeological data meaningful, and the use and usefulness of archaeological data dissemination will be partly determined by

tacit assumptions about how such data should be leveraged in research and teaching.

The importance of tacit knowledge highlights how the meaning and significance of archaeological data are socially embedded. In this regard, data sharing across projects must confront the tacit knowledge issue. Open Context does this by asking contributors to document their projects and define meaningful relationships within their data sets during data import and the editorial review process. Other systems may ask users to fill out forms providing metadata expressed in some community-wide standard. Such efforts attempt to document tacit assumptions and meanings of archaeological data in a machine-readable fashion.

As seen in many cases across disciplines, these approaches can be very valuable. Once articulated, tacit assumptions are easier to examine and critique. This encourages introspection and “reflexivity” in a discipline. However, we should recognize that data documentation and metadata description have limits. Such efforts never wholly capture tacit knowledge. Nuance, experience, and background always sit behind a data set, and these are difficult for data-sharing systems to capture. Reasons behind gaps in recording, the relative reliability of different observers, or even subtle shifts in use of database fields over the history of a project are all part of the back-story that may add important nuance to the meaning and interpretation of field documentation. In addition, tacit knowledge is often distributed across the various members of a given community and may not be wholly available to the person filling out Web-based metadata forms. Moreover, tacit knowledge changes and helps shape social relations. Some participants in a given project may lack the institutional memory to know why the database has certain fields, or even if those fields were used consistently in the project.

Translating tacit knowledge into some explicit standard is somewhat “lossy” (meaning, information gets lost). Nuance will be lost and transformed, and any metadata or documentation standard will not be totally comprehensive. Thus the meaning of archaeological data is a moving target, varying from group to group and subdiscipline to subdiscipline. This is a tremendous challenge for Semantic Web approaches, which attempt to formally describe concepts and relationships. Efforts to express assumptions of meaning in machine-readable formalisms will never be comprehensive and are always contingent. Research should consider which kinds of archaeological meaning are stable and widely shared across the discipline, and which

areas are more dynamic and particular to certain subdisciplines or schools of thought. Such an investigation may make application of Semantic Web approaches more appropriate.

Open Context makes some initial guesses about archaeological meaning so the system can articulate with widely held tacit assumptions about the nature of archaeological data. Open Context lays heavy emphasis on spatial context and observational classes to organize information in a way useful to many practitioners. However, Open Context has limited explicit semantic modeling and avoids many of the formalisms often associated with Semantic Web approaches. Open Context expresses explicit semantic relationships only at the general level of noting relationships (especially spatial containment relations) among items, and notes that some data is descriptive information. Beyond that and some general descriptive Dublin Core metadata, Open Context has little explicit semantic formalism.

Different Web-based systems can lie at different ends of a spectrum with regard to how much they invest in using Semantic Web formalisms to transform tacit knowledge into explicit knowledge. Though Open Context is much more formal than many Web 2.0 systems, it invests less in semantic formalisms than European systems do, especially those that model data according to the **CIDOC-CRM**,<sup>14</sup> a major cultural heritage ontology (Doerr and Iorizzo 2008). Since encoding more specific semantic relationships is costly, the effort must be justified by benefits in terms of research outcomes, data preservation, or some other value (Isaksen et al. 2009). In our initial judgment, a limited and less costly level of semantic formalism seems appropriate thus far. Open Context shows it is possible to build useful systems to manage and work with aggregated data by using a simple, generalized, underlying semantic model.

Nevertheless, we acknowledge that more sophisticated, explicit semantic modeling may be useful for certain future applications. Because every item and set of items in Open Context can be retrieved from a specific address (URI), it will be easy to index and reference Open Context content using more sophisticated semantic models. Open Context uses Atom feeds as a standards-based approach for sharing URIs of resources returned from queries. By sharing lists of URIs, other applications and systems can identify, retrieve, and annotate precisely defined subsets of Open Context content.

---

<sup>14</sup> <http://www.cidoc-crm.org/>

This opens the door for incremental semantic enrichment of Open Context resources (Kansa and Bissell 2010).

This architecture helps “future-proof” Open Context with regard to emerging new ontologies. Also, by emphasizing data portability over elaborate semantic modeling, we believe Open Context better accommodates the socially embedded nature of data. Cultural heritage has multiple meanings in multiple contexts. It is neither feasible nor advisable to attempt to map all cultural heritage to some single master ontology, even if that ontology has great conceptual elegance and sophistication (as is the case with the CIDOC-CRM). Any attempt at classifying and organizing cultural heritage would necessarily prioritize certain perspectives and world views that favor certain agendas (Boast et al. 2007; Bowker and Star 2000). Even within the research community, appropriate choice of ontologies may depend on one’s research needs, the nature of available data sets, and practical concerns (Kansa and Bissell 2010; see also Martinez and Isaksen 2010). Thus, it makes sense to make data portable so that it can be moved and incrementally enriched, as needed. This approach helps keep a human researcher “in the loop” to determine exactly how different data sets and subsets of data should be related. For many scientific applications, a researcher’s domain knowledge may play an important role in resolving ambiguities in a given data set or, by extension, multiple data sets (see Palmer and Craigin 2008). Data portability helps ensure that investigators have choice in selecting and applying the ontologies and data sets best suited for their research.

## CONCLUSIONS

The discussion of tacit knowledge touches on larger issues relating to archaeological theory and data sharing. Theoretical assumptions are implicit in systems designs and underlie deeper and more fundamental goals. One important assumption is that more information is better, and loss of that information is a problem requiring a solution. This is defensible from many perspectives. A scientifically oriented archaeologist, or one interested in critical “reflexive” approaches, sees benefit in greater information access and transparency to field records. Nevertheless, even though many assume more data and increased transparency are better for the discipline, the benefits of open access to field data have yet to see much clear demonstration in terms of research outcomes. Though we call for greater information disclosure and

ask researchers to reveal primary data at unprecedented volumes and levels of detail, we do not have clear examples that such disclosure actually improves understanding.

Impacts and assessments of sharing more archaeological data may emerge in the next few years. Other disciplines see clear evidence of positive impacts for open data, especially in terms of citation frequency. However, these impacts are often found in disciplines with more uniform data sets than are typical of archaeology.<sup>15</sup> Even with extensive metadata, archaeological documentation may create data sets too complex for widespread use by people outside the intimate circle of a specific project. Such an outcome could question the whole data-sharing enterprise and the effort and expense of data archiving. Why save a resource that, even with extensive documentation, is difficult for others to understand? If the archaeological community does not find effective reuses for archaeological documentation, the discipline may be collecting the wrong kinds of primary documentation. If so, one can imagine that efforts to improve data access may influence the kinds of documentation archaeologists produce.

If archaeologists see rewards in publishing primary data, they will be more likely to shape those data into forms that their colleagues will demand. In the past, primary data was typically developed to serve the limited audience of the project director and a few specialists. Primary data had little role outside of a project, so it tended (and still tends) to be informally developed, maintained, and validated. In working to publish these primary data, we are attempting to make the data useful and relevant to much larger audiences. That effort will inevitably transform the data and archaeological field documentation practices. Greater transparency will likely result in more professionalism and formalism in database design. Quality and consistency will become more important. Sloppy record-keeping that was tolerable in the past will likely give way to more carefully constructed documentation in the future. Validation and other quality assurance processes will become more generally applied. Finally, to be useful, these data will need standardization in semantics and description (metadata), even if that standardization stays at a generalized and abstracted level, as in Open Context.

---

<sup>15</sup> One archaeological example of the success of a more uniform data set is the Canadian Archaeological Radiocarbon Database (CARD) (<http://www.canadianarchaeology.ca/>), which is now in its ninth year and sees active use and citation in print publications.

As discussed in this chapter, effectively managing and understanding masses of information poses new challenges. Making mountains of data informative and useful for creating new knowledge is a daunting task. It involves technology, information architecture, data modeling, and service design, areas in which archaeologists have little experience or theoretical guidance. Yet the possibilities and potentials for research hold great promise. We hope that the conversation will expand beyond the technically savvy to include the theoretically sophisticated, the practically oriented, and others who think about, produce, and want to share and reuse data.

## REFERENCES CITED

- Boast, R., M. Bravo, and R. Srinivasan  
 2007 Return to Babel: Emergent Diversity, Digital Resources, and Local Knowledge. *The Information Society* 23: 395.
- Bowker, G. C., and S. L. Star  
 1999 *Sorting Things Out: Classification and Its Consequences*. Cambridge, Mass.: MIT Press.
- Brody, T., S. Harnad, and L. Carr  
 2006 Earlier Web Usage Statistics as Predictors of Later Citation Impact. *Journal of the American Society for Information Science and Technology* 57/8: 1060–1072.
- Bughin, J.  
 2007 How Companies Can Make the Most of User-Generated Content. *McKinsey Quarterly*. Retrieved from [http://www.mckinseyquarterly.com/Marketing/Digital\\_Marketing/How\\_companies\\_can\\_make\\_the\\_most\\_of\\_user-generated\\_content\\_2041\\_abstract#registerNow](http://www.mckinseyquarterly.com/Marketing/Digital_Marketing/How_companies_can_make_the_most_of_user-generated_content_2041_abstract#registerNow) (accessed June 4, 2008).
- Christen, K.  
 2005 Gone Digital: Aboriginal Remix and the Cultural Commons. *International Journal of Cultural Property* 12: 315–345.
- Denning, K.  
 2003 “The Storm of Progress” and Archaeology for an Online Public. *Internet Archaeology* 15. Retrieved from [http://intarch.ac.uk/journal/issue15/denning\\_index.html](http://intarch.ac.uk/journal/issue15/denning_index.html) (accessed May 2004).
- Doerr, M., and D. Iorizzo  
 2008 The Dream of a Global Knowledge Network—A New Approach. *Journal on Computing and Cultural Heritage* 1: 1–23.
- Elliot, T., and S. Gillies  
 2009 Digital Geography and Classics. *DHQ: Digital Humanities Quarterly* 3. Retrieved from <http://digitalhumanities.org/dhq/vol/3/1/000031.html> (accessed January 6, 2010).



- Haas, M., and M. T. Hansen  
2007 Different Knowledge, Different Benefits: Toward a Productivity Perspective on Knowledge Sharing in Organizations. *Strategic Management Journal* 28: 1133–1153.
- Harley, D., S. K. Acord, S. Earl-Novell, S. Lawrence, and C. Judson King  
2010 Assessing the Future Landscape of Scholarly Communication: An Exploration of Faculty Values and Needs in Seven Disciplines. UC Berkeley, Center for Studies in Higher Education. Retrieved from [http://escholarship.org/uc/cshe\\_fsc](http://escholarship.org/uc/cshe_fsc) (accessed June 22, 2010).
- Hearst, M.  
2006 Clustering versus Faceted Categories for Information Exploration. *Communications of the ACM* 49: 59–61.
- Heath, S.  
2010 Diversity and Reuse of Digital Resources for Ancient Mediterranean Material Culture. In *Digital Research in the Study of Classical Antiquity*, ed. G. Bordard and S. Mahony, pp. 35–52. Farnham, UK: Ashgate. Retrieved from <http://sebastianheath.com/files/HeathS2010-DigitalResearch.pdf>
- Hodder, I.  
1999 Archaeology and Global Information Systems. *Internet Archaeology* 6. Retrieved from [http://intarch.ac.uk/journal/issue6/hodder\\_toc.html](http://intarch.ac.uk/journal/issue6/hodder_toc.html) (accessed May 2004).
- Hollowell, J., and G. Nicholas  
2008 Intellectual Property Issues in Archaeological Publication: Some Questions to Consider. *Archaeologies* 4: 208–217.
- Isaksen, L.  
2008 Pandora's Box: The Future of Cultural Heritage on the World Wide Web. Slideshow as PDF. Invited keynote paper published in *Digital Heritage in the New Knowledge Environment: Shared Spaces & Open Paths to Cultural Content (Proceedings)*, ed. M. Tsipopoulou. Athens.
- Isaksen, L., K. Martinez, and G. Earl  
2009 Archaeology, Formality & the CIDOC CRM. Paper presented at Interconnected Data Worlds: Workshop on the Implementation of the CIDOC-CRM, Berlin, Germany, 23–24 Nov. 2009. Retrieved from <http://eprints.soton.ac.uk/69707/> (accessed January 31, 2010).
- Jeffrey, S., J. Richards, F. Ciravegna, S. Waller, S. Chapman, and Z. Zhang  
2009 The Archaeotools Project: Faceted Classification and Natural Language Processing in an Archaeological Context. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 367: 2507–2519.
- Juola, P.  
2008 Killer Applications in Digital Humanities. *Lit Linguist Computing* 23: 73–83.

- Kansa, E.
- 2005 A Community Approach to Data Integration: Authorship and Building Meaningful Links across Diverse Archaeological Data Sets. *Geosphere* 1 (2).
  - 2007 An Open Context for Small-Scale Field Science Data. In *Proceedings of the International Association of Technical University Libraries Annual Conference* (Stockholm, Sweden). Retrieved from <http://www.lib.kth.se/iatul2007/abstract.asp?lastname=Kansa>
  - 2009 Indigenous Heritage and the Digital Commons. In *Traditional Knowledge, Traditional Cultural Expressions and Intellectual Property Law in the Asia Pacific Region*, ed. C. Antons, pp. 219–244. Alphen aan den Rijn: Kluwer Law International.
- Kansa, E. C., and A. N. Bissell
- 2010 Web Syndication Approaches for Sharing Primary Data in “Small Science” Domains. *Data Science Journal* 9: 42–53. Retrieved from [http://www.jstage.jst.go.jp/article/dsj/9/0/9\\_42/\\_article](http://www.jstage.jst.go.jp/article/dsj/9/0/9_42/_article) (accessed June 29, 2010).
- Kansa, E. C., J. Schultz, and A. Bissell
- 2005 Protecting Traditional Knowledge and Expanding Access to Scientific Data. *International Journal of Cultural Property* 12/3: 97–109.
- Martinez, K., and L. Isaksen
- 2010 The Semantic Web Approach to Increasing Access to Cultural Heritage. In *Revisualizing Visual Culture*, ed. C. Bailey and H. Gardiner, pp. 29–44. London: Ashgate.
- McAfee, A. P.
- 2006 Enterprise 2.0: The Dawn of Emergent Collaboration. *MIT Sloan Management Review* 47: 21–29.
- McManamon, F. P., and K. W. Kintigh
- 2010 Digital Antiquity: Transforming Archaeological Data into Knowledge. *The SAA Archaeological Record* 10/2: 37–40.
- Palmer, C. L., and M. H. Cragin
- 2008 Scholarship and Disciplinary Practices. *Annual Review of Information Science and Technology* 42: 163–212.
- Pierce, M. E., G. Fox, H. Yuan, and Y. Deng
- 2007 Cyberinfrastructure and Web 2.0. Paper presented at International Advanced Research Workshop on High Performance Computing and Grids, Cetraro (Italy), 2007. Retrieved from <http://grids.ucs.indiana.edu/ptliupages/publications/Web20ChapterFinal.pdf> (accessed April 10, 2009).
- Piowar, H. A., R. S. Day, and D. B. Fridsma
- 2007 Sharing Detailed Research Data is Associated with Increased Citation Rate. *PLoS One* 2/3: e308. Retrieved from <http://www.plosone.org/article/lookupArticle.action?articleURI=info:doi/10.1371/journal.pone.0000308> (accessed April 20, 2007).

Roberts, P. D., G. B. Stewart, and A. S. Pullin

- 2006 Are Review Articles a Reliable Source of Evidence to Support Conservation and Environmental Management? A Comparison with Medicine. *Biological Conservation* 132: 409–423.

Rubinoff, R.

- 2004 How to Quantify the User Experience. Sitepoint. Retrieved from <http://www.sitepoint.com/print/quantify-user-experience> (accessed February 15, 2009).

Schloen, D.

- 2001 Archaeological Data Models and Web Publication Using XML. *Computers and the Humanities* 35: 123–152.

Steward, G. B., A. S. Pullin, and C. F. Coles

- 2007 Poor Evidence-Base for Assessment of Windfarm Impacts on Birds. *Environmental Conservation* 34: 1–11.

Trant, J.

- 2006 Exploring the Potential for Social Tagging and Folksonomy in Art Museums: 2006 Proof of Concept. *New Review of Hypermedia and Multimedia* 12/1:83–105.

Vaughan, L., and D. Shaw

- 2005 Web Citation Data for Impact Assessment: A Comparison of Four Science Disciplines. *Journal of the American Society for Information Science and Technology* 56/10: 1075–1087.

Warwick, C., M. Terras, P. Huntington, and N. Pappa

- 2008 If You Build It Will They Come? The LAIRAH Study: Quantifying the Use of Online Resources in the Arts and Humanities through Statistical Analysis of User Log Data. *Lit Linguist Computing* 23: 85–102.

Wilkinson, D. M.

- 2008 Strong Regularities in Online Peer Production. *Proceedings of the 2008 ACM Conference on E-Commerce, Chicago, IL, July 2008*. Retrieved from <http://portal.acm.org/citation.cfm?id=1386837>(accessed May 1, 2009).



## SECTION II

# THE TECHNICAL AND THEORETICAL CONTEXT OF ARCHAEOLOGY ON THE WEB

One of the most salient aspects of the Web 2.0 model is that it is primarily user driven. The foundation of Web 2.0 content lies in the fact that it is community and crowd sourced, user generated, and user contributed. *Folksonomies* are perhaps one of the best-recognized aspects of this user-generated approach to content creation. The idea that a community of interest (either defined formally as a profession or more informally as self-organizing) creates, generates, and manages tags to annotate and categorize content is quite a powerful one. In the case of archaeology, the folksonomy model sometimes mirrors and sometimes works at odds with archaeological typologies. Like folksonomies, many archaeological typologies are “user created” and can be loosely adopted in a community of regional specialists. In other cases, archaeological typologies are more formalized and institutionally generated, often making them less flexible, from both an intellectual and a technical standpoint.

This tension represents something of a challenge for digital approaches to archaeology. How do we reconcile the need for institutionalized typologies with the possibilities of socially based and user-sourced annotation and categorization? How do we merge and manage the top-down, institutionally derived approach of a more *cyberinfrastructure*-oriented perspective with the bottom-up, community-sourced approach of Web 2.0? How do we navigate the complexities of clashing systems of meaning in online contexts, especially when these take place cross-culturally? For example, is it possible to simultaneously meet the needs and desires of indigenous populations for cultural integrity and local meaning while maintaining interoperable analytical frameworks for scientific analyses and exploration?

The chapters in this section discuss and explore some of these questions. The chapter by Dunn (Chapter 3) attempts to explore how archaeologists

can effectively articulate the bottom-up, individually oriented model of Web 2.0 with the top-down, institutionally oriented model of cyberinfrastructure. He very rightly recognizes that archaeologists and cultural heritage professionals desperately need to address the intellectual and technical complexity of archaeological data within the context of existing Web 2.0 models and techniques. He also very astutely points out that the value of unconstrained archaeological data (coming from such Web 2.0 sources as blogs or wikis) may well be lost due to lack of structure and standardization.

The chapter by Boast and Biehl (Chapter 4) addresses the same general question from a different perspective. Instead of exploring the technical and intellectual challenges of applying folksonomies to archaeological collections and merging cyberinfrastructure with Web 2.0, Boast and Biehl see Web 2.0 as an opportunity to create spaces that foster and encourage dialogues that emerge from the different traditions in which an artifact has traveled. Further, they see Web 2.0 as source of inspiration for new models to rethink representation of cultural heritage materials.

Ultimately, the conclusion that both chapters reach is that while the impact of a Web 2.0, crowd-based model of knowledge generation on traditional archaeological practice raises significant technical and intellectual issues, it also provides exciting new opportunities to engage data and collections in a potentially more meaningful way than was previously possible.

### CHAPTER 3

# POOR RELATIVES OR FAVORITE UNCLES? CYBERINFRASTRUCTURE AND WEB 2.0: A CRITICAL COMPARISON FOR ARCHAEOLOGICAL RESEARCH

*Stuart Dunn*

## INTRODUCTION

**H**istorically, archaeology and disciplines related to the study of the past have been at the forefront of the use and development of advanced information and communication technology (ICT) tools and methods. In one sense, this is at odds with the broader “digital humanities”: as Susan Hockey has noted, “[a]pplications involving textual sources have taken center stage within the development of humanities computing as defined by its major publications” (Hockey 2004). This is fundamentally due to the nature of the material that these fields are concerned with. Texts occupy an important place in archaeological research, chiefly in the form of so-called gray literature reports of excavations, which are often the only extant records of those excavations, along with secondary literature and publications. But the bulk of primary archaeological excavation data comes in the form of numeric, graphic, statistical, and formal descriptions of the material record. This mass of digital evidence is geographically distributed, fuzzy, incomplete, inconsistent, and difficult or impossible to access. This paper identifies some of the broad research questions involved in this “complexity deluge,” following on

from the author's previous discussion of this issue in relation to virtual research environments (VREs) (Dunn 2009), and it attempts a critical review of so-called Web 2.0 methods and technologies as a means of addressing this emerging problem, as distinct from—and compared with—the broader context of research cyberinfrastructures. As such, it seeks to provide a guide for archaeologists and research managers approaching research data management issues. The software applications commonly associated with Web 2.0 may have matured considerably over the past five years, but the principles of user-centered design, interaction, and information sharing remain constant, and “Web 2.0” holds as a useful term of contrast with more centralized approaches. Broadly, Web 2.0 may be defined as the “bottom-up” tools and content (often, but not always, open source) on the Internet that have been created, developed, and maintained by distributed user communities; whereas cyberinfrastructure consists of discrete repositories, proprietary programs, and attributable content and databases produced and maintained by individuals or individual research groups. Elements of one are, of course, potentially interchangeable with the other: for example, a Web 2.0 application designed by a sophisticated user to create maps from CSV data, and contributed to an open forum, might create output files in formats for manipulation in proprietary (cyberinfrastructure) Web GIS systems; and it could be argued that such an application becomes part of the cyberinfrastructure once it is available. But beyond the semantics of individual cases, the most useful distinction for the purposes of this paper is that of *source*. Web 2.0 comes from the many, cyberinfrastructure from the few.

Pre-Web 2.0 (and indeed Web 1.0), archaeologists used computers to meet data challenges in a variety of ways. Even then, a common motivation linking applications was the need to manage increasingly large volumes of complex data and the workflows associated with using them. In a 2004 review article, H. Eiteljorg states that “the importance of context to the archaeologist highlights the importance of good records. . . . Those records are truly crucial, and the potential utility of the computer for that record keeping is obvious today” (Eiteljorg 2004). Databases, and the fundamental role they have come to play in the record-keeping process, are critical to understanding the relevance of computing to archaeology. Elsewhere, it has been argued that the emergence of advanced ICT in archaeology—from its beginnings alongside the inception of computing in the 1950s through to today's use of web services, service-oriented architecture (SOA), and grid com-



puting<sup>1</sup> to deal with very large and very complex data—may be seen as a technological process that reflects archaeology’s historical willingness to press cutting-edge scientific techniques into use. It would be difficult, for example, to see how current archaeological interpretations of prehistory could be constructed without the support of radiocarbon and thermoluminescence, both techniques unknown before the middle of the twentieth century (Dunn et al. 2007).

It is often said that “cyberinfrastructure,” which broadly equates to what is known as “e-science” in the United Kingdom (<http://www.nesc.ac.uk/nesc/define.html>; last accessed 6/9/2010), will enable new research and new *kinds* of research, as well as making existing research better, bigger, and faster. In a passage critical to the present discussion, Eiteljorg states that the “early and lasting interest in databases stemmed not only from the need to control huge quantities of excavation data but also from the hope that data store-houses [could be used by] scholars to retrieve and analyze information from related excavations, thus permitting broader synthesis.” However, disparity across culture and geography means that “combining data from multiple sources remains an illusive goal” (Eiteljorg 2004). Eiteljorg is undoubtedly correct to state that a top-down imposition of standards for recording, meta-data, and archiving is no solution. This chapter argues that creative use of Web 2.0 platforms, such as the interactive mapping technologies and services that are now widely available and supported by freely available (but proprietary) standalone applications such as Google Earth, along with a continuous process of critical reevaluation of how that mass of data is created, analyzed, and disseminated, will allow the archaeological community to reap maximum benefit from the emerging research cyberinfrastructure.

One straightforward aim of research cyberinfrastructure is to achieve interoperability between data sets that were not created with interoperability in mind. An obvious archaeological manifestation of this would be finds databases from two different sites whose excavations and recording procedures had been developed independently. A recent example comes from epigraphy: The Linking and Querying Ancient Texts (LaQuAT) Project has successfully

---

<sup>1</sup> Grid computing, at least in the sense most widely understood in Europe, refers to multiple resources, be they data, applications, or processing capacity, being bought together using middleware to perform a single task. Because the task is centrally defined, it fits the present conception of “top-down.”

demonstrated the use of OGSA-DAI middleware to connect three databases—of Roman laws, papyrological evidence, and epigraphic transcriptions, respectively—in different proprietary formats (Access and FileMaker) and in different languages (English and German) (Blanke et al. 2009). In the LaQuAT system, a user submits a query to a single interface, the query is mapped to all three sources, and coherent results are returned. This responds to well-established problems in archaeological data integration: Snow et al. (2006) identify three classes of data that currently cannot be accessed simultaneously: databases (whether government, institutional, or private) on different computer systems; “gray literature”; and images embedded in catalogs and reports. They argue that a “service-oriented cyberinfrastructure” is needed bring these together (Snow et al. 2006). They also identify non-technological challenges, which “require archaeologists first to understand each other’s concepts and procedures in order to comprehend each other’s data, methods and results. . . . [W]e require an e-science that marries the interconnectedness of digital research tools with the introspection enabled by traditional record-keeping” (Snow et al. 2006).

This hints at a key problem of terminology. The definition of an “archaeologist” transcends most conventional concepts of the academic or professional user; and as a branch of learning, archaeology frequently cuts into other disciplines, such as history, the natural and physical sciences, and philology, all of which deal with differing types of evidence in differing ways. Archaeology may be practiced in commercial, governmental, academic, or purely private and amateur settings (or simultaneously across any combination of these). Availability of resources for comparable efforts differs vastly across thematic and geographical areas. Archaeology’s subject delineation itself is problematic. In the United States, many faculties align it with anthropology, far more so than in Europe. It leaches into areas such as geography. Nearly 30 years ago, Renfrew noted that “[i]n a number of ways the *methods* of the geographer both at the hard (i.e. physical) and softer (i.e. social or political) ends have already proved of great value to the archaeologist. . . . [B]ut when the geographer seeks to look more closely at the role of human action in the past, he or she must set that action in a context that is more than simply spatial” (Renfrew 1983). So archaeology is, in itself, a disciplinary mash-up, needing support from a range of technological infrastructures, at all levels of scale and complexity. The emergence of Web 2.0 platforms therefore feeds into a long-standing discussion of where “the archaeologist” is situated intellectually.

## THE RESEARCH LIFE CYCLE

It is useful in this context to divide the process of using archaeological data into three high-level sections, beginning with the *collection and harvesting* of data. This can occur via digital capture in the field, or it can refer to the generation or scanning of digital representations of museum objects, or from secondary archival sources. Secondly, *analysis, integration, and interpretation* of archaeological data is considered. How can these techniques bring about new knowledge and link data in new ways (in the way Eiteljorg describes)? The last section explores new *social research processes*, based on digital dissemination and publication of the ever-increasing corpus of data and the workflows associated with excavation and survey. The creation and publication of data and secondary material through university libraries, Sites and Monuments Records, and cultural resources management organizations,<sup>2</sup> among others, has profoundly influenced the way users interact with archaeological data. The wider availability of digital infrastructure has greatly expanded upon—although, of course, not replaced—these sources. Yet the ability to create and share, as well as read information (Web 2.0 versus Web 1.0), adds an extra layer to these. E-science tools and methods allow the archaeologist to add, create, annotate, compartmentalize, and organize those data on their own desktops and in virtual server-side cyberspace in a so-called architecture of participation (see Batty et al. 2010).

## COLLECTION AND HARVESTING

The history of gathering data from archaeological excavation is long and complex. Until the middle of the nineteenth century, the retrieval of artifacts was little more than a money-making or prize-chasing activity. In the 1850s, the scientific recording of sites, artifacts, and features—the recovery of knowledge, rather than of simply valuable objects—emerged. Augustus Lane Fox Pitt Rivers, excavator of prehistoric remains at Cranborne Chase, Dorset, United Kingdom (Bowden 1984), adapted Darwinian principles of evolution to the material record. Describing science in 1875 as “organized

---

<sup>2</sup> In the United Kingdom, Sites and Monuments Records are official lists of ancient monuments maintained by local authorities; government organizations in other countries maintain equivalent data sets; cultural resources management organizations are (often commercial) repositories that hold information about historic sites.

common sense,” Pitt Rivers stated that “[w]e see from the facts that both growth and decay, the two component elements of evolution, are represented in the study of the material arts” (Pitt Rivers 1875).

Pitt Rivers recognized that systematic, or “scientific,” creation of content in the course of excavation enabled the development of typologies of material culture. The significance of such typologies in archaeology has endured. For example, the relative chronological framework of the prehistory of the Aegean basin is based on the study and evolution of pottery. In his seminal work, *The Palace of Minos at Knossos*, Arthur Evans divided the Bronze Age into the tripartite chronological system of Early, Middle, and Late, which has informed subsequent approaches to pottery typology (Evans 1928). His scheme was based on an evolutionary process of form, style, and technology, with the various sequential periods defined by the types of pottery produced in them.

Given that this conceptual approach has endured, a discussion of Web 2.0 in archaeology gives rise to the key question: who creates such typologies, and on what evidence are they based? If data are gathered and recorded, for example, through the **Portable Antiquities Scheme**<sup>3</sup>—a national database of small artifacts located by the general public and recorded online using standardized metadata—rather than through a systematic top-down excavation and interpretation, then the construction of typologies and vocabularies is not constrained in the same way. It is obvious that this raises fundamental issues for archaeological thought.

One such issue is the rethinking of what might be termed “industry standard” recording and publishing procedures, as outlined by Eiteljorg. Cyberinfrastructure has a crucial role to play in supporting and implementing such procedures, not only during post-excavation data processing, but also during the field data-gathering phase. A well-known example from the United Kingdom is the Virtual Research Environment for Archaeology (VERA) project<sup>4</sup> (for a full discussion of this project, see Rains, Chapter 5, this volume; see also Warwick et al. 2009: 222). The large-scale Silchester excavation has well over 100 staff working on site at any one time during the annual six-week excavation period, all of whom are, in some way, engaged in a complex series of interlocking workflows. All of the data produced by the excavation

---

<sup>3</sup> <http://www.finds.org.uk>

<sup>4</sup> See Clarke and O’Riordan 2009.

is stored in an integrated archaeological database (IADB). In the normal excavation process, descriptions, plans, dimensions, and records of artifacts are written on context record cards and then manually transcribed to the IADB in the post-excavation data-processing phase—a laborious and time-consuming process. A key element of infrastructure that the first phase of the project aimed to provide was a broadband network for the site, delivered by a wireless broadband aerial on a farm building some 600 m to the south of the main trench. Wireless-enabled devices such as personal data assistants (PDAs) and “ruggedized” laptops could then be used to enter data directly into the IADB system. Of course, the hardware itself presented a range of straightforward environmental problems. Sun glare, for example, severely impeded the use of liquid screen PDAs and laptops. There were instances where the PDA devices were found to be insufficiently robust for the all-weather outdoor work of field archaeology. And although this was not a problem with the ruggedized laptops, their high cost precludes any widespread adoption, even in well-funded excavations. In the project’s second phase, a system is being tested whereby digital pens are used in conjunction with the existing context card data framework to capture the fieldworker’s writing electronically. At the time of writing, preliminary results indicate a considerable workflow improvement (see <http://www.nesc.ac.uk/action/esi/download.cfm?index=4062> especially slides 11–16; last accessed 7/9/2010). In other words, the technology has been integrated seamlessly into existing practices such that the researcher, or in this case the excavator, need not even be aware of the change. The transcription process remains as a means of quality control: with a “second pair of eyes,” this process acts as a filter for archaeological, as well as textual, errors.

This useful lesson demonstrates that the successful introduction and adoption of Web 2.0 tools (with the term “tool” encompassing both hardware and software) and applications in archaeological field practice must be a process of evolution, not revolution. In some ways, this reflects the development of Web 2.0 from Web 1.0 (see Anderson 2007). The evolutionary process will select what is useful in terms of the archaeological research process: in VERA’s case, the pens are useful because they withstand the physically demanding environment *and* speed up the data-entry process. One may observe a similar process, over a much longer time scale, with (for example) the uptake of geographic information systems (GIS) technologies in archaeology: these are useful because they allow archaeologists to extract and

analyze secondary information from spatial data more easily than could be achieved with analog mapping technologies (see discussion in Wheatley and Gillings 2002). The emphasis on practicality, and user-friendliness, is paramount. This has been noted elsewhere. The vision articulated by a recent National Science Foundation workshop held at Arizona State University, entitled 'Enabling the Study of Long-Term Human and Social Dynamics: A Cyberinfrastructure for Archaeology,' for example, records that "we have encountered broad interest in, and very little resistance to, the development of a system for data sharing and cyberinfrastructure, *if it can be made practical*" [my emphasis] (Kintigh 2006a). This report is primarily concerned with the storage and integration of archaeological data, but the emphasis on practicality at the field excavation and survey stage is paramount. One of the very few things that the different types of archaeological communities discussed above—commercial, academic, amateur—have in common is a relative scarcity of resources and a resulting lack of scope for risking those resources on untried and untested cyberinfrastructure hardware (or methods). Therefore, for any kind of uptake to be possible in the field, it is essential that, in addition to delivering tangible benefits at some stage of the research cycle, any new piece of cyberinfrastructure should cost no more (in financial, human, or any other resource terms) than any existing component of the research cycle that it seeks to replace or supplement. Web 2.0 interfaces can be modularized in a way that, conceptually, reflects Snow et al.'s "service-oriented cyberinfrastructure"; the ability to combine photographs with Google Earth's representation of the earth's surface, for example, can be a powerful multimedia (or rather inter-media) mash-up, which can both combine data in new and useful ways and facilitate investigation of user interaction with those data (see below).

Like Web 2.0, archaeological e-science or cyberinfrastructure is a collaborative concept. In 2006, the United Kingdom's AHDS Arts and Humanities Data Service produced a report entitled "**Grand Challenges; Grand Opportunities? Archaeology, the Historic Environment Sector and the E-Science Programme,**"<sup>5</sup> which dealt with "Archaeology as a Virtual Research Organization." As this report notes, "Characteristically, E-science is about collaboration. Superficially, therefore, archaeology is well placed to contribute and benefit from the [e-Science] programme since almost all archaeological re-

---

<sup>5</sup> <http://www.ahds.ac.uk/e-science/documents/Archaeology-grand-challenges.pdf>

search projects are shared endeavours” (Kilbride 2006). The example given above, the VERA Project, shows that a number of people can, within certain constraints, use cyberinfrastructure to achieve a common aim—in this case the entry of data on contexts into the IADB—faster and more efficiently, thus freeing up valuable resources. The next section looks at how the archaeological community can build on cyberinfrastructures that enable collaboration in this way once the data have been gathered.

### ANALYSIS, INTEGRATION, AND INTERPRETATION

The number of obvious success stories in the development of digital tools for analyzing and interpreting archaeological data is relatively small. One example is **OxCal**,<sup>6</sup> a program for calibrating radiocarbon data. OxCal allows the user to enter an uncalibrated radiocarbon determination, and then it performs the necessary calculation and comparison with the radiocarbon tree-ring calibration curve to give the appropriate calibrated date BC. The software required to perform the algorithm resides on a server and requires only Web 1.0 technology to access and run it. In many ways, this highlights what Web 2.0 is *not*—it is not, primarily, a mechanism for creating new research *data*, but is more for enabling and understanding distributed research *processes*. A brief survey of the field of archaeological computing over the past decade makes it clear that the primary impact of Web 2.0 on archaeology will be on the recursive development of data by the community: one has only to think of the concept of “multivocality” developed by Ian Hodder and others at Çatalhöyük (see [http://www.catalhoyuk.com/TAG\\_papers/ian.htm](http://www.catalhoyuk.com/TAG_papers/ian.htm); last accessed 25/8/2009): logically, social software enables any number of users to contribute their “voices” to a body of data, rather than just reading it. It matters relatively little if those data are “born digital”—that is, gathered from field inquiry using hardware systems integrated with existing research practices, as described above—or comes from data sets digitized from extant nondigital sources. It is equally clear that infrastructurally and technologically, the field is still at a relatively early stage in this regard.

It has been argued that top-down e-science and cyberinfrastructure approaches are likely to have the largest impact in terms of integration. For example, Keith Kintigh has called for the establishment of a discipline of

---

<sup>6</sup> <http://c14.arch.ox.ac.uk/embed.php?File=oxcal.html>

“archaeological informatics” (Kintigh 2006a; 2006b<sup>7</sup>), in parallel with comparable information developments in the natural sciences. A top-down approach naturally implies the presence of consistent standards and metadata (see discussion in Dunn and Isaksen 2007). According to Kintigh, this comprises essentially three elements, the first of which is systematically collected primary data. Given that the processes of primary data collection are almost certain to be an effort shared between two or more people, this crosses over into the collaborative aspect of e-science, as described in the previous section. Second, Kintigh notes the importance of the voluminous but sparsely available “gray literature,” an issue that is being addressed in the United Kingdom and is discussed below. Third, the significance of integrating geo-spatial data is noted. The most significant challenge identified by Kintigh and his colleagues is the need for semantic integration, to “reconcile the semantic requirements of a user query with the semantic content of the available data sources” (Kintigh 2006b). A full exploration of the relationship of Web 2.0 (however so defined) with the so-called Semantic Web is beyond the scope of this paper, but the creation of user-generated semantics is a key feature of the former, and its contrast with what might be termed “professionally generated semantic typologies” as described above, supported by traditional publication channels, is critical.

In contrast, grid technologies, which do not rely on user-generated semantics, have a clear function to play in the future development of archaeological cyberinfrastructure. A paper outlining the proposed ArchaeoGrid Project notes that virtual organizations reflecting existing institutions in the academic, cultural heritage, and government sectors could subscribe to such a grid (Pelfer and Pelfer 2004), increasing the collaborative possibilities *alongside* semantically integrated data (which underlines the need described above for archaeological cyberinfrastructure to be integrated with existing systems). The achievements of the U.K.’s Archaeology Data Service should be mentioned here. Having accessioned over a million records, the ADS is the country’s principal provider of digital archaeological data and has pioneered the application of ontologically driven retrieval techniques in this area, including faceted classification (see <http://ads.ahds.ac.uk/>).

It is interesting to note that this is of concern elsewhere in the humanities (see, e.g., <http://www.nesc.ac.uk/action/esi/contribution.cfm?Title=773>). Inte-

<sup>7</sup> I thank Keith Kintigh for sending me a copy of this paper.



gration may require a semantic layer on top of existing data collections, one that allows formal (that is, machine-readable), sharable (containing knowledge that has been accepted by a group), and domain-specific representation of knowledge. It is also encouraging to note that the computer science community is responding to the technical challenges (Zhang et al. 2002). However, a less formal mode of integration could be invoked whereby two or more sets of data are bought together by a user for whom they simply share some kind of conceptual similarity. This could be easily supported by the kind of “user-profile” approach familiar from social networking sites; but, of course, it would be very difficult to support with any machine-readable or automated search-and-retrieval technology.

As outlined in the previous section, structured archaeological information is, in the main, described semantically. In the Cretan example given, “Late Minoan” and its counterpart from the Greek Mainland, the “Late Helladic,” are modern constructs—linguistic terms that describe certain decorative and chronological attributes, or comparable attributes, on (chiefly) ceramic objects, thus allowing typological classification, spatial provenancing, and attribution of date. In many cases, these semantic structures are deeply entrenched and decades old. However, a contrasting approach is provided by Gilboa et al. (2004). They take up the theme that such typologies are, in and of themselves, subjective and that the sheer volume of data (sometimes running into many millions of individual finds) makes selective publication difficult, expensive, and time-consuming, and complete print publication impossible. Furthermore, the vocabularies used to construct such typologies are often inaccessible to nonspecialists. Gilboa et al. (2004) describe an approach to typology that is based on shape geometry and computational techniques. Although there are certainly archaeological limitations on this—one can hardly discount the importance of well-recognized elements of cognitive material culture, such as vessel decoration, which most “traditional” classifications are based on—it is surely critical to guard against the development of data integration systems for archaeological cyberinfrastructure that deal *only* with semantic integration. Gilboa et al.’s study should be taken as an indication that cross-data-set assimilation and comparison on ever more complex quantitative and mathematical grounds may well have a far more significant role to play in the future.

There is also a well-recognized need for transparency in standards and metadata, which are constantly evolving in response to changes in data

curation and preservation practice (Kintigh 2006b). Descriptive metadata are needed for *data creation processes* as well as for the data itself. In order for data to be trusted, and therefore usable, users need to know how that data has been created, including all processes associated with its recording, context, and current status. The Silchester VRE example described above illustrates this perfectly. Just like the data itself, data creation and collection processes need to be recorded, standardized, described, and have appropriate metadata attached. A relevant project in the scientific community is the “myExperiment” virtual research environment,<sup>8</sup> a community web space where scientists can create, upload, and share workflows. As the project states, “myExperiment introduces the concept of a workflow bazaar; a collaborative environment where scientists can safely publish their creations, share them with a wider group and find the workflows of others. Workflows can now be swapped, sorted and searched like photos and videos on the web.” Publishing archaeological workflows in a comparable environment clearly brings issues that are peculiar to archaeology, but at its core, a discipline-specific workflow publication system in this area would need to formalize the relationship between the digital and analog components of the data creation process. This is an area of broad significance: the report of a 2006 NSF workshop on the challenges of scientific workflows recommended that the community should “integrate workflow representations with other forms of scientific record” (Deelman and Gil 2006). Bringing together existing forms of archaeological process and visualization, such as the Harris matrix, to create a formal system for recording how data are gathered and presented, would greatly support the kinds of cyberinfrastructure envisaged by Kintigh, Pelfer, and others.

There is broad agreement in the literature that data integration and description—whether qualitative or quantitative—holds the most immediate potential as an area to be addressed by archaeological cyberinfrastructure and, I believe, by Web 2.0. But it is useful to review some other areas where the medium-term impact is likely to be just as great. It is almost certain, for example, that agent-based modeling (ABM) will emerge as a key area of research that, indeed, could not be delivered *without* e-science/cyberinfrastructure. ABM uses a set number of software entities to predict what will happen to those entities within a given set of parameters. For example, the family

---

<sup>8</sup> <http://www.myexperiment.org>

units of an Iron Age farming community could be represented by a set of agents. It would be possible to simulate how those agents would act if there were, for example, a sudden lack of access to water, or if crops failed. ABM is, in other words, a powerful simulation tool that requires a great deal of computational power and, in many cases, federated data resources to build the parameters and frame instructions being given to the agents. However, it is important to note that ABM simulates; it does not attempt to replicate or reconstruct actual historic scenarios. As such, it needs to be treated with some caution: as the AHDS E-Science Scoping Study report noted, “[Previously] behavioural simulation [essentially ABM] . . . was not widely used because the computing power was insufficient but also because it fell out of favour intellectually. A firm prediction that this would become more popular again was qualified by the clear sense that the theoretical underpinnings would have to be included as part of the processing” (Kilbride 2006). Now that the computing power *is* sufficient to support ABM across federated data sets, its theoretical basis becomes an important subject for discussion. In this regard, the importance of documenting the process, or workflow, of how the data object—in this case the simulation—was created is at the fore. Just as the cyberinfrastructure in the field at Silchester only works if it is integrated into, and respects, existing analog systems, so ABM can only fulfill its potential if it is rooted in, and informed by, existing theoretical approaches to interpreting the past. Against this background, few in the United Kingdom would disagree ABM has also had a significant impact in the social sciences (e.g., Crooks et al. 2007). But the importance of combining ABM with technologies such as gaming cannot be understated in archaeological and historical e-science and cyberinfrastructure as well. For example, a recent study simulated the Battle of Trafalgar, compared simulative and analytical approaches, and found that the simulative approach agreed very well with the available historical data (Trautteur and Virgilio 2003). A project at the Institute of Archaeology and Antiquity at the University of Birmingham, entitled “Medieval Warfare on the Grid: The Case of Manzikert,” provides another example. The AD 1071 battle of Manzikert marked a critical strategic turning point in the history of the Byzantine Empire, but despite its historic importance, the historical (written) evidence for it is ambiguous in a number of ways. The Birmingham project will “provide a fundamental re-analysis of the Manzikert campaign and illustrate the use of Grid-aware distributed simulation techniques to model movement and sustainability of historic

armies” (see Gaffney et al. forthcoming). The fundamental point here is that this is a decidedly non-Web 2.0 research approach that *would not be possible* without HPC (high performance computing) infrastructure. The volume of data involved and the number of simulations needed constitute, essentially, a grid computing problem (as defined above). The development of technologies such as this throw up new kinds of requirements for archaeologists using them: Web 2.0 tools and applications can meet some of these by themselves; others require HPC and/or grids.

A further problem that has frequently been cited in support of the development of archaeological cyberinfrastructure is that of “gray literature,” reports that are compiled and then frequently left unpublished, or at least with very minimal circulation (Falkingham 2004). Practical and financial considerations often come between archaeologists and the ethical convention that as much data from excavations be published as soon as possible. Two projects are dealing with this at the time of writing. The **Archaeotools Project**<sup>9</sup> at the University of York (see also Jeffrey et al. 2009: 2507–2511) has explored the use of natural language processing (NLP) tools with the ADS library of gray literature, allowing users to browse in a more concept-based fashion founded in natural language processing. The concepts targeted are those of “what,” “where,” and “when,” which are of primary interest to users searching for archaeological information. In the Netherlands, the Open Boek Project is pioneering new methods of extracting and contextualizing numerical information from archaeological reports (Paijmans and Wubben 2007). These cyberinfrastructure developments provide and highlight new ways in which archaeologists can approach the mass of online archaeological data, and derive new data from it.<sup>10</sup> The application of advanced computational methods is stimulating new ways of conceiving of archaeological data. The application of NLP technologies highlights that information can be automatically extracted from digitized text and then treated as a database—such as a gazetteer of place-names, a list of chronological elements, or a thesaurus of objects or labels. This amounts to a fundamental reconsideration of the nature of evidence, of text versus data, and highlights that the boundary between them is becoming less clear.

---

<sup>9</sup> <http://ads.ahds.ac.uk/project/archaeotools/>

<sup>10</sup> Typically the gray literature as used in these applications is in the form of excavation reports, but in theory the approach is applicable to any kind of digital (or digitized) report arising from research or presenting results.

## SOCIAL RESEARCH PROCESSES

In contrast to the somewhat nebulous definitions of Web 2.0 currently available, the preceding sections have suggested that cyberinfrastructure and Web 1.0 applications in archaeology are relatively easy to define. However, the two are inextricably linked by complex Internet-based social research processes.

A service-oriented cyberinfrastructure for archaeology must support those processes. A full discussion of the theory behind service-oriented architecture (SOA) is beyond the scope of this chapter, but in brief terms, SOA may be seen as a set of technological procedures to enable existing components and services to function together for new or diverging purposes, without having to rebuild or redesign the system from scratch. Web 2.0, on the other hand, is an articulation of the vision that the flow of information on the Internet is a two-way process. This negates the concept of a clear divide between data users and data providers (and perhaps even prompts a rethink of what is meant by the word “data”). Given the range of possibilities and activities outlined above, there are clear implications for how data are created in archaeology, and by whom. This process of democratization gives us an opportunity to understand the social research process and to identify themes of popular concern (archaeology, of course, is an academic discipline with wide popular appeal). One simple, yet effective, example of this comes from the readily available mash-up that Google provides between its mapping service and the online photograph platform **Panoramio**<sup>TM</sup>.<sup>11</sup> This allows users to upload their own photos to Google Maps, where the photos are georeferenced (in most cases through being taken with a GPS-enabled camera). A small thumbnail of the image is placed on the Google Maps interface. If one examines, for example, the area north of Haydon Bridge, Northumberland, the thumbnail images form a clear linear representation of Hadrian’s Wall (Figure 3.1). This is the feature of greatest overall interest to the content-creating photographers in this geographical area; and so a composite and unified representation of the Wall emerges without any one of them making a conscious or unified intervention to create such a composition. Web 2.0 can, in other words, shape the content-creation process around archaeological features in the landscape.

---

<sup>11</sup> <http://www.panoramio.com/>

This democratization process is important for a number of reasons, and it can mean a number of things. However, giving users unfettered means for data creation can leave the representation of the archaeological record open to deliberate distortion and abuse, particularly in politically or economically volatile regions. This, of course, has always been the case with any archaeological discourse. However, a number of factors, including the (potential) anonymity of Web 2.0 environments and social software, combined with the ubiquity of access to them, alter the dynamics. For example, in the past, claims for the repatriation of cultural objects from one country to another have tended to be expressed (if not only felt or pursued) at the level of governments and/or national institutions: one thinks immediately of famous cases such as the Greek claim on the Parthenon Marbles, or Nigeria's on the so-called Benin Bronzes. In the future, however, the democratization of archaeological discourse by Web 2.0 tools and applications is likely to lead to loose, interested *non-governmental* groups (which could well be cross-border and cross-cultural in modern terms) gaining far more influence. This mirrors the globalization process visible in the curation of cultural heritage: as James Cuno has written, “[a]ntiquities are the cultural property of all humankind—of *people* not *peoples*—evidence of the world’s ancient past and not that of a particular modern nation. They comprise antiquity, and antiquity knows no borders” (Cuno 2008: 146; emphasis in original).

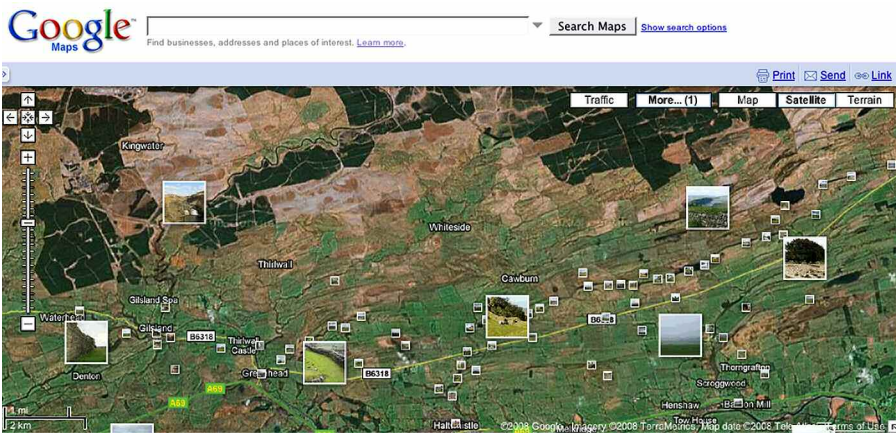


Figure 3.1. Thumbnail images on the Google Maps interface form a clear linear representation of Hadrian's Wall.

The issue of data quality is discussed in detail above; but where users are adding data, whether from a field excavation, secondary interpretations from desk research, or from anywhere else, there have to be mechanisms to make the data traceable. There are also practical problems with—theoretically—unconstrained content creation on, for example, blog sites. It can be difficult to structure or keep track of contributions temporally, although it has been pointed out that a simple but effective solution to this is to develop naming conventions for time, date, month, and year in the permalink URLs of such postings (McGrath 2007). The importance of providing adequate metadata at both the object and collection levels has long been recognized by curators of archaeological data (and indeed by those charged with curating and preserving any kind of humanistic data: see <http://www.ahds.ac.uk/metadata>). But in order to deal with unconstrained information created online by archaeologists (and the problems of defining that term have already been dealt with), we must add to this a need to document workflows as well as the data itself. A group of archaeologists using a blog to disseminate and discuss information about a project's progress are creating a collaborative digital object using a collaborative workflow. The need to source, provenance, and document that collaborative object is just as great as with any more traditional data object—a database record, a digital image, or a digitized context card. Although most users would not consider it necessary to archive all ephemera added by all users in the Web 2.0 milieu, the possibilities for documenting digital resources extend far beyond simply “adequate metadata”: user-selected blog feeds and personalized tag systems, for example, can enable users to select *what* ephemera they are interested in.<sup>12</sup>

A particular strand of the social research concept that will prove relevant for collaborative archaeological research is semantic social tagging. There are social websites for tagging many routine kinds of data objects: URLs (Delicious: <http://delicious.com/>), videos (YouTube: [www.YouTube.com](http://www.YouTube.com)), PowerPoint presentations (SlideShare: <http://www.slideshare.net>), citations (Connotea: <http://www.connotea.org>), academic blog and community information portals (<http://arts-humanities.net>), and related document-sharing services such as Google documents (Hannay 2007). Such sites create “tag clouds,” aggregated lists of the terms users have associated with each object,

---

<sup>12</sup> For a more complete discussion of this issue, see D. Wheatley, Rapporteur's Report: Heterogeneity, in Dunn and Isaksen 2007: 16.

in which the font size of each tag increases proportionately with the number of times it has been used overall. This “collective ontology,” or “folksonomy,” approach contrasts significantly with those familiar formal, top-down semantic structures, in which archaeological information has traditionally been stored (see above; also Guy and Tonkin 2006 for a review of the folksonomies in research). The importance of these semantic structures for data objects has been noted several times above. A common understanding of terms and agreed-upon descriptions of material culture is fundamental to the exchange of knowledge in archaeological communities. Misidentification and misclassification of individual objects can have serious implications for broader general theories and understanding. However, a folksonomy approach could work well *alongside* existing formal semantic structures and could indeed bring significant benefits in situations where, for example, a particular artifact is of disputed provenance: a vase may be tagged as dating from the Greek Geometric period by a site’s excavator, but a second researcher may come to the view that it is of Etruscan origin. If the latter were able to tag the vase as “Etruscan” alongside the original researcher’s “Greek Geometric” tag, then future information searches using either term would return the object, *without prejudice as to which of the original researchers was correct*. The person doing the search would then be able to view the data and come to his or her own conclusion.

## CONCLUSION

Both cyberinfrastructure and the application of Web 2.0 approaches are, to a greater or lesser extent, responses to the data and complexity deluges that archaeology has experienced. This article has attempted a synthesis of the distinction between the two as “top-down” (cyberinfrastructure) as opposed to “bottom-up” (Web 2.0). But the distinction is only a starting point; and it is clear that they should be regarded simply as different layers in the same structure. There are three main conclusions to be drawn:

1. It has long been recognized that integration of archaeological data based on common semantic descriptive frameworks is an important goal. But there is far more that can be achieved. The whole point of both Web 2.0 and cyberinfrastructure for archaeology is that the field can be taken beyond conventional archaeological computing, and that social research processes can be engaged more meaningfully with existing data, pro-



cesses, and conventions. Not only can the “data deluge” problems long identified by Eiteljorg and others be addressed along with questions of complexity deluge, but also whole new ways of doing research are opening up to us. It is not enough to simply extol the benefits of a framework for generating or integrating data (whether that integration is semantic, qualitative, or quantitative). The data will not become more meaningful, or more useful to society as a whole, unless they are accessible and reusable.

2. Ever more innovative means utilizing digital tools and resources are being developed within archaeology, and these may be regarded as components of “cyberinfrastructure.” Typically, where these have a network or online element, they are reliant on Web 1.0 technologies. However, knowledge derived from these form—or rather *should* form—part of the social processes inherent in Web 2.0. One must be conversant with the other at each stage.
3. The archaeological research cycle has traditionally been thought of as a process where data are produced via excavation and the identification of artifacts, processed during post-excavation analysis, prepared for publication, and then consumed by the wider community. As noted above, this reflects the Web 1.0 paradigm. Cyberinfrastructure, allied with Web 2.0 components, has enormous potential to add value to that. Although archaeologists who are not directly engaged with a particular project have always been able to contribute to the intellectual process of that project by commenting on its secondary publications, either informally or in print, or (where available) on its data records, data publication in interactive environments means that the relationship will be fundamentally changed. A “spade to screen” documentation process is needed so that the methods used to create every object and collection are transparent and attributable.

## REFERENCES CITED

Anderson, P.

- 2007 What is Web 2.0? *Ideas, Technologies and Implications for Education*. JISC Technology and Standards Watch. Electronic document, TSW0701. Retrieved from <http://www.jisc.ac.uk/media/documents/techwatch/tsw0701b.pdf> (accessed November 1, 2008).

- Batty, M., A. Crooks, A. Hudson-Smith, R. Milton, S. Anand, M. Jackson, and J. Morely  
2010 Data Mash-Ups and the Future of Mapping. JISC Technology and Standards Watch Horizon scanning report 10\_01. Retrieved from [http://www.jisc.ac.uk/media/documents/techwatch/jisctsw\\_10\\_01opt.pdf](http://www.jisc.ac.uk/media/documents/techwatch/jisctsw_10_01opt.pdf) (accessed September 7, 2010).
- Blanke, T., G. Bodard, S. Dunn, M. Hedges, M. Jackson, and S. Rajbhandari  
2009 LaQuAT: Integrating and Querying Diverse Digital Resources in Classical Epigraphy. In *Proceedings of Computer Applications and Quantitative Methods in Archaeology, Williamsburg, VA, 22–26 March 2009*.
- Bowden, M.  
1984 *General Pitt Rivers: The Father of Scientific Archaeology*. Salisbury: Salisbury and South Wiltshire Museum. (Reprinted 1995.)
- Clarke, A. S., and E. O’Riordan  
2009 Managing Change: Introducing Innovation into Well Established Systems. In *Proceedings of Computer Applications and Quantitative Methods in Archaeology, Williamsburg, VA, 22–26 March 2009*.
- Crooks, A., C. Castle, and M. Batty  
2007 Key Challenges in Agent-Based Modelling for Geo-Spatial Simulation. *UCL Centre for Advanced Spatial Analysis Working Papers Series* no. 121, September 2007. Retrieved from <http://www.casa.ucl.ac.uk/publications/workingPaperDetail.asp?ID=121> (accessed October 25, 2008).
- Cuno, J.  
2008 *Who Owns Antiquity? Museums and the Battle over Our Ancient Heritage*. Princeton: Princeton University Press.
- Deelman, E., and Y. Gil  
2006 Workshop on the Challenges of Scientific Workflows. Retrieved from [http://vtcpc.isi.edu/wiki/index.php/Main\\_Page](http://vtcpc.isi.edu/wiki/index.php/Main_Page) (accessed November 1, 2008).
- Dunn, S.  
2009 Dealing with the Complexity Deluge: VREs in the Arts and Humanities. *Virtual Research Environments: Issues and Opportunities for Librarians*, special issue of *Library Hi Tech* 27/2: 205–216.
- Dunn, S. and L. Isaksen  
2007 Space/Time: Methods for Geospatial Computing in Mapping the Past. Report of AHRC ICT Methods Network Expert Seminar, Edinburgh, July 2007. Retrieved from [http://www.arts-humanities.net/eventresources/spacetime\\_methods\\_geospatial\\_computing\\_mapping\\_past](http://www.arts-humanities.net/eventresources/spacetime_methods_geospatial_computing_mapping_past) (accessed September 7, 2009).
- Dunn, S., N. Gold, and L. Hughes  
2007 CHIMERA: A Service Oriented Computing Approach for Archaeological Research. In *Layers of Perception. Proceedings of Computer Applications and*

- Quantitative Methods in Archaeology, Berlin, 2–6 April 2007*, ed. A. Posluschny, K. Lambers, and I. Herzog. Bonn: Rudolf Habelt.
- Eiteljorg, H. II  
 2004 Computing for Archeologists. In *A Companion to Digital Humanities*, ed. S. Schreibman, R. Siemens, and J. Unsworth, pp. 20–30. Oxford: Blackwell.
- Evans, A. J.  
 1928 *The Palace of Minos at Knossos II*, Vol. 1. London: Macmillan.
- Falkingham, G.  
 2004 A Whiter Shade of Grey: A New Approach to Archaeological Grey Literature Using the XML Version of the TEI Guidelines. *Internet Archaeology* 17 (Winter 2004). Retrieved from <http://intarch.ac.uk/journal/issue17/> (accessed November 2, 2008).
- Gaffney, V., P. Murgatroyd, B. Craenen, and G. Theodoropoulos  
 Forthcoming “Only Individuals”: Moving the Byzantine Army to Manzikert. In *Digital Classicist: A Supplement of the Bulletin of the Institute of Classical Studies*, ed S. Dunn and S. Mahony. London: Wiley-Blackwell.
- Gilboa, A., A. Karasik, I. Sharon, and U. Smilansky  
 2004 Towards Computerized Typology and Classification of Ceramics. *Journal of Archaeological Science* 31: 681–694.
- Guy, M., and E. Tonkin  
 2006 Folksonomies: Tidying Up Tags? *D-Lib Magazine* 12/1 (January 2006). Retrieved from <http://www.dlib.org/dlib/january06/guy/01guy.html> (accessed October 28, 2008).
- Hannay, T.  
 2007 Web 2.0 in Science. In *CTWatch Quarterly, August 2007: The Coming Revolution in Scholarly Communications and Cyberinfrastructure*, ed. L. Dirks and T. Hey. Retrieved from <http://www.ctwatch.org/quarterly/articles/2007/08/> (accessed September 26, 2008).
- Hockey, S.  
 2004 The History of Humanities Computing. In *A Companion to Digital Humanities*, ed. S. Schreibman, R. Siemens and J. Unsworth, pp. 3–19. Oxford: Blackwell.
- Jeffrey, S., J. Richards, F. Ciravegna, S. Waller, S. Chapman, and Z. Zhang  
 2009 The Archaeotools Project: Facetted Classification and Natural Language Processing in an Archaeological Context. *Philosophical Transactions of the Royal Society A* (2009) 367: 2507–2519.
- Kilbride, W.  
 2006 Grand Challenges, Grand Opportunities? Archaeology, the Historic Environment Sector and the E-Science Programme. AHDS e-Science Scoping Study report, July 2006. Retrieved from <http://www.ahds.ac.uk/e-science/e-science-scoping-study.htm> (accessed November 2, 2008).

Kintigh, K.

2006a The Promise and Challenge of Archaeological Data Integration. *American Antiquity* 71/3: 567–578.

2006b The Challenge of Archaeological Data Integration. Paper presented at Technology and Methodology for Archaeological Practice: Practical Applications for the Reconstruction of the Past, Lisbon.

McGrath, S.

2007 Implementing Time by Expanding Space in Web 2.0. Retrieved from <http://www.itworld.com/Man/nlsebiz070917/pfindex.html> (accessed November 2, 2008).

Paijmans, H., and S. Wubben

2007 Open Boek: A System for the Extraction of Numeric Data from Archaeological Reports. *Proceedings of the All Hands Meeting 2007*. Retrieved from <http://www.allhands.org.uk/2007/proceedings/papers/804.pdf> (accessed November 2, 2008).

Pelfer, G., and P. G. Pelfer

2004 From WEB to GRID: A New Perspective for Archaeology. *Nuclear Science Symposium Conference Record, 2003 IEEE*, Vol. 2, p. 834.

Pitt Rivers, A.L.F.

1875 On the Evolution of Culture. A Lecture delivered at the Royal Institution of Great Britain on Friday, May 28, 1875, published in *Proceedings of the Royal Institute*, vol. 11, pp. 496–520; reprinted in M. W. Thompson, *General Pitt Rivers*, Moonraker Press, 1977, Bradford-on-Avon, Wiltshire.

Renfrew, C.

1983 Geography, Archaeology and Environment. *The Geographical Journal* 149/3 (November 1983): 316–333.

Snow, D. R., M. Gahegan, L. Giles, K. G. Hirth, G. R. Milner, A. Mitra, and J. V. Wang

2006 Cybertools and Archaeology. *Science* 331 (17 February 2006): 958–959.

Trautteur, G., and R. Virgilio

2003 An Agent-Based Model for the Battle of Trafalgar: A Comparison between Analytical and Simulative Methods of Research. In *Proceedings of the Twelfth IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE'03)*.

Warwick, C., C. Fisher, M. Terras, M. Baker, A. Clarke, M. Fulford, M. Grove, E. O'Riordan, and M. Rains

2009 iTrench: A Study of User Reactions to the Use of Information Technology in Field Archaeology. *Literary and Linguistic Computing* 24/2: 211–223.

Wheatley, D. and M. Gillings

2002 *Spatial Technology and Archaeology: Archaeological Applications of GIS*. London: Taylor and Francis.

Zhang, C., C. Cao, F. Gu., and J. Si

- 2002 A Domain-Specific Formal Ontology for Archaeological Knowledge Sharing and Reusing. In *Proceedings of Practical Aspects of Knowledge Management, Vienna, 2–3 December 2002*, ed. D. Karagiannis and U. Reimer, pp. 213–225. Heidelberg: Springer.



## CHAPTER 4

# ARCHAEOLOGICAL KNOWLEDGE PRODUCTION AND DISSEMINATION IN THE DIGITAL AGE

*Robin Boast and Peter F. Biehl*

## INTRODUCTION

This paper is part of an ongoing exploration located at the intersection of a number of related areas of inquiry, including digital field archaeology, information and communication technology (ICT), knowledge management, and the sociology of knowledge. At the core of each of these areas is a concern with the processes by which knowledge is produced and represented. This chapter presents several projects that are concerned with the ways such processes operate in the context of archaeological information as a means of sharing diverse forms of knowledge across communities. We write from a perspective that is informed by conceptions of knowledge as performance, of objects as citations, and of the potential of the Web as a contact zone; we identify the critical need to construct environments that support the generation and representation of knowledge in, by, and for different communities; and we evaluate the potential for the narratives, values, and interests of multiple knowledge communities to be appropriately represented with archaeological information that is created using the technologies and practices of social computing. Much of the work currently being done in these areas necessarily remains exploratory, and this chapter is a contribution in that vein.

## Knowledge Representation

Our assertion is that knowledge is a practice; it is a knowing of how to adjust to specific social-material settings (Smith 1996; Brown and Duguid 2000). Knowledge is performance: it is embodied in practice, not something we have, nor even something we can name consistently, but something we do. Moreover, a necessary condition for the generation of knowledge is engagement with other agents—other people and other things. However, engagement involves more than perception and cognition; it involves purposiveness and interpretation—that is, intentionality. Traditionally, the performance of archaeological knowledge tends to use two modes of representation, the interpretive and the classificatory. Archaeology generally treats interpretation and analysis as separate practices, interpretation being the representational mode that contextualizes the analysis, while the analysis is largely concerned with classification. The interpretive draws on a broad range of authoritative practices—method, institutional association, peer justification, and theoretical orientation, among others—that historically and practically permeate the discipline, while the classificatory—typology, stratigraphy, mapping, etc.—acts both as the first port of call for ordering the chaos of data and the last port of call for justification. We argue that there is a conflict between these two approaches. The systematic classificatory approach denies, fundamentally, the role of an object as citation. It gives fundamental primacy to the definitive account upon which all other secondary accounts are placed. The interpretive, on the other hand, engages with the classificatory largely as a mode of access to objects as illustrations. While archaeology has become increasingly open to grassroots access and to social computing's ability to provide for greater audience participation, an important step of reconsidering object citation and representation still has yet to be fully taken. This paper explores the possibility of this further step.

We argue that representation must involve a consideration of the diverse ontological frameworks associated with different expert communities, each with its own informed experience and interaction with the object. Archaeologists, cultural preservationists, curators, and, critically, key stakeholders must all interact around the object, and influence its selection, acquisition, classification, and presentation. This allows for online information systems to perform as “contact zones,” spaces that foster the divergent and incommensurable perceptions of objects and incite dialogues that emerge from the dif-



ferent traditions within which the object has traveled (Pratt 1992; Clifford 1997). Though these “contact zones” are not unproblematic (Boast 2011), they remain potentially powerful spaces.

Artifacts and sites, as pieces of tangible cultural heritage, are gateways to a number of intangible, yet critically connected practices: the telling of a story, the recitation of a prayer, the process of research, the history of the exhibition, the relation to other objects, and so on. Therefore, we wish to re-expose these intangible processes around the object through the consideration of “multiple ontologies” (Boast et al. 2007; Srinivasan 2007; Srinivasan and Huang 2005). We find this goal particularly pertinent and negotiable in the context of digital spaces and the possibilities of social computing to create new models for rethinking representation.

Museums have been experiencing many changes over the past three decades, beginning most significantly with a reorientation of the primary goal of museums, called by some the “new museology” (De Varine 1978; Vergo 1989). At the core of the new museology is an assumption that the museum is neither a center of research nor primarily a collecting institution, but rather an educational instrument. The goal of the new museology was, and largely still is, the transformation of social practices through the transformation of the museum from the display of singular expert accounts to a site of diverse educational engagements. However, no matter how much museum studies have argued for a pluralistic approach to interpretation and presentation, the intellectual control over the informational core of the museum—namely, its catalog of objects—has largely remained in the hands of the museum and its staff of elite experts. The extension of the new museology into museums, over the past 30 years, has introduced a regime where the educator and the marketing manager control the voices of the museum’s presentations for a relatively narrow, selective view of “public” interest. The maintenance of the museum as academic gatekeeper has been replaced by the museum as educational gatekeeper, focusing increasingly on simply supporting current educational programs and standardizing documentation of collections only to support their role as educational illustrations. This change is clearly represented in the dichotomy between the diversity of educational performances in museums (talks, guides, school tours, and exhibitions) and museum documentation, the methodical recording of information about the museum’s objects through careful study. While the museum allows many voices to be expressed—from different experts, authorities, and even

the public—rarely do these voices pass beyond a local and temporary educational performance, and rarely are they recorded in an enduring way in the museum’s catalog. Despite the numerous recent technological innovations that encourage contributions from a wide variety of distributed groups of users, traditional museum documentation practices persist, with narrowly descriptive catalog entries written by a small, select group of “expert” contributors.

In Macdonald’s reassessment of the “new museology” (2006), she argued that an undertheorized core of museum practice remains that fails to recognize the fundamental biographies of objects and their uses. That is, digital museums have done little to classify or annotate objects according to the different narratives and uses to which they are connected (Curtis 2006). This paper explores the hypothesis that this problem can be addressed by re-engaging objects with different expert accounts, and by reviving objects as agents within an ongoing exploratory dialogue (Boast et al. 2007). We assume that at least one of the principal motivations that people have when deciding to interact with an online catalog of museum objects—or of any objects, for that matter—is the goal of engaging with the objects themselves. Our understanding is that enabling users to directly engage with the objects themselves is the ultimate goal, but that resource discovery is a very important prerequisite for achieving that aim. We further argue that to engage with those objects, a mere technical description is not only insufficient, but also counterproductive. This paper therefore explores the theory that users will engage more deeply with digital museum objects when, alongside those objects, they are also presented with varied and even contradictory expert narratives (Turnbull 2003).

Numerous well-established museums are starting to experiment with the Social Web, the distributed, open-source, grassroots movement of Web users who are creating, modifying, and subverting online resources to an unprecedented degree (O’Reilly 2005, 2006). When applied to the museum context, social computing technologies have the potential to address the shortfalls of the single static object description, which has garnered a lot of criticism for traditional museum catalogs (Phillips 2005; Srinivasan and Huang 2005; Boast et al. 2007). Several notable projects are exploring the application of recent technological innovations to cultural heritage objects, in particular tagging and commenting. The **Steve.museum project**<sup>1</sup>—a partnership be-

---

<sup>1</sup> <http://steve.museum/>

tween several U.S. museums, *Think Design*,<sup>2</sup> and *Archives and Museum Informatics*<sup>3</sup>—is an ongoing exploration of whether, and in what ways, social tagging is applicable to describing works of art. By drawing upon the descriptions, impressions, and vocabulary of non-experts, the partners in the Steve.museum project are hoping to ultimately improve access to, and engagement with, works of art (Chun et al. 2006; Trant 2006). The *Reciprocal Research Network*,<sup>4</sup> a partnership between the Museum of Anthropology (MOA) at the University of British Columbia, the Stó:lō Nation Tribal Council, the U'mista Cultural Society, and the Musequeam Indian Band, is a collaborative project designed to extend collections-based research to source communities. While the project is still very much in development, the first iteration of their system (<http://www.rrnpilot.org/>) takes advantage of the commenting capability built into many web applications, allowing users to comment on objects in MOA's collections. The Recontextualizing Digital Objects around Cultural Articulations Project is a collaboration between the A:shiwi A:wam Museum and the Heritage Center at Zuni (New Mexico, U.S.A.), the Museum of Archaeology and Anthropology (MAA) at the University of Cambridge (U.K.), and the Department of Information Studies at UCLA (California, U.S.A.), designed to explore how digital repositories can be developed to recognize diverse forms of expertise, including the expertise of source communities, in describing museum objects. Their goal is to create a Web-based system that permits Zuni accounts to be directly incorporated into MAA's catalog, but that also functions according to local cultural protocols about the sharing of certain types of sacred or sensitive knowledge (Srinivasan, Boast, Becvar, and Enote 2009; Srinivasan, Enote, et al. 2009).

While these projects demonstrate the potential of recent technological innovations to engage stakeholder groups to participate in digital museum projects, what is still unclear about implementing social computing technologies into museum catalogs is whether these efforts can sufficiently balance the museum's account of objects with the input from the different sets of users in a way that yields a useful system for experts and non-experts alike. Our study aims to interrogate the very basis of the museum's classification scheme and knowledge base, its catalog. We hypothesize that two basic design errors limit

---

<sup>2</sup> <http://www.thinkdesign.com/>

<sup>3</sup> <http://www.archimuse.com/>

<sup>4</sup> [http://www.moa.ubc.ca/RRN/about\\_overview.html](http://www.moa.ubc.ca/RRN/about_overview.html)

the usefulness of most existing online catalogs of collections: (1) the requirement that catalog users search using concept labels drawn from a single, pre-defined set of vocabularies, usually following the traditional standards and vocabulary of the museum, and (2) the more general failure to provide catalog users with opportunities to truly engage with and manipulate the content of the records representing museum objects, let alone engage with the objects themselves. These design errors are likely the result of misunderstandings of the nature and roles not just of online museum catalogs, but also of museum objects and their removal from the consideration of practices of knowledge production (Bowker and Star 1999). Building on the case studies discussed above, we further hypothesize that the extent to which an online museum catalog provides a positive experience to its users depends on the extent to which its users are allowed to engage directly with museum objects through active participation in the discourses about those objects (Srinivasan and Huang 2005; Chun et al. 2006). Examples of the kinds of direct engagement we aim to explore in these studies include (1) generating and assigning uncontrolled accounts to objects' records; (2) discovery of objects of interest by navigating through the accounts and resources of other users rather than through the stagnant, monolithic structures within traditional museum classification; and (3) providing visual representations of objects, not just verbal ones. However, as is increasingly apparent (Kipp and Campbell 2006; Shirky 2008; Srinivasan, Enote, et al. 2009), such social computing interactions are not, in and of themselves, panaceas. Careful attention must be paid to the informational content of, and the modes of access to, the information. The following projects seek to initiate an inquiry into the power of social computing, but also to critically examine the imperative and classificatory modes of archaeological justification and representation.

## VIRTUAL REPRESENTATION

Virtual representation for producing and communicating archaeological knowledge has become increasingly important in the field of archaeology and heritage management in the past few decades. But it is a given fact that there are great potentials and serious dangers in using such multimedia technologies as virtual reconstructions and 3D animations to popularize archaeology (Biehl 2005); we discuss two case studies to illustrate this. Visual representations reproduce knowledge, whether by reproducing likenesses of

objects, places, or people, by organizing recorded data into visual formats for better communicability, or by reproducing the various interpretations of archaeologists and heritage managers. Van Dyke (2006) stresses that “visual representations are integral to the production of knowledge and scholarly authority.” Visual representations are often used by archaeologists and heritage managers not only to communicate information to one another, but also to make their interpretations available to the public. In recent years, one way to do this has been through outreach programs using digital media. It is true that computers have been used by archaeologists for a long time (see Boast 2002), but highly sophisticated and fast computer graphics have been available to archaeologists only in the past two decades. The 1980s marked the beginning of this trend, starting with the digital production of site plans, artifact illustrations, and the results from archaeological analyses. Computer graphics are a valuable tool, allowing researchers to represent and manipulate large amounts of complex data. Labeled “virtual archaeology” (Lehtonen 2005), this technology includes everything from reconstructions of sites and artifacts that can be created graphically from this amassed data to virtual reality reconstructions and 3D animations.

Virtual (or digital) archaeology is a powerful tool for visualizing and understanding archaeological data as well as for producing and communicating it to the public (Evans and Daly 2006: 253). It is also an educational resource for the general public and students in archaeology and heritage management. Many re-creations of greatly detailed archaeological sites have been created with standard modeling, rendering, and animation techniques. Digital archaeology makes possible increased rates of publication of archaeological materials through the use of the Internet. Its “open-source knowledge” allows researchers to quickly and inexpensively produce and communicate archaeological knowledge to a broad community of international specialists, schools, and the interested public alike and even to get them interactively involved in this process.

As funding for universities, heritage management groups, and museums becomes ever more limited, the Internet is increasingly pivotal for communicating archaeology (Biehl 2005). As such, archaeological knowledge needs to be efficiently produced and performed with multimedia applications so that it can be easily accessed by the public. The public, through tourism, represents one of the world’s most powerful sources of revenue. Visits to archaeological sites are often greatly educational. Unfortunately, the nature of tourism

is such that, even while economically beneficial to archaeology, heritage management, and the local economy, it sometimes also threatens the archaeological remains (Renfrew and Bahn 2008: 545–574).

One way to accomplish the dual goals of public outreach and preservation of archaeological remains is through digital archaeology and the Internet. The Internet has greatly expanded communication networks and the distribution of educational materials. The rate at which archaeological information is available online is ever-increasing. Site reports, virtual museums, digital reconstructions, and ideas are available almost instantaneously. Some even argue that the Internet is increasingly becoming the most important way to publish archaeological sites because of the wide distribution of knowledge and the frequency and ease of updates and new editions. The open-source quality of the archaeological knowledge on the Internet allows users to interactively modify, improve, and redistribute the knowledge. “The speed, range, and low cost of the internet have created new possibilities for dissemination and participation in knowledge construction and acquisition” (Hodder 1997). It offers access to raw data and the ability to form one’s own conclusions about archaeological materials. This has been seen as a move away from a hierarchical structure of interpretation to a more networked or multivocal approach.

These innovations bring with them the great potentials described above as well as serious dangers. Though multimedia presentations are a powerful tool for visualization, understanding, and communicating to the public, visual representations are biased—that is, they encourage one particular interpretation over another (Van Dyke 2006). Levy points out that “it is impossible to decide objectively between ‘good’ and ‘bad’ uses of the past; furthermore, there has been so much human movement, cultural mixing, and culture change in Europe that continuity from the past is a fiction” (Levy 2006). And there is a final danger with digital archaeology: its Euro-American perspective. Not all countries offer speedy broadband connections to their universities, museums, or heritage management services, not to mention access from public schools or private households.

However, we would like to discuss briefly two case studies that illustrate the popularization of archaeology in the digital age and one to discuss how archaeological knowledge is produced and communicated about online-museum collections.

PROJECTS: “MULTIMEDIA ARCHAEOLOGY” IN  
ÇATALHÖYÜK/TURKEY, EXCAVATION WEBSITES FOR  
POPULARIZING ARCHAEOLOGY IN GOSECK/GERMANY, AND  
ACCESSING DIGITAL CATALOGS—“BLOGGECTS”

Case Study: Multimedia Applications at Çatalhöyük—Digital Places

An important and influential website is that of [Çatalhöyük](http://www.catalhoyuk.com),<sup>5</sup> Turkey, a significant Neolithic site discovered in 1958 in Central Anatolia and excavated from 1959 to 1963 by James Mellaart and continued by Ian Hodder from 1992. The website features archived reports, databases, site management plans, illustrations, reconstructions, photographs, and video documentation, among other items. These allow interested parties to study and analyze the archaeological materials. The video documentation tracks not only the excavation processes, but also the views of the excavators. These videos are put on the website to ensure some sort of multivocality and have proven to be a good means to popularize the site and its archaeology, on the one hand, and to foster a better understanding of it among the public, on the other (Biehl and Gramsch 2002). Also included are lists of researchers and excavators, contact information, visitor instructions, forums, and blogs to encourage open communication networks.

Çatalhöyük exemplifies the methodological turn digital archaeology offers for producing and communicating archaeological knowledge. Video cameras and other multimedia equipment (Brill 2000; Stevanovic 2000; Wolle and Tringham 2000) bring to a large-scale excavation project a reflexive and fluid methodology and promote a pluralistic and “open” access to archaeological knowledge. Through this technology, knowledge producers can disentangle “the dichotomies between past and present, theory and method, interpreter and interpreted, subject and object, specialist and public, which are so troubling today” (Biehl 2002: 151). The latest trends in public outreach can also be studied at the Çatalhöyük project.

These cutting-edge and innovative projects are directed by Ruth Tringham and range from “remixing” (“[Remixing Çatalhöyük](http://okapi.dreamhosters.com/remixing/mainpage.html)”)<sup>6</sup> to “remediating”

---

<sup>5</sup> [www.catalhoyuk.com](http://www.catalhoyuk.com)

<sup>6</sup> <http://okapi.dreamhosters.com/remixing/mainpage.html>

(see *Senses of Places*,<sup>7</sup> the digital mediation of *Cultural Heritage and Second Life*).<sup>8</sup>

Still, documentation is one of the most important aspects of archaeology—that is, the listing of artifacts, mapping of site locations, and recording of positions and contexts of the artifacts within the strata. To create a detailed representation of an archaeological site or artifact, detailed measurements, observations, and other types of collections of data need to be accumulated (Lehtonen 2005). The digital tool Total Station—a combination electronic transit and electronic distance-measuring device—increases the speed at which finds and features can be recorded, allowing for many more finds to be recorded in much less time. This speed and efficiency increases the accuracy and thoroughness of excavations.

Still, big challenges remain. Archaeology frequently depends on archival data produced by other archaeologists or by researchers in other fields. Often, the archival data were recorded differently than those in the current project, causing noncomparable units of measurement and incompatible data formats between the two data sources. Project databases may be selective, and even when they are assessable, they may differ in size, format, or structure. Databases that have been compiled separately and are controlled by museums, government agencies, universities, or individuals may have been created on different computer platforms (Snow et al. 2006). In addition, there is a huge corpus of unpublished literature consisting of limited-distribution reports and so-called gray literature that has been mainly produced by commercial excavation firms and government agencies, as well as images, maps, and photographs embedded in museum catalogs and archaeological reports both published and unpublished. Standardized protocols are needed as well because of the confusion caused by modern political boundaries which are nevertheless irrelevant when talking about prehistoric, early historic, or environmental contexts.

Virtual excavations are constructed using a computer tablet and a GPS unit. Visitors to a virtual site see what the archaeological site would have looked like in the past. Not only can visitors see a site in its original state, they can also change their perspective or view the site without degradation

---

<sup>7</sup> <http://chimeraspider.wordpress.com/>

<sup>8</sup> <http://slurl.com/secondlife/Okapi/128/128/0>



by natural or human processes. And, of course, many more people can visit a virtual site on the Internet than can visit an actual site “in person.”

Computer programs also help archaeologists to reconstruct artifact assemblages by “finding adjoining pieces in a large collection of irregular fragments by comparing their shapes” (Da Gama Leitao 2002). Documentaries, too, are very important tools, utilized in communicating archaeology to the public. They can be viewed on TV as well as through the Internet (Van Dyke 2006). As an excavation progresses, the archaeologist never sees more than a single reference frame. As portions of a site are uncovered, they are recorded as data and a new reference frame is revealed while the first is forever destroyed by virtue of the second being revealed. By modeling the data, both artifacts and the matrix of associated soils, rocks, floral, faunal, and other documented finds, the researcher can essentially paint a motion picture of the excavation and the past.

### Case Study: “Multimedia Archaeology” in Goseck/Germany— Popularizing Archaeology

Archaeology as practiced in the digital age creates many more “artifacts” than simply the objects unearthed by traditional excavation methods. The recording system must accommodate *multimedia* in the true sense of the word—physical forms, plans, sketches, journals, slides and negative film images, video files, digital stills, audio recording, 3D models, GIS data, and satellite imagery. Multimedia is one way of addressing the representation problem by expanding the range and diversity of performances of the inscriptions from an excavation (Figures 4.1–4.4).

There are numerous technical solutions to this situation, for these are common problems in web and database design. However, the challenge is to create a solution that does not require the end users (archaeologists and the public) to become information technology specialists. It is essential that archaeologists be involved in the design process from inception through execution, and this means the solution must be understandable and operable by archaeologists. However, the solution also needs to be easily modifiable and must be robust and stable enough to sustain scrutiny from a worldwide user base. The **Goseck project’s website**<sup>9</sup> is built as an “*open knowledge*” (Open

---

<sup>9</sup> [www.praehist.uni-halle.de/goseck/](http://www.praehist.uni-halle.de/goseck/) or the main home page without flash animation: <http://www.praehist.uni-halle.de/goseck/index2.htm>

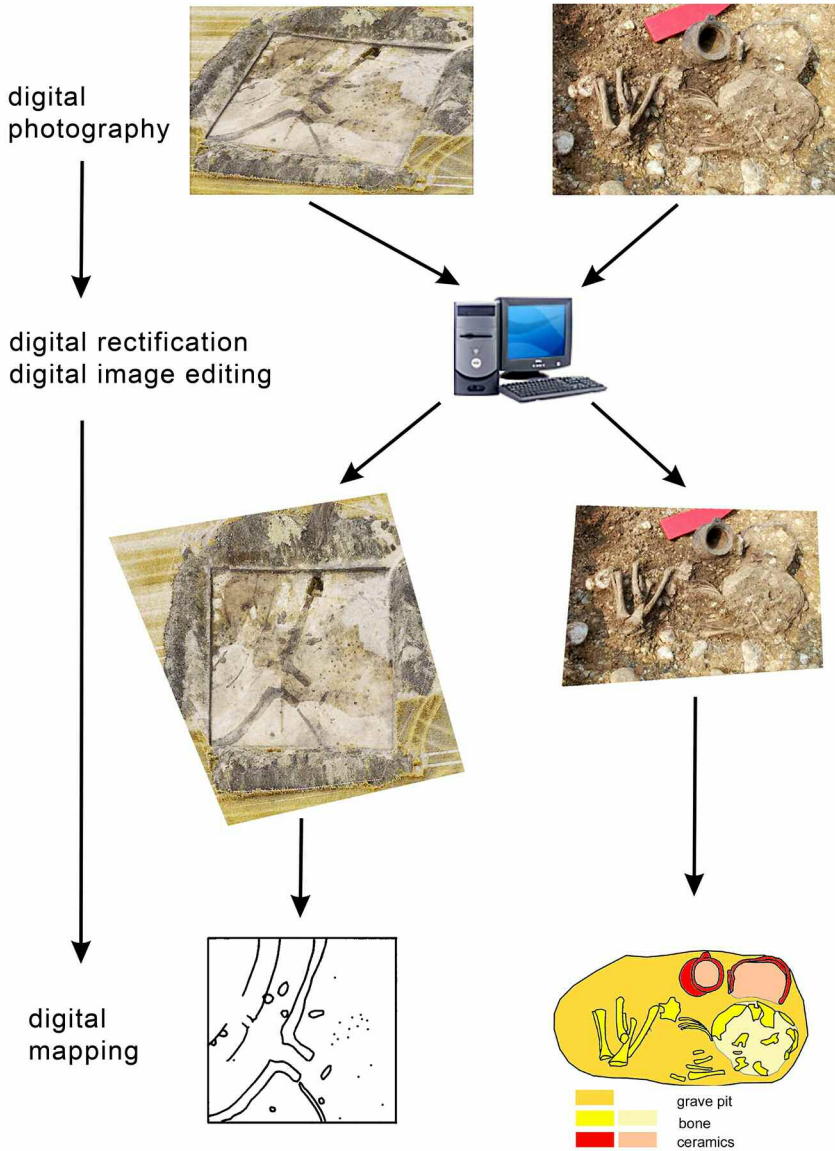


Figure 4.1. Multimedia applications in photo, video, and excavation documentation and digital reconstructions and visualizations.

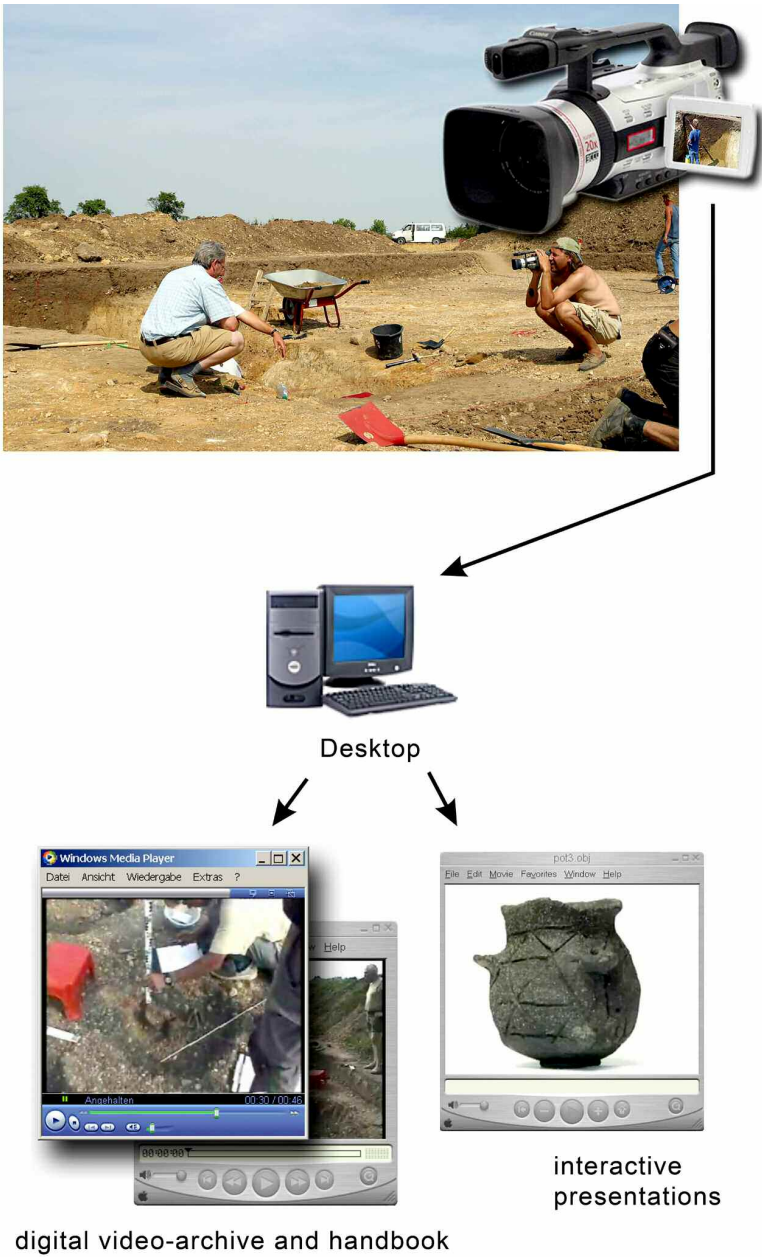


Figure 4.2. Multimedia applications in photo, video, and excavation documentation and digital reconstructions and visualizations.

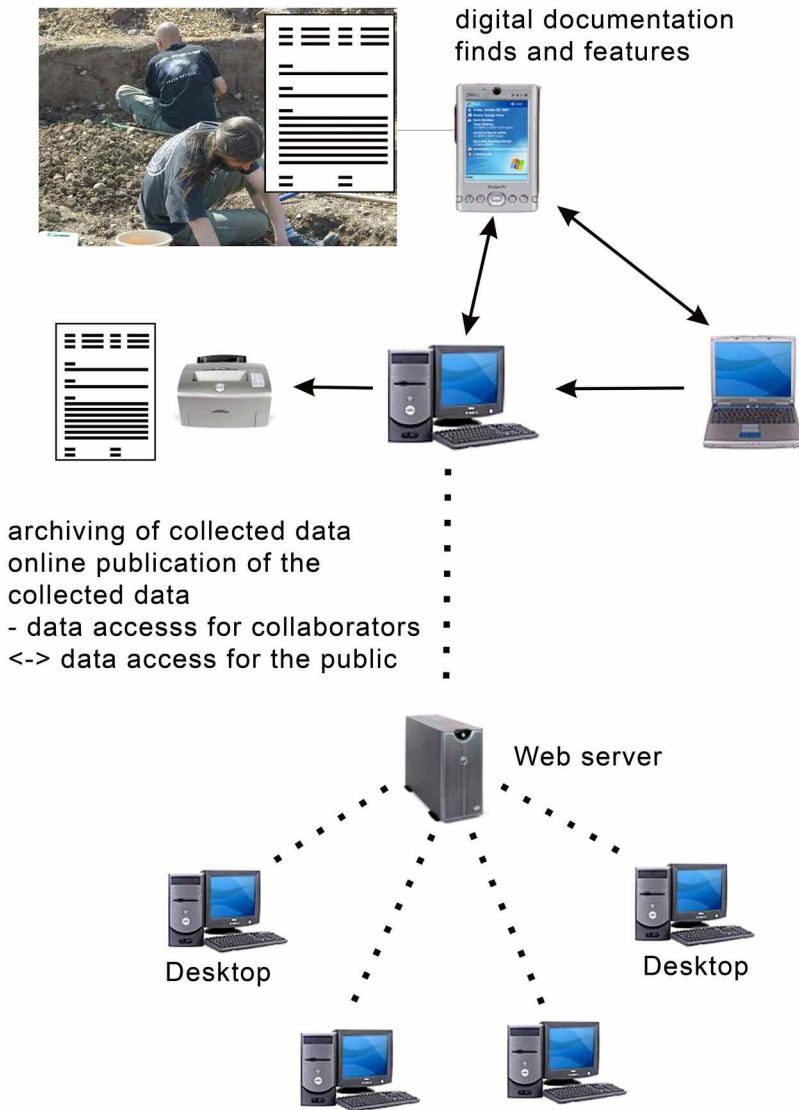


Figure 4.3. Multimedia applications in photo, video, and excavation documentation and digital reconstructions and visualizations.

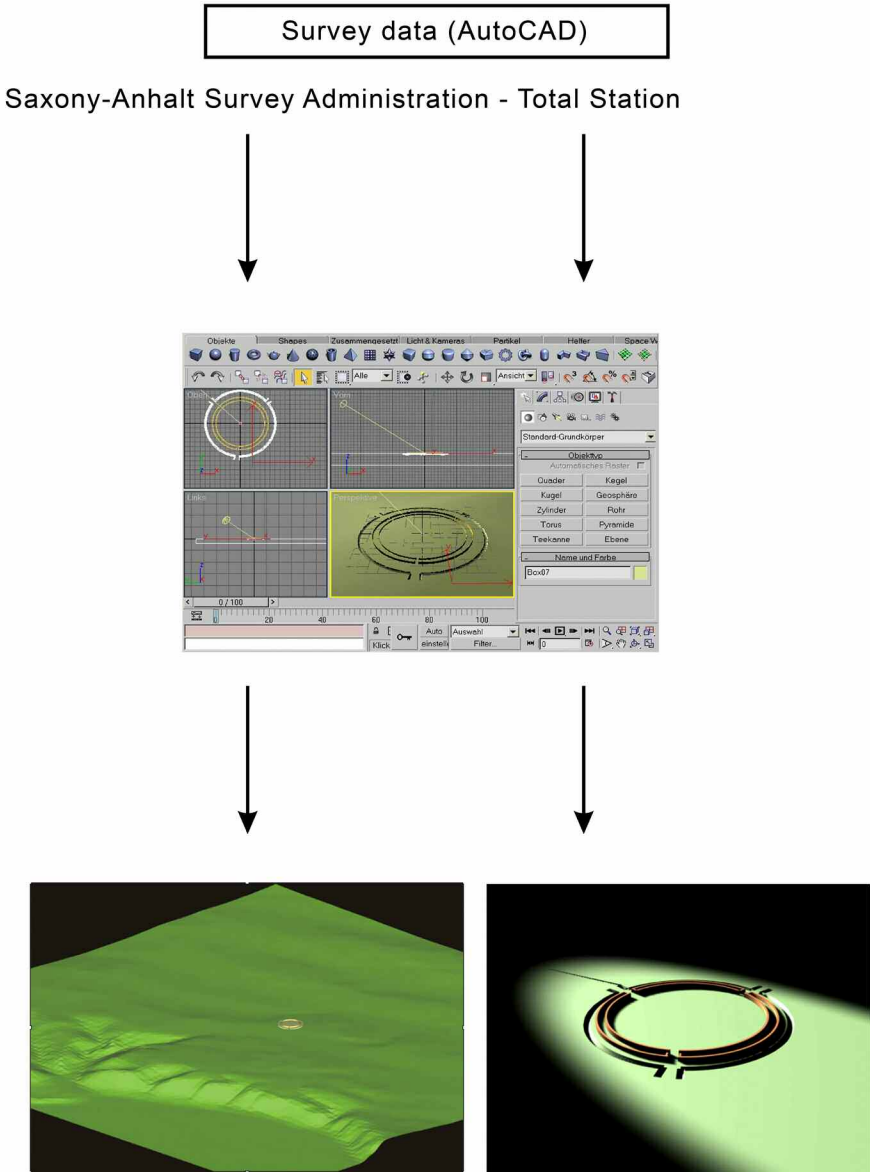


Figure 4.4. Multimedia applications in photo, video, and excavation documentation and digital reconstructions and visualizations.

Knowledge Foundation 2007) source that offers information both to the interested public (who may have no previous knowledge) and to archaeologists. It consists of differentiated levels of information, beginning with short introductory texts written in a popular manner, and extending through to detailed scientific reports supported by photos and videos, detailed descriptions, and illustrations of the archaeological data. Though all levels are accessible—which guarantees a general transparency—only the “deeper” levels of the website maintain a “scientific standard” of archaeological publication and provide the archaeologist-user with all available information of the excavated artifacts and their contexts (plans, photos, videos, and descriptions of finds and findings).

Elsewhere Biehl (2002) has discussed the enormous epistemological potential of hypermedia for archaeology. Rather than following an author’s linear argumentation in traditional forms of publication such as books and journal articles, readers/users of the Goseck website can browse through the information in a nonlinear way, approaching the data any way they want (Biehl 2002, 2005). Another advantage is that all data can be made available, which is normally not possible in traditional publications. Yet, despite the website’s universal access to all excavation data, in practice it is the virtual-reality objects that enjoy great popularity (see also Rieche and Schneider 2002; Samida 2004, 2006).

The fact that the layperson *and* the professional archaeologist can access the data from the Goseck excavation creates a new form of “knowledge transfer,” not only within the community of archaeologists, but also from the sciences to the public. In Goseck, the activities of archaeological excavation were transmitted via a webcam live on the World Wide Web (Figures 4.5–4.6).

The user can “look over the student’s shoulder” and vicariously participate in the archaeological training. The user can also learn about the daily work of archaeologists and see the first results of the excavation on the website. Communicating archaeology with interactive websites and live webcams can help us to make archaeology understandable and interesting to the public. Further, it helps archaeologists accept the responsibility for scientific transparency and sustainability in the research of regional history and monuments.

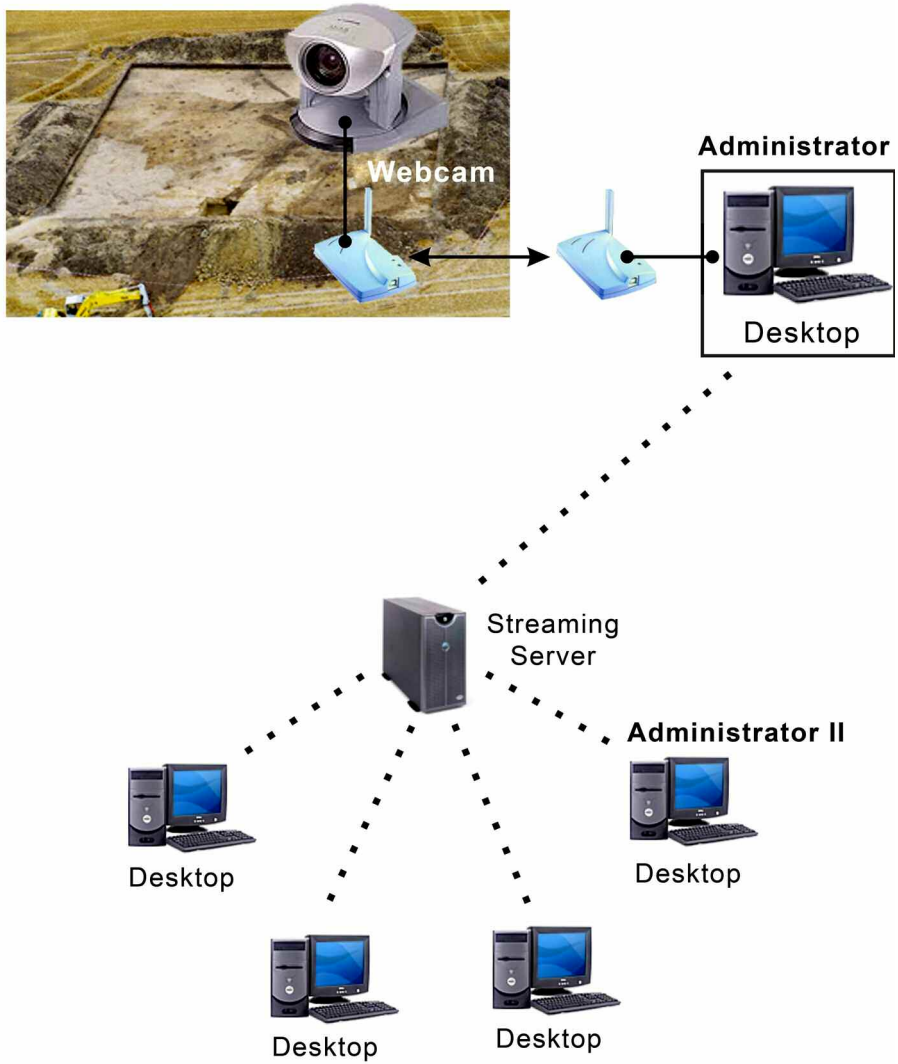


Figure 4.5. Webcam and videostreaming and their practical application on site.



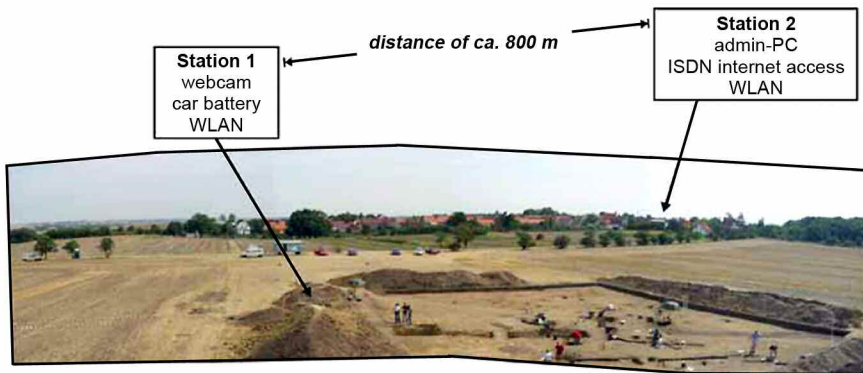


Figure 4.6. Webcam and videostreaming and their practical application on site.

## Case Study: Blobgets

**Blobgets**<sup>10</sup> was created at Cambridge University's Museum of Archaeology and Anthropology (MAA) to explore how people access and make sense of (or not) museum catalog entries online. The name "Blobgets" is a mash-up of the words "Blog" and "Object," just as the system itself is a mash-up of the functionality of a blog as applied to a catalog of museum objects. To this end, the study was focused on exploring how people would engage with relatively conventional catalog entries, but in a format that was familiar to most people but unfamiliar in a catalog—that is, in a blog. The study focused on how certain features of access—tagging and commenting—might impact the means by which users engaged with catalog entries as digital objects.

In particular, the study was designed to explore the role of unmediated catalog descriptions—that is, how well does a catalog description function as an *accurate* and *accessible* description of the object. All images were therefore intentionally omitted from the catalog entries, to ensure that the catalog descriptions were used without other mediating descriptions and so to test their validity, and to see how responses to these descriptions might perform in a social computing setting.

<sup>10</sup> <http://museum.archanth.cam.ac.uk/blobgets/>



The catalog entries used in Blobjects were drawn directly from the MAA's Collections Management System using the approximately 11,000 accessions (objects and photographs) from the Arctic. The vast majority of the material comes from collections made during the Wordie Arctic Expeditions of the 1930s to Greenland and Baffin Island. The material is not particularly contentious, as it was largely traded for openly during the expedition. However, there is a small proportion of the material that was excavated from sites during the expeditions. The data presented from the MAA catalog, which conforms to the **SPECTRUM documentation standard**,<sup>11</sup> included the usual public information (see example below). This information was not rewritten or modified for the Blobjects system—for instance, the original use of “Eskimo” was retained throughout the records—in the hope of prompting discussions of the nature of existing museum records.

IDNO: Z 45064 G

DEPT: Anth/Arch

NAME: Bone; Carving

KEYWORD: Tools; ?Art

MATERIAL: Bone

DESCRIPTION: Worked

Note with the objects reads: “These seven specimens were part of the priests collection from Abverdjar but from their appearance are obviously different from the rest of the collection and are probably either surface finds or mixed in by mistake by the Eskimo or at the priests house.” This record originally said this was a slate point. The slate point is marked A. The object marked G is bone. It has a dot pattern on the curved upper surface. The under side is flat. This object resembles a broken carving of a figure. S-J Harknett 23/1/2001

LOCAL NAME:

MAKER:

CULTURE GROUP:

---

<sup>11</sup> <http://www.collectionstrust.org.uk/stand>

SOURCE: Rowley.Graham.W (collector and donor)

SOURCE DATE: ? 1938; ? 1939

PLACE: Americas; North America; Arctic; Canada; Northwest Territories; Fox Basin; Abverdjar

PERIOD: Eskimo

CONTEXT: Date: ? Recent —; Collected by: Rowley.Graham.W.

The system has been inspired by the idea of creating a blog that would allow museum objects to be commented upon and tagged online. The Blobgects “experimental” version simply made the same metadata possible as the MAA’s standard catalog, but allowed users to modify, tag, comment, and so on. The results of the study confirmed that it is not simply the presence of social computing technologies that mattered, but the nature of the voices that use those technologies, ultimately allowing users to engage with multiple perspectives around the object. What was most apparent was the necessity, from the first encounter, for users to begin to create their own engagements with the objects unencumbered by excessive protocols or rules. In this regard, the initial prototype of Blobgects was considered a very successful failure: while it was not satisfactory as a standalone system, due to the nature and form of the information, the reactions gathered from users indicated a clear path forward to further developing digital museums that focus on making social computing capacities present while concurrently working actively to include direct interactions by relevant voices to provide context to the object in the form of a set of diverse perspectives.

This study was designed to compare results between two different user populations: a group of masters-level students in the Department of Information Studies at the University of California, Los Angeles (UCLA; U.S.A.), and a group of Inuit students at Inukshuk High School in Iqaluit, Nunavut Territory (Canada). Each of these groups is representative of an “expert community” interested in museum objects and their representation in catalogs, in that each maintains a distinct but important connection to the objects presented online, whether as part of cultural education of traditional objects from one’s community (Inukshuk) or as an object that must be shared with the public—and, in particular, with museum studies professionals—via a cultural institution (UCLA Information Studies students).

Each of the two user populations was divided into an experimental group and a control group. The experimental groups interacted with the fully functioning Blobjects system (see Figure 4.7), which displays a “tag cloud,” a set of hyperlinked descriptive terms used for navigation and access to groups of objects (for instance, clicking on “ivory” would bring up all objects with the term “ivory” in their catalog entry). This group could also search the system via a “simple” search from the home page or from a separate full-search page. Experimental group members were also allowed to add comments to entries if they wished. Importantly, the Blobjects tag cloud, rather than being user-generated (as is the case for many tagging sites like Flickr and delicious), was instead derived from terms found in the actual museum catalog records. This feature was designed to reveal whether a system identical to the MAA’s standard catalog system, in terms of the basic metadata provided, would prove superior if it allowed for social computing capabilities (in this case, navigating the Blobjects system via tags).



Collections Access Innovation

## Welcome to BLOBJECTS

BLOBJECTS is an innovative access system to the collections of the [Museum of Archaeology & Anthropology at the University of Cambridge](#). BLOBJECTS is based on a weblog where the blog entries are objects in the Museum's collections. This allows the visitor to search, comment, link via tags or RSS to any object or any search.

There are several ways of searching BLOBJECTS. One is to use the single entry SEARCH found on each page. This will return all objects that match the term entered anywhere in the record. A more accurate way of searching is to use the FULL SEARCH FORM. This may be accessed via the PAGES list found on the sidebar of every page. Finally, objects can be listed by clicking on the CATEGORY or on a TAG.

Help pages covering searching, categories, tags, RSS and commenting are listed under PAGES on each page.

SEARCH:

Find

PAGES:  
ABOUT BLOBJECTS  
• Collections  
• Emergent Databasing, Emergent Diversity (E02)  
Full Search

CATEGORIES  
object  
photograph  
document

TAGS:  
B/W\_print, arrow,  
arrowhead, baleen, blade,  
bone, bow, carving, expedition,  
figurine, film\_negative, fitting, flint,  
gelatin\_silver\_print  
glass\_negative, handle,  
harpoon, ivory, knife, lamp,  
lantern\_slide, pendant,  
photograph.point,  
scraper, sled, socket, spear,  
spearhead, stone, toggle, tool,  
transport, ulu, wood

RECENT COMMENTS:  
Laura: I don't know, but what's a baleen?  
Jennifer: What's a toggle?  
Tim Wilson: I believe there was already a comment on this, but what does "pierced" refer to?  
Jennifer: This bone entry is under the sled category but it has no internal sled tag. :(  
Tim Wilson: Why is there no reference to whales in descriptions of baleen objects?

ADMIN:  
Login

Figure 4.7. Screen-shot of the experimental Blobjects interface. Note the tag cloud and recent comments displayed on the right side.

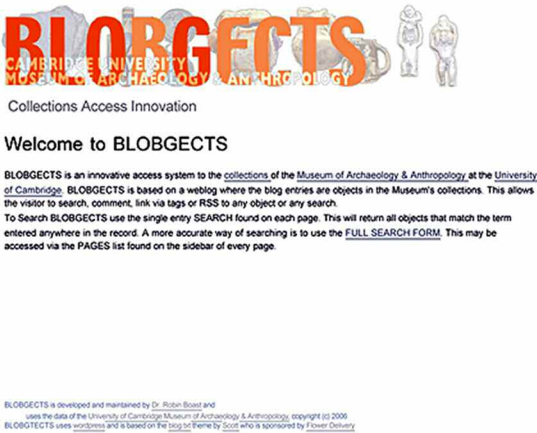


Figure 4.8. Screen-shot of the control Blobgets interface.  
Note the lack of the tag cloud and comments.

The control groups, by contrast, were presented with an identical version of Blobgets but without the tag cloud or commenting capability. These users had access to only three broad category terms as hyperlinks from the main page—“photograph,” “document,” and “object”—which meant they were restricted to interacting with the catalog alone (no user-generated information was available) and search was the primary mode of accessing objects in the system. This “control” system had the same functionality and content as Cambridge’s existing online catalog, but via an interface designed to resemble the experimental version (Figure 4.8).

Because part of the research study was meant to explore whether participants were interested enough in the items that they were engaging with to bookmark them for future exploration, participants were also encouraged to make use of the social bookmarking site [Delicious](http://www.delicious.com/)<sup>12</sup> during the study. Delicious is a Web-based bookmarking utility that allows users to tag sites with one-word descriptors, and those tags can be shared with other users. Delicious is one of several sites that Blobgets allows users to directly tag or link to (others include [digg.com](http://digg.com/),<sup>13</sup> [Technorati](http://technorati.com/),<sup>14</sup> [StumbleUpon](http://www.stumbleupon.com/),<sup>15</sup> and

<sup>12</sup> <http://www.delicious.com/>

<sup>13</sup> <http://digg.com/>

<sup>14</sup> <http://technorati.com/>

<sup>15</sup> <http://www.stumbleupon.com/>

**Bloglines.**<sup>16</sup> Tagging was not provided *within* Blobjects, though it could have been. The reason for this was to limit the test to see if the “raw” catalog entries would be sufficient to encourage further tagging within the user community. It could be argued that tagging within Blobjects would have better tested this premise, which may be a fair criticism. However, project designers felt that, in a preliminary study, the possible variables should be minimized. The results of this study are presented elsewhere (Srinivasan, Boast, Becvar, and Furner 2009).

As noted earlier, the study focused on how tagging and commenting might impact the means by which users engaged with catalog entries for digital objects. The most interesting outcome of this study was that the main feature of the Blobjects system, the ability to tag and to comment, had little to no effect—existing museum catalog metadata are simply too specialized to engage many different publics and “expert” communities. Through an extensive set of online questionnaires, before and after focus groups and during in-use discussions, what both sets of students told us—in particular, the students from Inukshuk—was that the classificatory order of the catalog as well as the imperative disciplinary idioms were the primary hindrances to use. It is not only that they found the classificatory structure inapplicable to their use of the objects, nor that they lacked comprehension of the navigational terms of the catalog; both of these skills can be acquired to a sufficient degree in a small amount of time. What we wish to suggest is that the students found these modes to be, first and foremost, barriers to access, forestalling any true grasp of an object until a deep understanding of the museum’s classifications and justifications had been gained. The study thus revealed the importance of narrative, dialogue, and image in contextualizing the objects, independently of catalog descriptions, and the potential these have for enabling users to move beyond definitive accounts. It also suggested that the many social computing tools of personalization and local description are not very useful without these complementary means of contextualization. More specifically, we note the following findings from this study:

- *The power of narratological tags:* Despite the rich, multiplicitous, dynamic nature of cultural knowledge production, we continue to create systems that mediate our interactions and preserve practices that are static, still

---

<sup>16</sup> <http://bloglines.com/>

focusing on retrieval questions that are disconnected from our actual interests and from ideas that would encourage active engagement. Even though the presence of Social Web software has positively opened up our categories from meta-ontologies, within the domain of multicultural systems and publics these systems fall short of actually sharing knowledges according to the contexts in which they are produced. We find in our study a possible way to reweave systems and cultures—that is, through narratological tags, not mere terms, but short accounts that connect through *citations* that can better contextualize and negotiate themselves into diverse knowledge practices.

- ▶ *Diverse users with diverse inputs add meaning to the online catalog:* Diverse inputs are often ambiguous relative to a descriptive perspective. Diverse expert communities add to these objects with concepts, images, and contextual information that may not be easily explanatory of the object for a layperson. Yet this ambiguity represents the reality of varied perspectives toward objects, and these ambiguities provide potential for inductive discoveries around the objects. As more users add to the digital object, the context of these seemingly ambiguous perspectives begins to become clearer and stimulate further insight.
- ▶ *Tagging must fit within a conversation:* We found that this process works within the online catalog system when it is embedded within a discursive conversation, a conversation between different social contexts and actors who have a connection to the object being presented. Diverse tags can serve as a mechanism by which the objects can stimulate new interactions between expert communities, and between museum visitors and expert communities. The tag is therefore not the exhaustive representation of the object, but the conduit for interaction among users and for a deeper sharing of context behind the object. The development of more extensive interactive systems is the subject of ongoing research (<http://collaborativecatalogs.blogspot.com/>).
- ▶ *The power of images:* Digital objects and digital museums may stimulate this cross-cultural dialogue when images are presented. The Blobgett experiment uncovered evidence that users are interested in interacting with, browsing, and retrieving objects via images and not just textual categories.

- *Blogs versus tags*: Participants are largely uninterested in status-quo tagging systems around digital objects, but the presence of the tagging system stimulates a reaction among participants to share different reactions that are not merely categorical and descriptive. Participants are interested in presenting social contexts, conversations, narratives, and images around the object, a process that may emerge more from a “blogging” framework than from a “tagging” one.

## RECONCEPTUALIZING DIGITAL OBJECTS AROUND CULTURAL ARTICULATIONS

Going beyond Blobjects, and to put into practice some of the lessons learned in the study described above, the MAA has joined with the Graduate School of Education and Information Studies (GSE&IS) at UCLA and the A:shiwi A:wam Museum and Heritage Center in Zuni (New Mexico, U.S.A.) on a project called “Reconceptualizing Digital Objects around Cultural Articulations” (RDO) (Srinivasan, Boast, Becvar, and Enote 2009; Srinivasan, Enote, et al. 2009). The project, funded by the National Science Foundation (NSF), is designed to bring distant collections back together with their source communities, but also, and primarily, to explore how to rejoin the source communities’ expertise with the objects in museums while maintaining individual and community intellectual property rights and rebalancing the museum’s editorial intervention over expert accounts.

A primary goal in this project is to explore both the similarities and differences between how local communities associate knowledge with objects versus how standardized museum systems do so. The publicly available stories, comments, and descriptions about objects from the Zuni participants in the study are compared here with the catalog entries about those same objects, forming the foundation for an analysis and the recommendations for further research. Specifically, the objects used in the study were originally excavated from the Kechiba:wa site at Zuni, New Mexico, during the early 1920s, as part of the larger Hendricks-Hodges Expedition directed jointly by the National Museum of the American Indian, Heye Foundation, and the National Museum of Natural History at the Smithsonian Institution (Isaac 2005). At the time of the excavations, the majority of the uncovered artifacts went to Washington, DC, but because of the participation at the time of

Cambridge's curator Louis C. Clarke, some of the artifacts crossed the Atlantic and became part of the collection at the MAA at University of Cambridge (Ebin and Swallow 1984).

As the Blobjects study argued, traditional museum catalogs have explicitly omitted the multiplicity of accounts and contexts that can be shared. This is partly due to an uncritical and largely hidden application of technology to the representation of cultural materials. That is, because all museum objects must now have descriptive metadata, catalog descriptions have inherited that focus and emphasize content standardization over all other issues (Crofts et al. 2009). The argument has been that such standardization is necessary to facilitate access and interoperability. However, content standardization comes at a high cost to the diverse local meanings of objects (Boast et al. 2007). Therefore, as part of the ongoing project "Emergent Databases: Emergent Diversity (ED2)," the RDO project has explored ways that museums can develop access systems that are able to accommodate and develop multiple ways of engaging with and understanding digital objects.

A fundamental component of this project is its collaborative intent. Every aspect of the research design and implementation has been enacted with the leadership of our Zuni colleagues at the A:shiwi A:wan Museum and Heritage Center (AAMHC) at Zuni, in order to ensure that the research process is relevant to local priorities, participation, agendas, and goals. Collaborative, participatory methodologies are gaining increasing acceptance across several social science disciplines (Robinson 1995; Bishop et al. 2001; Marshall 2002), and the proliferation of participatory methodologies in social science research reflects a fundamental decentering of the research paradigm (Tuhiwai Smith 1999). Moreover, this project is situated within a growing body of indigenous new media research that is based on local needs and agendas (Christie 2000; Salazar 2003; Christen 2006; Hughes and Dallwitz 2007). The preliminary set of objects to be circulated was selected by Zuni colleagues during a trip to Cambridge in 2006 and vetted by Octavious Seoutewa, a Zuni religious and cultural expert. This study also excluded objects with solely religious associations, because knowledge of a religious nature is sensitive in Zuni and is held by a few individuals on behalf of the community, making it an inappropriate topic for public inquiry and discussion (Isaac 2005; Jim Enote and Octavious Seoutewa, pers. comm.).

Over one hundred Zuni participants have been interviewed by the team of Zuni researchers, in a variety of locations, and these participants were sam-



pled from the larger population around the demographics of gender, age, and occupation, decided upon by AAMHC staff to be the most relevant social categories within the community. The stories collected in Zuni are, and remain, the property of the individuals first, and the Zuni second. The RDO project has access to only those accounts where the Zuni participants have decided that the content may be public. Many of the accounts collected will not be made available to the study, but will remain within the community for community use.

This study and its preliminary results are presented elsewhere (Srinivasan, Boast, Becvar, and Enote 2009), but several outcomes are relevant here. The structure and content of the MAA catalog conforms to the UK SPECTRUM Museum Documentation Standard (McKenna and Patsatz 2009). What was most interesting about the preliminary results of the RDO study was the extreme disparity, even incommensurability, between the MAA catalog description and the many descriptions and accounts arising from the Zuni participants.

Figure 4.9 shows the descriptions gathered about four objects representative of the larger collection: a fragment of a basket (MAA Z42472), a digging stick (MAA 1924.122), a pottery bowl (MAA 1924.473), and a rock with a naturally occurring lumpy shape (MAA 1924.101B). The size of the text corresponds to the number times that the study participants used that term or concept to describe the object they were looking at, and the “clouds” are clustered by general type of description—that is, “name,” “material,” “uses,” and so on. We have used a Venn diagram to represent that there are a few overlaps and similarities between how our Zuni participants described an object and how the Cambridge catalog did the same (shown in the center). But significantly, the majority of descriptions given by the Zuni participants (left side), relating as they do to past and present uses of objects and to stories and narratives about objects, do not have a corresponding description in the Cambridge catalogue (right side).

This disparity points to more than a difference in attention to different aspects of the objects. In such contexts, where different descriptions arise around the same objects, the traditional argument is that the descriptions are focusing on different aspects of the object. This is the “elephant in the room” argument. However, Figure 4.9 suggests that there is not a single “elephant” in the room, but rather quite different contexts of description, which lead to quite different objects being discussed (Law 1999). The object

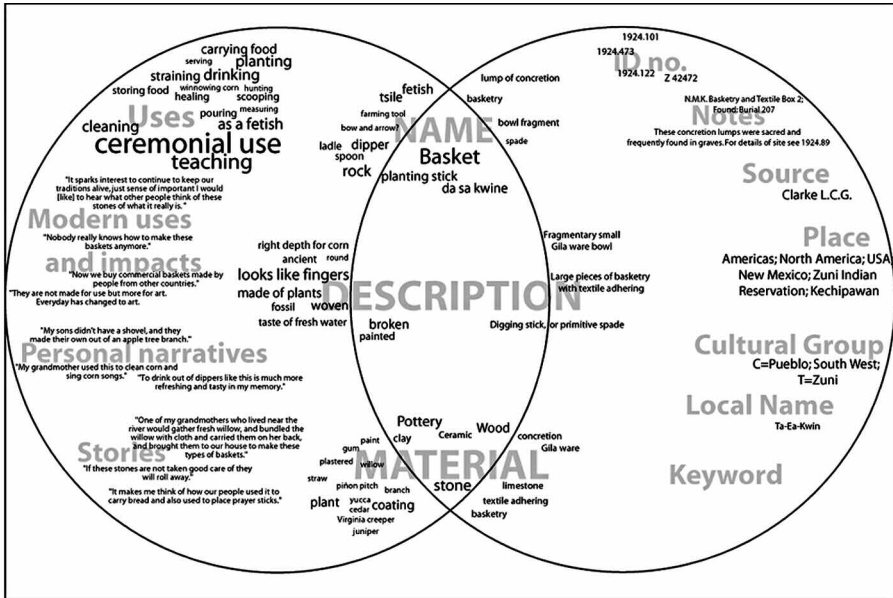


Figure 4.9. Comparison of Zuni comments (left) and MAA catalog records (right) for four sample objects, grouped according to metadata field or general category. The size of the text corresponds to the frequency with which that term appears in the data.

representations circulating around Zuni signify social practices that are completely different from those at the museum and, hence, represent fundamentally different social actors. The objects circulating at Zuni are participants in descriptive practices that differ most significantly in three ways from those descriptive practices found in the museum catalog.

### Stories and Narratives

An important part of the data gathered, which has no corollary in the MAA catalog, comes out of the stories shared by the Zuni when presented with the objects. Everyday experience of objects, around which we tell stories, is a key component of our understanding of the meaning of objects. Such stories, we argue, are crucial to cultural revitalization and for eliciting participation in the kinds of emergent cultural heritage systems that integrate and share multiple ontologies (Salazar 2003; Christen 2006).

Stories and narratives about objects can also be a way to discuss Zuni cultural objects in a way that aligns with Zuni ideas about the appropriate circulation of cultural knowledge. By focusing on personal experiences with objects, people are able to talk about important aspects of their lives as Zunis and still avoid revealing esoteric areas of knowledge. Isaac notes a similar approach to the discussion of cultural topics via personal experiences in her earlier analysis of the AAMHC, noting that the information that the staff chose to present—in other words, the “public sphere of local knowledge”—drew primarily from “personal, familial, or clan experiences” (Isaac 2005: 10–11). Excerpted below are some selections of the stories participants shared with our researchers when presented with the images of the objects from Cambridge:

I have [in my mind] an image of people using a basket to clean wheat and the smell [of] the wicker plants it is made of. [viewing MAA Z42472]

[M]y mother has a similar one [set of tweezers] that was used in the plaza ceremony with the bear dance and another dancer who had a yucca plant on him, and the bear tries to get them, but it was my mother who had the tweezer and took the fruits. We also used to make our own tweezers while staying at our Nutria farming village. [viewing MAA 24.119]

[This mortar reminds me of] grandfather making black paint. This same grandfather also survived the smallpox epidemic in the early 1900s, and [he] was passed for being dead but came back to life after three days of being comatose, [which] proved how strong he was but [he] was forever scarred by the smallpox. [viewing MAA Z42477]

The role of narrative here should not be seen as trivial or traditional—as somehow counter to “scientific” or classificatory data. Though this study focuses on a comparison between the museum catalog and the accounts of one expert source community, the Zuni, it can equally be extended to other specialist/expert communities that have developed knowledge practices around these same objects, such as archaeologists, anthropologists, historians, artists, or ecologists. These other specialist communities also give meaning to these objects through narratives, often narratives of use, but of a different nature and purpose. Like the others, these other specialist narratives also rarely

make it into the museum's documentation and are also, often, relegated to the exhibition or display, without a permanent association with the object in the museum.

## Uses, Both Historical and Modern

Based on our data, we confirm the hypothesis that objects-in-use are a critical way in which objects of cultural heritage are ascribed with meaning by source communities such as the Zuni. We argued elsewhere that people make sense of an object by how it is used, not merely by its physical description and characteristics (Srinivasan, Boast, Becvar, and Enote 2009). People are interested in considering objects in the context of practice—that is, within the rituals, activities, and lived experiences that support the object. Narratives, described above, fit within this parameter, but so do the uses connected with the object. In total, nearly two-thirds of the interviewees referred to the uses of objects when we asked the preliminary question about describing objects and stories related to the them (99 out of 158 object interviews). Later on in the interviews, we did ask questions specifically oriented toward the uses of objects, but the fact that our participants discussed objects-in-use when asked the open-ended question “How would you describe this object?” reveals the central importance of use and usefulness in describing and understanding objects.

The Zuni are interested in how these museum objects compare to things that are in contemporary use, a type of description that is also absent from the current catalog. As shown in Figure 4.9, our respondents made frequent references to modern impacts on the production and use of traditional objects like the ones we were showing them, especially on human-made objects (baskets, pottery, etc.). Topics that frequently emerged in discussions about the human-made objects include a loss of quality and knowledge in the production of objects, the fact that people no longer make or use these objects, and also the fact that people now purchase commercial products instead of making objects for their own use.

## Mobilizing Objects for Contemporary Agendas

It is clear that the Zuni see their traditional objects as important catalysts to vitalize contemporary social and political programs in the community, in par-

ticular teaching and learning (Clifford 1997). At some point in nearly every interview, interviewees expressed a link between the objects that they were looking at and learning more about Zuni culture and history, mentioning a desire to learn more, or something similar regarding the relationship between cultural education and objects like the ones we were showing them. This finding is consistent with the link between objects and learning upon which museums are built and which they have reinforced for decades. Because of the limited nature of the catalog entry, it is clear that merely providing access to catalog entries written for specialists does not mean that non-specialists can definitely learn from those entries. However, the absence of contextualization, and of comparisons to other objects, means that the “scaffolding” that is so important to the process of learning cannot take place when systems merely extract metadata from a museum’s catalog and make it accessible to the public, specialist communities, and the museum.

The RDO project is returning objects to the community, in the form of images and associated museum data, and eliciting accounts through local institutions. This is not an open grassroots commenting forum, though the project is being extended to include such programs, but it recognizes the importance, and existence, of diverse forms of expertise. Accounts are elicited from those members of the community, as identified by the community, who have a direct and deep understanding of the objects. The collected expert accounts are the property, first, of the individual, and, second, of the community. What information returns to the museum, to be associated with the objects, is in the hands of the local community.

The next phase of the project, “Creating Collaborative Catalogs: Using Digital Technologies to Expand Museum Collections with Indigenous Knowledge,” funded by the Institute for Museum and Library Studies (IMLS), has begun with the implementation of a system based on **PubSubHubbub**<sup>17</sup> where data is PuSHed from the museum to key stakeholder systems where it is incorporated into local knowledge systems. In phase, museum data is being posted into a local Zuni Knowledge System and all the collected Zuni expert accounts will be kept and managed in Zuni, and shared only under license from the Zuni. This phase of work is developing the publish, subscribe and hub services to automatically distribute all museum information of interest to

---

<sup>17</sup> <http://code.google.com/p/pubsubhubbub/>

the Zuni directly into the local Zuni Knowledge System held at the A:shiwi A:wan Museum and Heritage Center.

## CONCLUSIONS: SOME SUGGESTIONS OF WHERE TO GO NEXT

We end this paper not so much with a conclusion as with a postscript. These studies raise several issues that have always been there, but have been largely neglected. There is clearly a need for information in narrative form and for diverse contextualization to be developed. This suggests two major stages of access.

The first stage entails understanding how to present digital objects to multiple publics. Though this was not a study of semantics, we do feel that semantics is not, in itself, a useful way forward. Semantics, in the sense used by the W3 Semantic Web (Berners-Lee 1998), starts from the assumption that syntax is the bridge between ontology and epistemology. The work presented here suggests that understanding requires a consensus and participation from those using the information; that the relevance of the digital object arises not from the semantic designation of the object, nor from its role as an illustration of some definitive story, but from a context of use; that the context of these rich representations must be made apparent; and that through this dialogue with diverse images, accounts, and descriptions, others can begin to construct a meaningful understanding of these objects, sites, and practices. It is also through the process of meaningful use that others can begin to expand these understandings.

The usual response to this need has been to create interfaces to the information. Much of social computing operates on this assumption, with some real success: provide users with a platform for interaction and use, and leave them to it. However, this ignores the problem of context. Social computing offers a space for exploring the power of appropriation and reuse of digital objects, but this must be extended to consider the ability to contextualize and engage local and vernacular accounts of digital objects from multiple communities. Future research will continue to probe these critical issues and enable digital performance to serve as environments that support the generation and representation of knowledge in, by, and for diverse communities.

## REFERENCES CITED

- Berners-Lee, T.  
 1998 Semantic Web Road Map. Retrieved from <http://www.w3.org/DesignIssues/Semantic.html>.
- Biehl, P. F.  
 2002 Hypermedia and Archaeology: A Methodological and Theoretical Framework. In *Multimedia Communication for Cultural Heritage. Proceedings of the Multimedia Conference in Prato, Italy 2001*, ed. Franco Niccolucci, pp. 147–153. Florence: All'Insegna del Giglio.  
 2005 Archäologie Multimedial: Potential und Gefahren der Popularisierung in der Archäologie. In *Archäologisches Nachrichtenblatt* 10: 240–252.
- Biehl, P. F., and A. Gramsch  
 2002 Book Marks. Communicating Archaeology. *European Journal of Archaeology* 5/2: 249–250.
- Bishop, A., I. Bazzell, B. Mehra, and C. Smith  
 2001 Afya: Social and digital technologies that reach across the digital divide. *First Monday* 6/4. Retrieved from [http://www.firstmonday.org/issues/issue6\\_4/bishop/index.html](http://www.firstmonday.org/issues/issue6_4/bishop/index.html).
- Boast, R.  
 2011 Collaboration as Neocolonialism: Museum as Contact Zone Revisited. *Museum Anthropology* 34/1: 56–70.  
 2002 Computing Futures: A Vision of the Past. In *Archaeology: The Widening Debate*, ed. B. Cunliffe, W. Davies, and C. Renfrew, pp. 567–592. London: British Academy.
- Boast, R., M. Bravo, and R. Srinivasan  
 2007 Return to Babel: Emergent Diversity, Digital Resources, and Local Knowledge. *The Information Society* 23/5: 395–403.
- Bowker, G. C., and S. L. Star  
 1999 *Sorting Things Out: Classification and Its Consequences*. Cambridge, MA: MIT Press.
- Brown, J. S., and P. Duguid  
 2000 *The Social Life of Information*. Boston: Harvard Business School Press.
- Christen, K.  
 2006 Ara Irititja: Protecting the Past, Accessing the Future: Indigenous Memories in a Digital Age. A Digital Archive Project of the Pitjantjatjara Council. *Museum Anthropology* 29/1: 56–60.
- Christie, M.  
 2000 Galtha: The Application of Aboriginal Philosophy to School Learning. *New Horizons in Education* 103: 3–19.
- Chun, S., R. Cherry, D. Hiwiler, J. Trant, and B. Wyman  
 2006 Steve.museum: An Ongoing Experiment in Social Tagging, Folksonomy, and Museums. In *Museums and the Web 2006: Proceedings*, ed. J. Trant and D.

- Bearman. Toronto: Archives and Museum Informatics. Retrieved from <http://www.archimuse.com/mw2006/papers/wyman/wyman.html> (accessed January 28, 2008).
- Clifford, J.
  - 1997 Museums as Contact Zones. In *Routes: Travel and Translation in the Late Twentieth Century*, ed. J.Clifford, pp. 188–219. Cambridge, MA: Harvard University Press.
- Crofts, N., M. Doerr, T. Gill, S. Stead, and M. Stiff
  - 2009 *Definition of the CIDOC Conceptual Reference Model, Version 5.0.1*. ICOM/ CIDOC Documentation Standards Group.
- Curtis, N. G. W.
  - 2006 Universal Museums, Museum Objects, and Repatriation: The Tangled Lives of Things. *Museum Management and Curatorship* 21: 117–121.
- Da Gama Leitao, H. C.
  - 2002 A Multiscale Method for the Reassembly of Two-Dimensional Fragmented Objects. *Pattern Analysis and Machine Intelligence, IEEE Transactions* 24/9: 1239–1251.
- De Varine, H.
  - 1978 A Grass-Roots Revolution: Community Initiative in Culture. *Cultures* 5/1: 62–86.
- Ebin, V., and D.A. Swallow
  - 1984 *Proper Study of Mankind*. The Great Anthropological Collections in Cambridge. Cambridge: University of Cambridge Museum of Archaeology & Anthropology.
- Evans, T. L., and P. Daly (eds.)
  - 2006 *Digital Archaeology. Bridging Method and Theory*. London: Routledge.
- Hodder, I.
  - 1997 “Always Momentary, Fluid and Flexible”: Towards a Reflexive Excavation Methodology. *Antiquity* 71: 691–700.
- Hughes, M., and J. Dallwitz
  - 2007 Ara Irititja: Towards Culturally Appropriate IT Best Practice in Remote Indigenous Australia. In *Information Technology and Indigenous People*, ed. L. E. Dyson, M. Hendriks, and S. Grant, pp. 146–158. Hershey, PA: Information Science Publishing.
- Issac, G.
  - 2005 Mediating Knowledges: Zuni Negotiations for a Culturally Relevant Museum. *Museum Anthropology* 28/1: 3–18.
- Kipp, M. E. I., and D. G. Campbell
  - 2006 Patterns and Inconsistencies in Collaborative Tagging Systems: An Examination of Tagging Practices. *Proceedings Annual General Meeting of the American Society for Information Science and Technology, Austin, Texas, 3–8 November 2006*. London, ON. Retrieved from <http://dlist.sir.arizona.edu/1704/>.



- Law, J.  
1999 *Objects, Spaces, Others*. Lancaster: Centre for Science Studies, Lancaster University.
- Lehtonen, H.  
2005 Virtual Archaeology—What Is It? *Mirator* Theme Issue 2005: Proceedings from the Symposium “Virtually Medieval?” Retrieved from <http://www.glossa.fi/mirator/themeissue2005/vmlehtonen.pdf>.
- Levy, J. E.  
2006 Prehistory, Identity, and Archaeological Representation in Nordic Museums. *American Anthropologist* 108/1: 135–147.
- Macdonald, S.  
2006 Expanding Museum Studies: An Introduction. In *A Companion to Museum Studies*, ed. S. Macdonald, pp. 1–12. Malden, MA: Blackwell.
- Marshall, Y.  
2002 What is Community Archaeology? *World Archaeology* 34/2: 211–219.
- McKenna, G., and E. Patsatzi  
2009 SPECTRUM: The UK Museum Documentation Standard. Cambridge: Museum Documentation Association.
- Open Knowledge Foundation  
2007 Open Knowledge Definition. v1.0. Retrieved from <http://opendefinition.org/1.0/>.
- O'Reilly, T.  
2005 What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. Tim O'Reilly [blog], September 30. Retrieved from <http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html> (accessed November 27, 2007).  
2006 Web 2.0 Compact Definition: Trying Again. O' Reilly. Radar [blog], December 10. Retrieved from <http://radar.oreilly.com/2006/12/web-20-compact-definition-tryi.html> (accessed November 27, 2007).
- Phillips, Ruth  
2005 Re-placing Objects: Historical Practices for the Second Museum Age. *Canadian Historical Review* 85: 83–110.
- Pratt, M. L.  
1992 *Imperial Eyes: Travel Writing and Transculturation*. London: Routledge.
- Renfrew, C., and P. Bahn  
2008 Whose Past? Archaeology and the Public. In *Archaeology: Theories, Methods and Practice*, ed. C. Renfrew and P. Bahn, 5th edition, pp. 545–574. London: Thames and Hudson.
- Rieche, A., and B. Schneider (eds.)  
2002 *Archäologie virtuell. Projekte, Entwicklungen, Tendenzen seit 1995. Beiträge zum Colloquium in Köln, 5–6. Juni 2000*. Bonn: Habelt.

- Robinson, M.  
 1995 Shampoo Archaeology: Towards a Participatory Action Research Approach in Civil Society. *The Canadian Journal of Native Studies* 16/1: 125–138.
- Salazar, J. F.  
 2003 Articulating an Activist Imaginary: Internet as Counter Public Sphere in the Mapuche Movement, 1997/2002. *Media Information Australia incorporating Culture and Policy* 107: 19–29.
- Samida, S.  
 2004 Virtuelle Archäologie—Zwischen Fakten und Fiktion. In *Text und Wahrheit. Ergebnisse der interdisziplinären Tagung, Fakten und Fiktionen' der Philosophischen Fakultät der Universität Mannheim, 28–30. November 2002*, ed. K. Bär et al., pp. 195–207. Frankfurt am Main: Peter Lang.
- 2006 *Wissenschaftskommunikation im Internet. Neue Medien in der Archäologie*. Munich: Verlag Reinhard Fischer.
- Shirky, C.  
 2008 *Here Comes Everybody: The Power of Organizing without Organizations*. New York: Penguin Press.
- Smith, B. C.  
 1996 *On the Origin of Objects*. Cambridge, MA: MIT Press.
- Snow, D. R. G., L. Giles, K. G. Hirth, G. R. Milner, P. Mitra, and J. Z. Wang  
 2006 Cybertools and Archaeology. *Science* 311: 958–959.
- Srinivasan, R.  
 2007 Ethnomethodological Architectures: Information Systems Driven by Cultural and Community Visions. *Journal of the American Society for Information Science and Technology* 58/5: 723–733.
- Srinivasan, R., and J. Huang  
 2005 Fluid Ontologies for Digital Museums. *International Journal on Digital Libraries* 5/3: 193–204.
- Srinivasan, R., R. Boast, K. Becvar, and J. Enot  
 2009 Blobjects: Digital Museum Catalogs and Diverse User Communities. *Journal of the American Society for Information Science and Technology (JASIST)* 60/4: 666–678.
- Srinivasan, R., R. Boast, K. Becvar, and J. Furner  
 2009 Digital Museums and Diverse Cultural Knowledges: Moving Past the Traditional Catalog. *The Information Society* 25/4.
- Srinivasan, R., J. Enot, K. Becvar, and R. Boast  
 2009 Critical and Reflective Uses of New Media Technologies in Tribal Museums. *Museum Management and Curatorship* 24/2: 169–189.
- Trant, J.  
 2006 Exploring the Potential for Social Tagging and Folksonomy in Art Museums: Proof of Concept. *New Review in Hypermedia and Multimedia* 12/1: 83–105.

Tuhiwai Smith, L.

- 1999 *Decolonizing Methodologies: Research and Indigenous Peoples*. New York: Zed Books.

Turnbull, D.

- 2003 Assemblages and Diversity: Working with Incommensurability: Emergent Knowledge, Narrativity, Performativity, Mobility and Synergy. Paper presented at the annual conference of the Australasian Association for the History, Philosophy and Social Studies of Science, Melbourne, Australia, June.

Van Dyke, R. M.

- 2006 Seeing the Past: Virtual Media in Archaeology. *American Anthropologist* 108/2: 370–384.

Vergo, P.

- 1989 *The New Museology*. London: Reaktion.

Wolle, A. C., and R. Tringham.

- 2000 Multiple Çatalhöyüks on the World Wide Web. In *Towards Reflexive Method in Archaeology: The Example at Çatalhöyük*, ed. I. Hodder, pp. 207–217. McDonald Institute Monograph—British Institute for Archaeological Research. Cambridge: McDonald Institute for Archaeological Research.



### SECTION III

## ARCHAEOLOGICAL DATA MANAGEMENT AND COLLABORATION

Archaeologists face daunting complexity in managing data. Fieldwork often takes place in remote locations, far from communications and electrical infrastructure. Archaeological projects also often involve the coordination of many specialists, some of whom may work together in the field, but others who work asynchronously on stored collections long after fieldwork has terminated. Finally, archaeology (and its many subdisciplines) is a global enterprise. Archaeological conferences often attract international participants, and publications, reference collections, and expertise are all scattered across the globe.

No single technology can meet all needs of archaeologists working in these diverse settings. As a result, archaeologists often adapt commonly used hardware and software to meet their needs. Unfortunately, many “off-the-shelf” tools are designed for use in conditions and contexts quite different from those that often characterize archaeological fieldwork. Similarly, some popular social media platforms may not support the more specialized needs of a research community.

Archaeologists have made various efforts to adapt and develop technologies to better suit the particular requirements of this discipline. A handful of these undertakings are presented in this section. Each of these attempts represents an important and needed experiment in how to best use the Web and web technologies. Projects presented in this section explore using the Mobile Web, such as handheld devices and Internet-connected applications, to streamline work in the field; developing a comprehensive solution for organizing active field projects and coordinating among specialists; and leveraging the Social Web (interactions between people on the Web) to improve social and collaborative processes in archaeology.

All of these projects attempt to adapt web services, mobile and location-based services, and social media tools to improve collaborative processes in archaeology. They draw from a well-developed set of tools, standards, and design patterns pioneered in the Web 2.0 era. This experimentation is ongoing, and even more “casual” tools such as blogs and wikis are useful examples of how researchers are attempting to use the Web to improve scholarly communication.

The chapters in this section discuss tools and projects aimed at facilitating the collection, documentation, and communication of archaeological data. Lessons can be learned from each, and each will improve as user participation helps to hone designs for archaeological use. Simply put, there is no one “right” way to do this, and all of these projects demonstrate the need for continued iterative experimentation.

## CREATING A VIRTUAL RESEARCH ENVIRONMENT FOR ARCHAEOLOGY

*Michael Rains*

Development of the **Integrated Archaeological Database (IADB)**<sup>1</sup> began at the Scottish Urban Archaeological Trust (SUAT) in the mid 1990s and has continued in recent years at York Archaeological Trust. The original concept of the IADB was to make available digital versions of excavation records as an easily accessible integrated resource for use in post-excavation analysis and to provide a framework within which that analysis would be undertaken. Initially, the IADB dealt only with simple artifact records and stratigraphic unit or context records. Over time, the scope of the IADB has widened to include other digital resources, including single-context plans, photographs, and stratigraphy diagrams. Technically, the IADB began as a desktop database application written in dBase III/Clipper and has developed through several intermediate stages into what it is today, a Web-based user interface written in PHP and JavaScript which acts as a front end to a MySQL database. The IADB is currently in use in a number of commercial archaeological units and U.K.-based archaeological research projects in England, Albania, Romania, Iran, and Jordan.

In 2005, the Joint Information Systems Committee (JISC), part of the U.K. higher education funding framework, announced funding for a number of projects under its **Virtual Research Environments program**.<sup>2</sup> The JISC described the purpose of a virtual research environment (VRE) as being to “help researchers in all disciplines manage the increasingly complex range of

---

<sup>1</sup> <http://www.iadb.org.uk>

<sup>2</sup> <http://www.jisc.ac.uk/whatwedo/programmes/vre.aspx>

tasks involved in carrying out research” by providing a “framework of resources to support the underlying processes of research on both small and large scales” (JISC 2007). In practice, this translates into providing access, which in current circumstances means online access, to the resources (data) and applications (tools) necessary for research. Although not stated explicitly, the concept of collaborative working is also central to the JISC’s definition of a virtual research environment. In its earlier days, phrases such as “digital workbench” and “computerized desktop” (Rains 1995) were used to characterize the IADB; but by 2005 it was clear that the newly coined term “virtual research environment” could be applied equally well to it, and a joint bid by the Department of Archaeology at the University of Reading and York Archaeological Trust was awarded funding by the JISC for a two-year project under the acronym OGHAM (Online Group Historical and Archaeological Matrix). In 2007, an additional two years of funding was awarded for a continuation project entitled VERA (Virtual Environments for Research in Archaeology). This project included the original partners plus the School of Systems Engineering at the University of Reading and the School of Library, Archive and Information Studies at University College London.

Both OGHAM and VERA are centered around the **Silchester Town Life Project**<sup>3</sup> based at the University of Reading. This is a large-scale, long-running, and ongoing excavation of part of the abandoned Roman town of Calleva Atrebatum at Silchester, which lies approximately 80 km west of London. Silchester has used the IADB as its data management system since the start of the project in 1996. The OGHAM and VERA projects were aimed at developing specific aspects of the IADB, particularly in the areas of improved data flow and collaborative working.

The Silchester project has a small core team based in Reading and a larger group of specialists studying, among other topics, animal bones, pottery, and glass. Most of these specialists do not live or work in or near Reading or Silchester. They all have only a part-time involvement in the project and either have other day jobs in museums or other institutions or work as freelance archaeological specialists contracted for a wide range of different projects. Many will also have their own research interests. As a result, specialists often feel isolated or semidetached from the project. Key aims of the two JISC-funded projects have therefore included improving the flow of information from excavation, through analysis and research, to publication and

---

<sup>3</sup> <http://www.silchester.rdg.ac.uk>



dissemination, and developing a collaborative working environment through which all members of the project team can feel truly involved. Many of the technologies we commonly associate with the term “Web 2.0” have enabled and, to some extent, driven these developments in the IADB.

Developments toward a collaborative working environment within the IADB have taken a number of routes, including, for example, the creation of internal messaging and chat systems, and the provision of online collaborative document-editing facilities. Significant among these is the concept of the virtual research domain (VRD), which was developed as a way of encapsulating the key features of a virtual research environment within the IADB. Each VRD is designed to address a particular research issue or activity within an archaeological project, such as the stratigraphic analysis of a phase of the site or the analysis of the coin assemblage from the site. Some key features of a VRD are:

- ▶ The VRD should provide simple, direct access to all the key resources required to address the particular research issue.
- ▶ Recognizing that the end product of most archaeological research is the production of one or more documents, the VRD must provide for collaborative online document creation and editing.
- ▶ Access to and use of the VRD must require minimal user training.

At the heart of a virtual research domain are one or more structure diagrams, which are interactive graphical representations or visualizations of part of the project database. The starting point for a structure diagram of, for example, a particular phase in the development of the site, would normally be a standard archaeological stratigraphy diagram showing the excavated contexts in the phase and the stratigraphic relationships between them (Figure 5.1). To this are added, for example, a plan of the contexts, one or more photographs, and any other resources from the project database considered relevant to the research topic being addressed. Most significantly, one or more documents are also added to the VRD (Figure 5.2). These will most likely be blank initially. They will be completed by the researchers working in the VRD and can be thought of as the “factory floor” of the virtual research domain. A VRD addressing the research topic of excavated finds might contain less stratigraphy and more artifact and image resources (Figure 5.3).

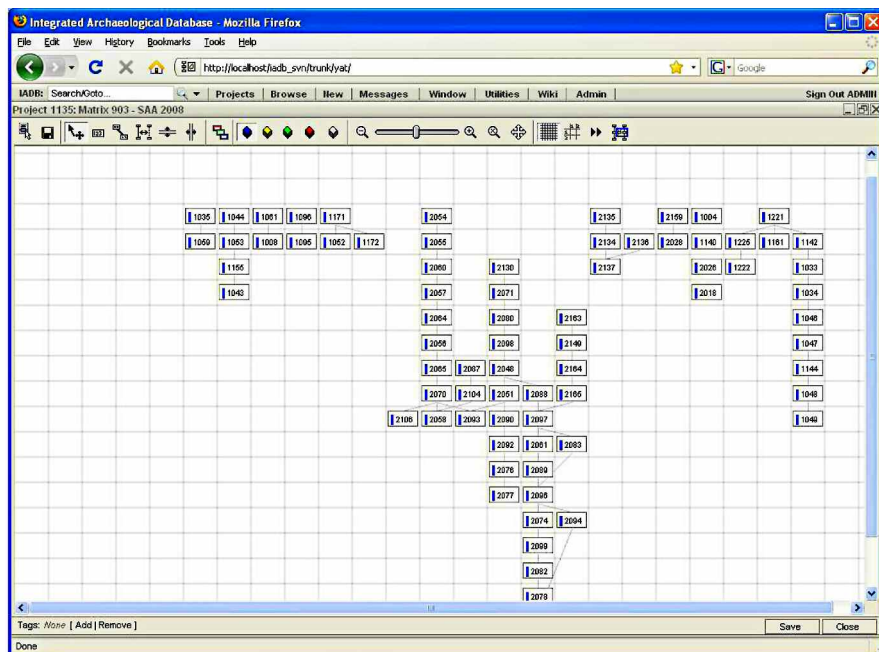


Figure 5.1. A simple diagram showing stratigraphic units (contexts) and the stratigraphic links between them.

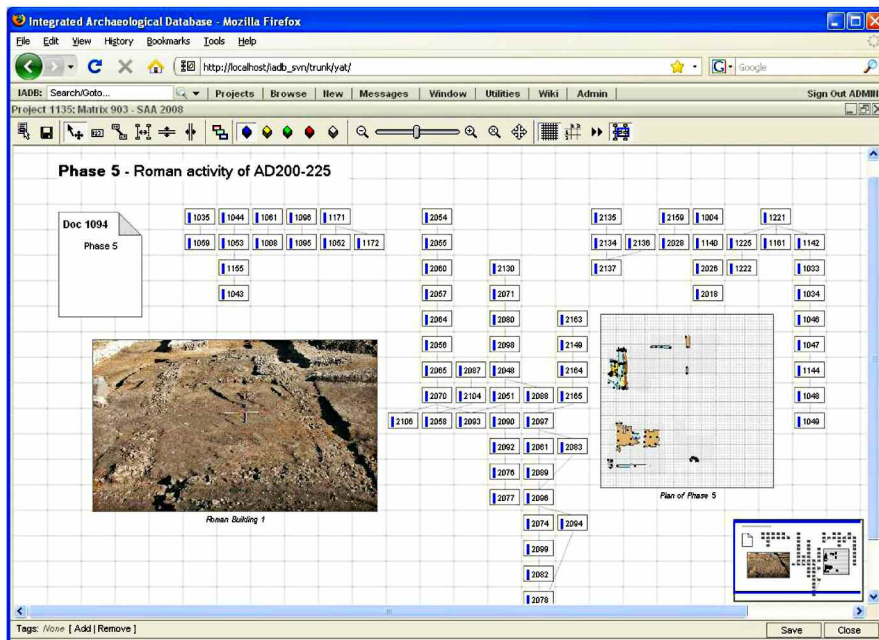


Figure 5.2. Structure diagram with added image, plan, and document.

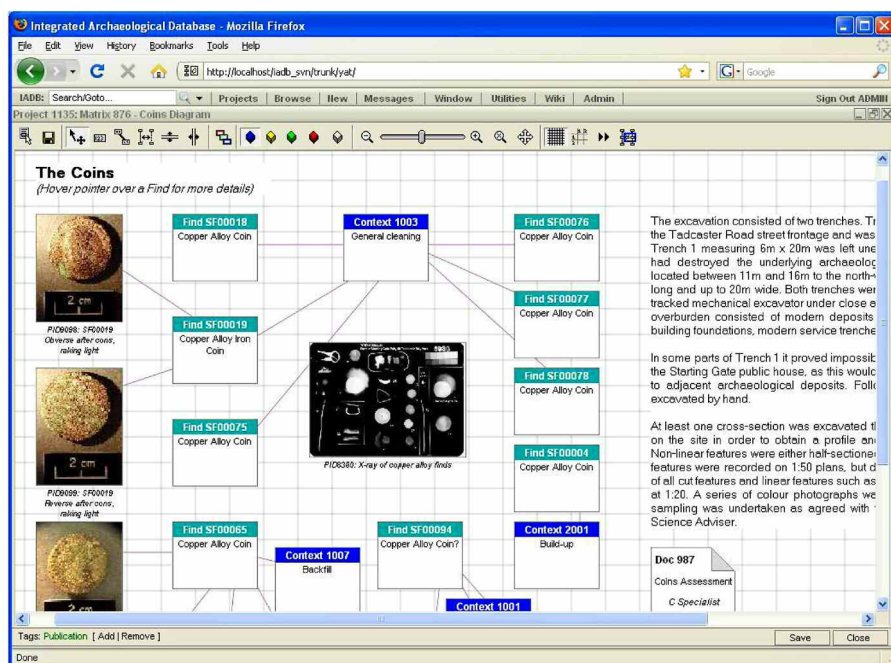


Figure 5.3. Researching a coin assemblage within a virtual research domain (VRD).

Researchers logging into the virtual research domain are presented with the interactive structure diagram as their interface to all the resources relevant to the topic of the VRD. A simple double-click on any item, whether a stratigraphic unit (context), plan, or photograph, will take them straight to that resource. For example, double-clicking on a plan will open the plan in the IADB's interactive plan browser, which is, in effect, a small-scale GIS system within the IADB enabling the detailed manipulation of individual context plans and the drawing elements within them. If permitted, researchers can add annotations and other resources to the structure diagram. They can add their contributions to the VRD documents mentioned above and see the contributions of others (Figure 5.4). They can also use the IADB's messaging facilities to communicate directly with other researchers. All of this is possible with minimal user training, while researchers who are more familiar with the system still have full access from within the VRD to all the facilities of the IADB.

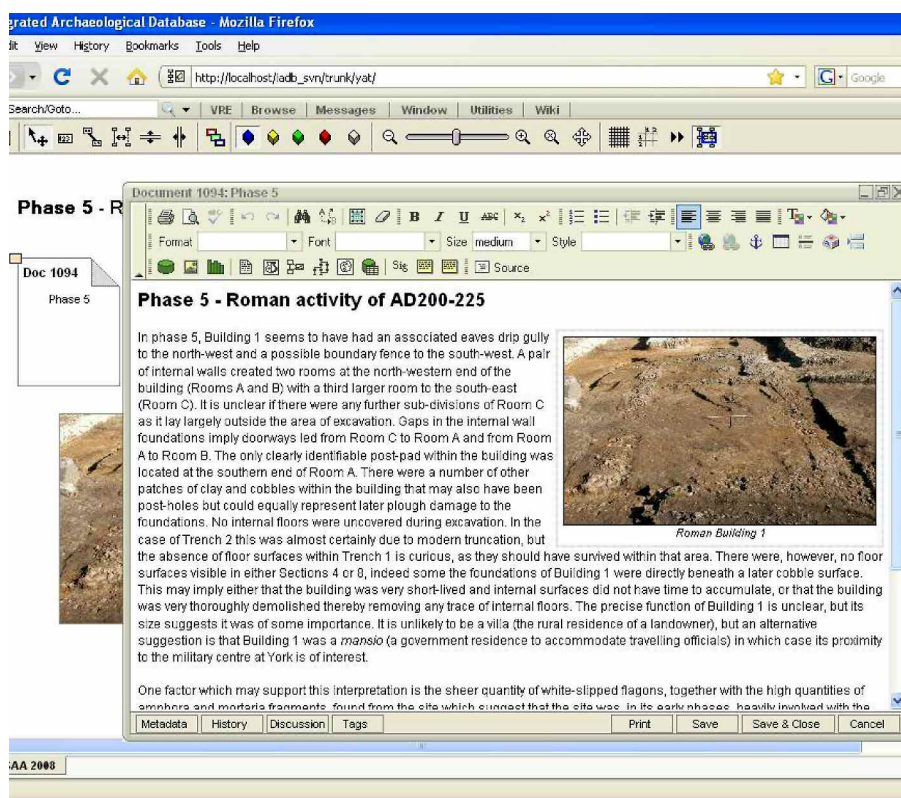


Figure 5.4. Online collaborative document editing within a virtual research domain.

The first practical use of the VRD concept during its initial development was to produce an online publication as part of the **Linking Electronic Archives and Publications (LEAP) program**<sup>4</sup> funded by the U.K. Arts and Humanities Research Council and administered by the U.K. Archaeology Data Service. The archaeologists involved in the production of this paper were based in Reading and York but were able to collaborate effectively online through the VRD, which simplified and streamlined the research process and, most importantly, fostered teamwork. The resulting paper was published in *Internet Archaeology* 21 in 2007 (Clarke et al. 2007). The data on which the paper was based were archived to the Archaeology Data Service (ADS) and are accessible from within the *Internet Archaeology* paper.

<sup>4</sup> <http://ads.ahds.ac.uk/project/leap>

Future improvements to the editing facilities within the VRD and the possible introduction of three-dimensionality will further enhance its effectiveness for researchers, while direct-to-Web publication of VRDs will allow faster dissemination of archaeological research results to the wider archaeological community.

All of this has been made possible in a web environment through the application of key web technologies which have become available over recent years. Perhaps the first of these was the use of cookies to enable preservation of state in web sessions, thus allowing server and browser to maintain a conversation. The development of the document object model (DOM), dynamic HTML, JavaScript, and server scripting languages such as PHP have together enabled the development of dynamic web pages. Techniques using the JavaScript XMLHttpRequest object have been used to provide behind-the-scenes communications between browser and server. Structured data technologies including XML and JavaScript Object Notation (JSON) facilitate the transfer of data between server and client. Taken together, these techniques have enabled the development of the IADB into a true web application with a rich user interface encompassing a wide and growing tool set.

The structure of the underlying IADB database, which we can think of as the “what” of the IADB, has changed little since the early days of the system. Most effort has gone into developing the system’s scope (the “why”) and the tools and user interface (the “how”). Expansion of the scope of the IADB—from relatively simple data management during the excavation and post-excavation analysis stages of a project through to data archiving, dissemination, and publication—have been enabled and driven by developments in the wider computing world. These include the falling costs and increased capacity of online data storage, the explosive growth of the Internet and the World Wide Web, and the move to online web publication of archaeological reports. Online publication of the results of the Silchester Town Life Project—as the work progresses, not just at the end of it—has always been seen as a key aspect of the project (Clarke et al. 2004). Developments that we might characterize as Web 2.0 have, in a similar way, both enabled and driven the development of the user interface and tool set of the IADB. Put simply, they have made it possible to do within the environment of a web browser things that it was not possible to do before.

However, development of the IADB as a web application has also raised a number of problems. For example, the question of cross-browser

compatibility is perhaps not a major issue for small-scale, closed applications used within a particular research team, but becomes highly significant when applications are opened out to a wider user base. Questions also remain over the long-term future of some of the new technologies often associated with Web 2.0—for example, Scalable Vector Graphics (SVG), which is a subset of XML designed to represent two-dimensional vector graphics, particularly in a web environment. Support for this standard from browser manufacturers and large software houses has so far proved patchy and, in some cases, short-lived, while key alternatives such as Flash remain proprietary.

Over recent years, the emphasis on user interface and its functionality seen in systems such as the IADB has led to a position where it is increasingly difficult to separate data from interface. For modern IADB projects, a simple dump of the raw data is of much less utility and value than that data when accessed through the IADB interface. Indeed, many of the tables within the IADB database are there only to support the functionality of the interface and have little intrinsic meaning or value when taken out of the interface. This has major implications for data archiving and questions of sustainability of data. Many of the issues surrounding long-term data archiving and preservation—for example, the durability of storage media and technological obsolescence (how many computers can now read 5.25-inch floppy disks?)—are well understood, if not completely resolved. On the other hand, software sustainability has tended to be viewed mainly at the platform level, which in database management terms might be Linux/Apache/MySQL/PHP (in the case of the IADB) or Windows/Access/Visual Basic, rather than at the application level of the user interface. York Archaeological Trust has digital records dating back over 30 years. It is interesting to note that those that remain most accessible today are the ones that have never been formally archived but, because they are part of the larger organization-wide database (the IADB), have been carried forward through each successive hardware and software migration and interface development. On the other hand, some records that, with the best of intentions, were “archived” to floppy disk or tape in earlier years are now largely inaccessible or indecipherable.

In developing any new system or application, it is important to keep in view the reason for doing it. What problem are we trying to address? Does the solution actually resolve it? With any new technology, hardware or software, it is easy to become blinded by its newness and lose sight of its original reason for being. An example of this can be seen in another strand of the

VERA project which aimed to improve the flow of data or information throughout the lifetime of the archaeological project. As part of this, extensive trials were undertaken to test the use of digital pens to speed up the digital capture of site context records and their accessioning into the project database. Previously, these records had been written onto pro-forma Context Recording Sheets on site and then manually transcribed into the IADB, normally as part of the post-excavation process. While the digital pens have been found to perform well from a technical point of view, it is not clear that their use has produced any significant time savings with regard to the production of digitized context records. This is particularly true when the extra resources allocated to the trials are taken into account. In fact, it can be argued that the digital pens are not addressing the real problem. The amount of actual text on a typical Context Recording Sheet is relatively small—rarely exceeding 100 words—and so transcribing this text has never been the most time-consuming part of the process. Irrespective of whatever decisions may be made in the future about the continued use of digital pens and other technologies, these issues have prompted an ongoing fundamental reappraisal of the structure and function of context records within the IADB. Recent trials have examined the use of scanned images of Context Recording Sheets (with appropriate metadata) in preference to full transcription of the context record.

Both the digital pen trials and the VRD development strands of the VERA project have highlighted another important issue with regard to the introduction of new techniques and technologies into long-running systems with well-established methodologies such as archaeological excavation projects. As mentioned above, for many years the Silchester Project, like most other excavations, has used hand-completed pro-forma Context Recording Sheets. Over time, a comprehensive system has been developed to manage and check these forms. For various reasons, the digital pens trial opted not to use digital versions of these forms, but to use free-form digital notebooks. This fundamental change to the recording system caused many problems for the management and checking of the data. For example, while it was easy enough to check what had been recorded in the digital notepads, it was much more difficult to check what had *not* been recorded, whereas empty fields on the Context Recording Sheets were easy to spot.

Virtual research environments will only be adopted into regular use if researchers feel that the solution offered truly addresses a need that they

themselves perceive. In other words, it's not enough for the solution to adequately address the problem; the problem itself has to be a real one.

In conclusion, it can be seen that the technical advances and programming developments that have made possible Web 2.0 applications such as Facebook and Google Mail have also allowed archaeological systems such as the IADB to develop into something much better than they were, say, five years ago, and this will hopefully continue. However, these same developments have also highlighted the need for a clear understanding of the problem being addressed, and the importance of detailed user needs analysis being undertaken alongside the technical development of new applications and approaches to archaeological data management.

In the two years since this paper was first written, significant development has taken place in two aspects of the IADB mentioned briefly above.

The use of the IADB as a web publication tool has been developed both within the Silchester Project and at York Archaeological Trust. By creating individual web pages as documents within the IADB, and then using the IADB to manage and publish them, the project teams were able to achieve a much closer integration of IADB data resources (such as context records, lists of finds, and the like) into the pages of the web report than was possible within the LEAP project. The recent Silchester Project web publication, "The City in Transition" (Fulford and Clarke 2010), demonstrates well this close integration of report and database publication.

The use of scanned images, or facsimiles, of Context Recording Sheets (CRS), along with appropriate metadata, in preference to a full transcription of the CRS, has been adopted as standard practice by both York Archaeological Trust and Canterbury Archaeological Trust. In contrast to the use of digital pens described above, this approach to context recording has been found to save considerable time and to enhance the overall quality of the record (Fisher and Rains in press).

When assessing the long-term significance of the VERA project (and its precursor project, OGHAM), it is clear that many of the developments that took place as part of VERA, as well as others such as facsimile context recording, which were prompted by it, have now been adopted widely by IADB users. They have fulfilled the stated aims of the VERA project by promoting lasting improvements in the processes of initial record creation and



post-excavation analysis and research. In addition, although this was not an explicit aim of the project, it can be argued that VERA has made a significant contribution to the ongoing development and refinement of approaches to Web-based archaeological report and database publication.

## REFERENCES CITED

Clarke, A., M. Fulford, and M. Rains

- 2004 Nothing to Hide—Online Database Publication and the Silchester Town Life Project. In *The Digital Heritage of Archaeology; Computer Applications and Quantitative Methods in Archaeology, Proceedings of the 30th CAA Conference*, ed. M. Doerr and A. Sarris, pp. 401–404.

Clarke, A., M. Fulford, M. Rains, and K. Tootell

- 2007 Silchester Roman Town Insula IX: The Development of an Urban Property c. AD 40–50 – c. AD 250 *Internet Archaeology* 21. Retrieved from [http://intarch.ac.uk/journal/issue21/silchester\\_index.html](http://intarch.ac.uk/journal/issue21/silchester_index.html)

Fisher, C., and M. Rains (in press)

- In press Preserving the Record—Context Recording in the Digital Age. In *Proceedings of the 38th CAA Conference, Granada, Spain, 2010*.

Fulford, M., and A. Clarke

- 2010 Silchester: The City in Transition: The Mid-Roman Occupation of Insula IX, c. AD 125–250/300. A Report on Excavations Undertaken since 1997. Retrieved from <http://www.silchester.reading.ac.uk/cit/index.htm>

JISC

- 2007 Virtual Research Environments Programme. Retrieved from <http://www.jisc.ac.uk/whatwedo/programmes/vre2.aspx> (accessed October 28, 2008).

Rains, M.

- 1995 Towards a Computerised Desktop. In *Proceedings of the 22nd CAA Conference Held at Glasgow University, Glasgow, 1994*, ed. J. Huggett and N. Ryan. BAR International Series. Oxford: B.A.R.



# iAKS: A WEB 2.0 ARCHAEOLOGICAL KNOWLEDGE MANAGEMENT SYSTEM

*Ethan Watrall*

## INTRODUCTION

Methodologically speaking, field archaeology has been relatively content with traditional practices of inquiry whose lineage can be traced back to the beginning of the nineteenth century. However, as new generations enter the discipline, there has been a trend to accept a greater range of computational methodologies. As a result, techniques such as ground-penetrating radar, magnetometry, and geographical information systems (and the computational analysis of associated data) have become more commonplace. This willingness to accept new computational techniques into the discipline, however, has not generally been extended to systems that allow field archaeologists to digitally collect, archive, and access archaeological data. This is no great surprise, as the type, amount, and sheer scale of data collected during any given archaeological project is staggering. In addition, the type of data and the way in which those data are collected can vary widely from one archaeological project to another.

While archaeologists generally agree on what might be called methodological meta-standards, the discipline as a whole lacks a set of standards for the kinds of data that are recorded and for the exact ways those data are recorded. This is not always the result of theoretical differences, but rather can arise from the specifics of where archaeologists are working (geographically speaking) and what they are working on (both temporally and culturally speaking). The kinds of artifacts recovered from a Paleolithic site in France, for example, differ enormously from those recovered from the excavation of

a historical nineteenth-century farmstead in Nova Scotia. The consequent lack of collecting standards means that many archaeological projects still rely on paper record-keeping. When this is the case, the longer an archaeological excavation has been running, the more paper is generated. As a result, it becomes more difficult for the project's archaeologists to locate and analyze specific subsets of data. This problem is greatly intensified for archaeologists not associated with the project itself who might be interested in the data being recovered. Not only are they unfamiliar with the project's specific system of data collection, data management, and data archiving, but also the data itself (in the form of reams of paper) may be located in a box halfway around the world. One might also easily argue that, in addition, many archaeologists are mainly concerned with their own research and less in how their data might be used in conjunction with other data.

This should not suggest that archaeological projects have not stored data in databases; quite the contrary (Arndt and Coulson 1985; Kvamme 1989; Seger 1995; Dibble and McPherron 1988). However, these databases are usually designed specifically for individual archaeological projects and are, as a result, quite unique. Given this, there is very little chance that one data entry and archiving system can be adapted for use by another archaeological project. Further, the systems themselves are rarely, if ever, networked (or designed for robust and elegant user interaction). This means that both non-project archaeologists and project archaeologists who are not in the same location as the database are unable to access the data. This is unfortunate, as the lack of standardized data and interoperability of data sets and data systems has made it difficult to produce the kinds of "bigger picture" socio-cultural analysis that is the hallmark of both field archaeologists and theoretical archaeologists.

The evolution of the Web, and more specifically web services and web applications, offers an opportunity to address some of these problems. In recent years, there has been an increased desire in the archaeological community to explore semantically based web services and platforms to archive, publish, analyze, and share archaeological data (Kansa and Bissell 2010; Kintigh 2006; Schloen 2001; Snow et al. 2006). It is within this context that this chapter explores a Web 2.0 archaeological knowledge management system named iAKS (Interactive Archaeological Knowledge Management System). Developed as a preliminary proof of concept project at Michigan State University's MATRIX: The Center for the Arts, Letters, and Social Sciences

Online, **iAKS**<sup>1</sup> is specifically designed to take advantage of web services and Web 2.0 technologies and platforms in order to create a robust, scalable, extensible, stable, and reusable system that solves many of the data collection, data archiving, and data analysis problems encountered by archaeological projects. The ultimate goal of iAKS is to create an interactive system that archaeologists can use not only to collect, archive, and analyze excavation and artifact data, but to access and visualize data remotely from anywhere in the world. iAKS features a very flexible setup and install model that allows archaeologists to customize the specific types of data that they might want to collect and archive. In addition, iAKS has a variety of client/server models, making it an appropriate tool for a wide variety of archaeological settings, ranging from a small-scale, single-site field school to large, multi-site excavations. Further, iAKS features a variety of connectivity models, making it suitable for projects that have regular network connectivity, those that have intermittent network connectivity, and those that have no network connectivity. iAKS is also designed with a keen sense of usability, thereby making it friendly for a broad range of both experienced and inexperienced users. Finally, iAKS features a robust data visualization system, allowing current (and future) archaeologists to browse and creatively visualize data. It is important to note that at this stage of development, iAKS is a proof of concept. It is intended to explore the aforementioned challenges within the broader context of issues such as archaeological data standards, linked data, and user experience design within the domain of cultural heritage, many of which are addressed elsewhere within this volume.

## iAKS ARCHITECTURE

The Web 2.0 “movement” has seen the development of a variety of useful and very powerful technologies that have changed the Web from a one-way information medium populated with walled-off silos of data to a platform-based information ecosystem in which data flows (relatively) freely between applications (as opposed to sites). The iAKS project harnesses this medium and its associated technologies to deliver an interactive, data-centered Web- and client-based application to help make data collection, data archiving, and analysis easier for archaeologists.

---

<sup>1</sup> <http://iaks.matrix.msu.edu>

In late 2007, Adobe introduced developers to a product named AIR (Adobe Integrated Runtime). AIR is a cross-operating system (Mac, Windows, and Linux) runtime that allows web developers to deploy rich web applications to the desktop without the constraints of a browser, using HTML, JavaScript, Flex, or Adobe Flash. A rich Internet application (RIA) is a relatively new kind of web experience that is engaging, interactive, lightweight, and flexible. RIAs offer the flexibility and ease of use of an intelligent desktop application, and add the broad reach of traditional server-side web applications. RIAs typically transfer the processing necessary for the user interface to the web client while keeping the majority of data (and possibly content) on the application server.

Applications developed using AIR differ from traditional browser-based RIAs in that no client-side installation is required. However, AIR-based RIAs must be packaged, digitally signed, and installed to the user's local file system. The advantage of this model is that access to local storage and file systems is provided; in comparison, browser-deployed applications are more limited in where and how data are accessed and stored. Additionally, given the fact that AIR applications are developed with familiar and accessible web technologies such as HTML, JavaScript, Flex, and Flash, applications can be deployed both to the desktop (with AIR) and to the browser from a single code-base. While AIR certainly is not the only platform attempting to blur the line between browser-deployed applications and desktop-deployed applications ([Google Gears](http://gears.google.com),<sup>2</sup> [Mozilla Prism](http://prism.mozilla.com),<sup>3</sup> [Fluid](http://fluidapp.com),<sup>4</sup> [Titanium](http://www.appcelerator.com/),<sup>5</sup> and potentially even [HTML5](http://www.w3.org/html/logo/)<sup>6</sup> are of particular note), it has the advantage of being intimately linked with the Adobe Flash platform, an ecosystem of mature and nearly ubiquitous technologies, architectures, and development environments.

By taking advantage of AIR, the iAKS client runs on Mac, Windows, or Linux on virtually any desktop PC; accesses the local file system and resources when a connection to the server cannot be established; and runs as a traditional in-browser web application, all from the same code base. AIR is vital to the success of iAKS because it enables operation under the wide va-

---

<sup>2</sup> <http://gears.google.com>

<sup>3</sup> <http://prism.mozilla.com>

<sup>4</sup> <http://fluidapp.com>

<sup>5</sup> <http://www.appcelerator.com/>

<sup>6</sup> <http://www.w3.org/html/logo/>

riety of system and network conditions often found on archaeological projects. In addition to the flexibility of AIR, iAKS also leverages the power of Adobe's **Flex platform**.<sup>7</sup> Flex, a rich Internet application framework based on Flash, allows iAKS to present data in detailed, multicolored graphs, charts, and tables, giving end users almost immediate feedback and analysis of their data.

The downside to the Flash ecosystem is its complicated relationship with the open source movement. In 2009, Adobe released the full specification for the SWF file format. In addition, they have removed licensing fees for Adobe Flash Player and AIR, published several additional Flash-related protocols, and made efforts to develop an API for porting Flash to new devices. Adobe Flex, however, has made more concrete strides toward open source. Adobe has released Flex under the Mozilla Public License (MPL). This includes not only the source code to the ActionScript components from the Flex software developer kit (SDK), but also the Java source code for the ActionScript and MXML compilers, the ActionScript debugger, and the core ActionScript libraries from the SDK.

## iAKS Client/Server Models

One of the main goals of the iAKS project is to create a system that can be used by archaeologists in a wide variety of settings. iAKS can be configured to operate efficiently in different situations, ranging from small archaeological projects to large, multi-site excavations. To facilitate this, iAKS is broken up into two components: the iAKS Manager and the iAKS Client. The iAKS Manager is a server-side application (written in **Ruby on Rails**),<sup>8</sup> which is used to manage users, set up projects, store data remotely, and publish to iAKS Online. The iAKS Client is the AIR-based desktop application used to record, view, and query data. The iAKS Manager/iAKS Client can be set up using five different client-server models:

- *One Client/No Server model.* This configuration leverages the power of AIR and allows the iAKS client to store all project information and data on the user's local hard drive. This model is ideal for small projects or educational settings where minimal multiuser collaboration is needed.

---

<sup>7</sup> <http://www.adobe.com/products/air>

<sup>8</sup> <http://rubyonrails.org>

- *One Client/One Server Model.* This is the most basic of the client/server models that incorporates the iAKS Manager system into the configuration. With the server installed, this model allows the client to store data remotely, create multiple users for projects, and share data with other applications through an API. Both the iAKS Manager and the client application store project data in the same format, making it easy to migrate from local storage to server storage (Figure 6.1).

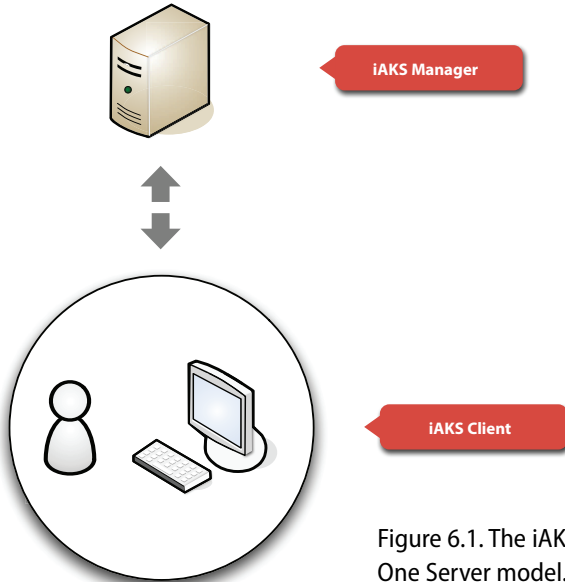


Figure 6.1. The iAKS One Client/One Server model.

- *Many Clients/One Server model.* This client/server model allows iAKS to manage multiple client applications accessing the same set of project data. This facilitates multiple users being able to access the system through different clients from different locations. iAKS manages each running client and syncs to the project that the user selects upon logging in to the system. This configuration also gives iAKS the ability to manage a pool of client applications, assigning roles and permissions for a particular project (Figure 6.2).
- *Many Clients/Many Projects/One Server model.* In this configuration, the iAKS Manager tracks multiple projects, each with multiple client applications. This is the most complex setup and would be used on a large



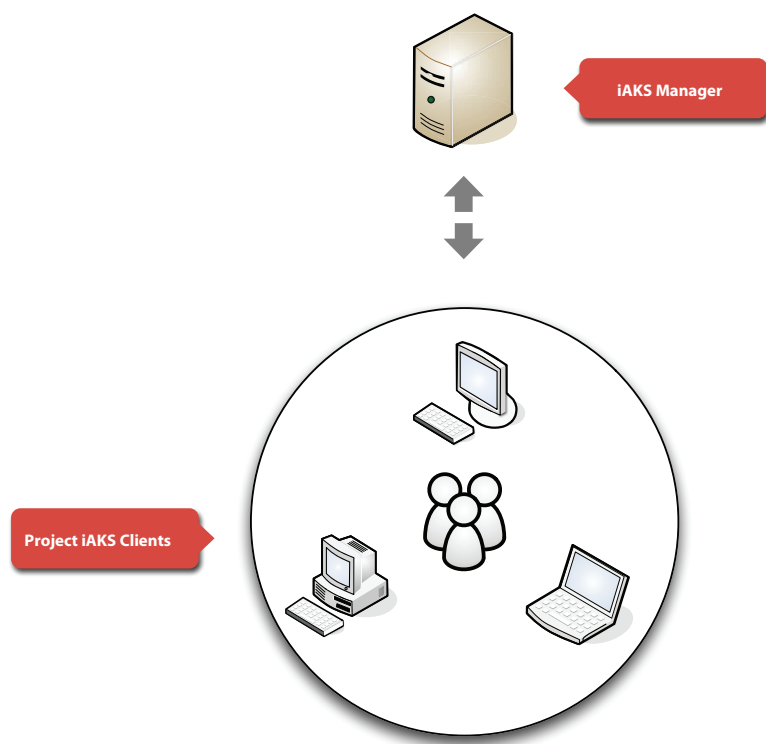


Figure 6.2. The iAKS Many Clients/One Server model.

project that might span multiple excavation sites and have many users accessing the data both locally and from remote locations via the iAKS Client (Figure 6.3).

Given the nature of the platform upon which iAKS is developed (Flex/Flash and AIR), users are able to run the client from within a web browser and have remote access to all their projects and data without the presence of the iAKS Client. This can be a flexible option for teams that need remote access, or a starting point for an iAKS-based online service that offers a “lite” version of the full installation available only through the browser. iAKS On-line is managed from the iAKS Manager server admin tools and, as such, is not available when users are employing the One Client/No Server model discussed above.

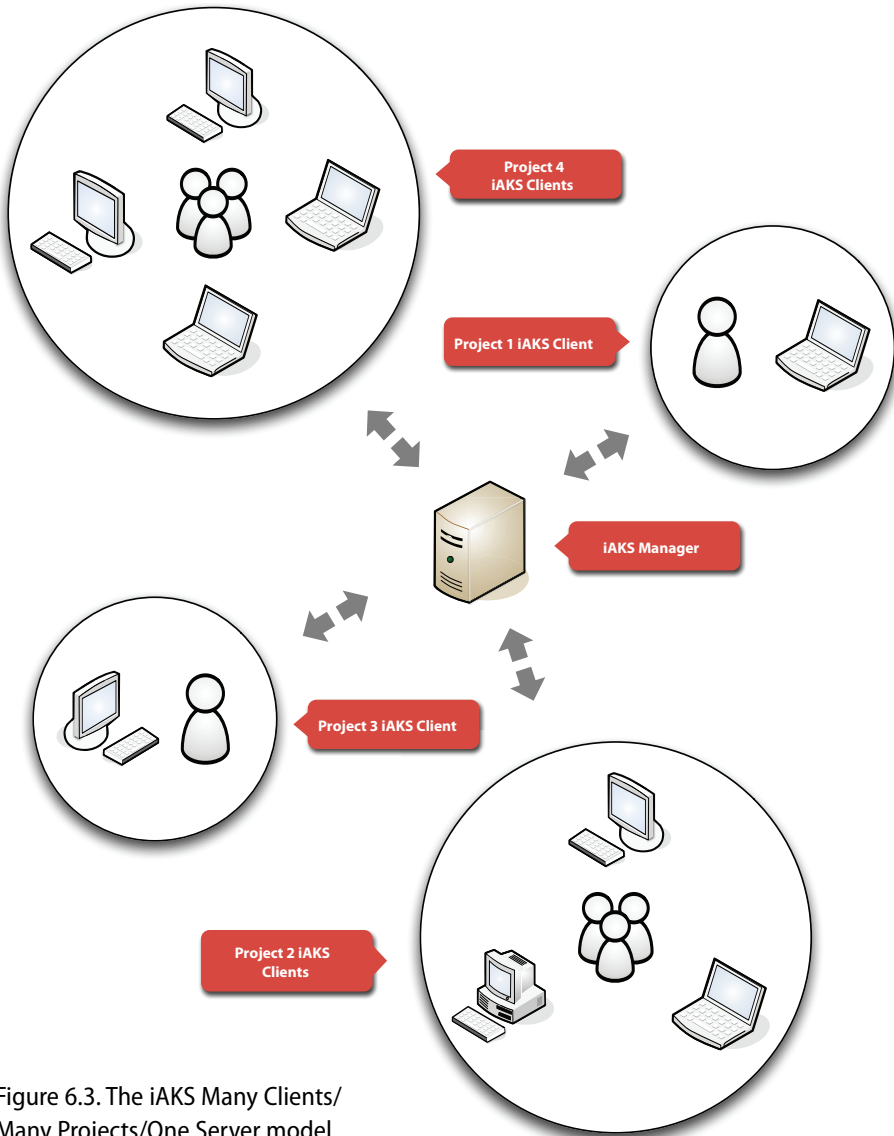


Figure 6.3. The iAKS Many Clients/Many Projects/One Server model.

## iAKS Setup Model

One of the most daunting hurdles in designing a system to support the creation, capture, storage, curation, and dissemination of archaeological information that is applicable in a wide variety of settings is the lack of standardization

among the types of data that different archaeologists collect. Taking an example from lithic analysis, one archaeologist may record the length, width, thickness, weight, and type of a tool. However, another archaeologist may collect these measurements and many others as well (such as platform width, platform thickness, number of dorsal scars, blank type, and retouch type, among others). In addition, the way in which two archaeologists undertake the measurement of the same characteristic may differ. As a result, a knowledge management system for field archaeology with hard-coded methods of data entry is wholly unsuitable for use across a wide variety of archaeological settings.

iAKS solves this problem by including a flexible setup model that allows archaeological projects to configure exactly the types of data (and the associated methods of data entry) they would like to include in their iAKS system. For example, an excavation of an eighteenth-century historic fort in England would have absolutely no need to collect and archive data on stone tools yet might have the need to collect, archive, and analyze data on pottery. An excavation working on an early Neolithic site in the Egyptian Western Desert, by contrast, would have nowhere near the same kind of need to collect, archive, and visualize data on pottery but would have a very pressing need to collect, archive, and visualize a vast array of data pertaining to stone tools. As a result, the knowledge management for the two projects would look radically different.

The core feature of the iAKS setup model is a robust library of data types and their associated data entry mechanisms. The libraries have been created based on consultations with different archaeologists working in a variety of areas (geographical, temporal, and methodological) in order to get a deep sense of what kinds of data archaeologists collect, and how they might want to enter, archive, and visualize that data in iAKS.

During the setup process, which is facilitated by a user-friendly wizard, the project archaeologists can pick and choose the types of data (and associated data entry mechanisms) that are important to their excavation, all of which are drawn from the aforementioned library. The result is that the project archaeologists can directly customize exactly how their iAKS clients are configured. During the setup process, archaeologists are also able to take a modular approach to setup. For example, one archaeologist may include a “Ceramic Data Module” (and then customize it further with specific data types and associated data entry mechanisms), while choosing not to add the

“Lithic Data Module” and the “Faunal Data Module,” based on their excavation’s specific needs. This setup model also configures the iAKS client interface itself, creating a custom user experience for each project.

It is very important to note that the setup model is identical regardless of the client/server model. The process by which a large archaeological project (with multiple sites) configures its iAKS clients is exactly the same as that of the individual archaeologist working on a smaller scale (a graduate student working on a small portion of an archaeological project, for example). The only difference is that during the setup and configuration of projects using the iAKS Manager, the server is connected, whereas the setup and configuration process for a single client install is completely self-contained.

The setup and install process creates an XML file that sits either on the server (if the archaeological project is using the client/server model) or on the client (if the archaeological project is using the single-client model). The XML file essentially acts as a configuration file for that project’s iAKS interface. As a result, whenever a new iAKS client connects to the server of a project that has already been configured, the server pushes the XML file to the client and configures the interface, thereby making it unnecessary to go through the lengthy setup process anew unless there is a need for a new configuration. In a single-client setup (where, as mentioned before, iAKS is set up on a single machine without connection to a central server), the XML file can be exported and moved to another machine, thereby enabling another new iAKS installation.

Ultimately, the goal of the setup model is to create an installation and configuration process that is flexible enough to result in a system that is both suitable and easy to use for the widest possible range of archaeological projects.

## iAKS Distribution Model

The final, but no less important, aspect of iAKS is the manner in which it is distributed.

The final, 1.0 build of iAKS will be released as open source software under the GPLv2 license. This will allow additional developers to extend and alter iAKS as their needs require. It is very important to note that Adobe Flex, the development environment in which iAKS is created, is itself open source and available for free. In 2007, Adobe released Flex under an open

source Mozilla Public License (MPL). This empowers both open source and commercial developers to extend and enhance the Flex framework to suit their own needs and to contribute to the evolution of the Flex framework. In addition, Adobe distributes the Flex SDK (the tool used to author Flex applications) at no cost to the public.

In addition, several additional distribution models exist for iAKS (some of which are currently in development and some of which are still in the proposal phase).

- ▶ The iAKS Client and iAKS Manager would be distributed freely and user hosted. In much the same way as the WordPress blogging system, an organization would be formed to host and distribute the software as well as build and host a user community.
- ▶ Free hosted service would be offered to those who might not have a server appropriate for mounting the iAKS Manager. A business model could be built around a scenario that offers upgraded services for a fee. This is not dissimilar to the business model of [37Signals](http://37signals.com)<sup>9</sup> or [Wordpress.com Premium](http://wordpress.com).<sup>10</sup>
- ▶ As mentioned before, the release on an iAKS API would facilitate the extension of the system as well as encourage the use of iAKS data in other web applications.

## iAKS Online

In addition to the iAKS Client and iAKS Manager, the iAKS project will ultimately include a robust online community-based site that will act as a central (and semi-public) repository into which iAKS users can upload the archaeological data from their local iAKS installation. Once uploaded by individual iAKS users, other community members can either search and browse the data online or download and import the data into their own local iAKS installation for standalone analysis or inclusion into an existing iAKS data set. For the purposes of security, iAKS online will have a two-tiered system of access. The first tier will be open and accessible by the general public and will feature data that is filtered by the original contributor.

---

<sup>9</sup> <http://37signals.com>

<sup>10</sup> <http://wordpress.com>

This way, sensitive data, such as site location or exact artifact provenience, can be hidden from the general public. The second tier will feature full and complete data sets and will only be accessible by those professional archaeologists who have registered with the site and whose credentials have been verified.

The interoperability of data between an iAKS installation and iAKS online, as well as between individual iAKS installations, is made possible by the fact that all iAKS data are stored as XML. The structure and logic of the XML used to format and store iAKS data is based loosely on the ArchaeoML schema developed by David Schloen and used by the University of Chicago **OCHRE Project**.<sup>11</sup> This is important, because it allows iAKS to become, despite its being primarily a knowledge management and data visualization tool, a vehicle for integrating and cross-referencing different sets of archaeological data between disparate iAKS installations or other archaeological web applications.

## CONCLUSION

It is extremely important to note that while iAKS is a project steeped in technology and web application development methodology, it is simply a vehicle to allow archaeologists to better collect, store, share, and visualize archaeological data. Most importantly, however, iAKS, through XML-based data standards, encourages and facilitates the interoperability of data, allowing archaeologists to engage in a level of research synthesis that is currently very difficult.

## REFERENCES

- Arndt, A., and W. D. E. Coulson  
 1985 The Development of a Field Computer for Archaeological Use at Naukratis in Egypt. *Journal of the American Research Center in Egypt* 22: 105–115.
- Kvamme, K. L.  
 1989 Geographic Information Systems in Regional Archaeological Research and Data Management. *Archaeological Method and Theory* 1: 139–203.

---

<sup>11</sup> <http://ochre.lib.uchicago.edu>

Seger, J. D.

1995 Lahav DIGMASTER. *The Biblical Archaeologist* 58/2: 116.

Dibble, H. L., and S. McPherron

1988 On the Computerization of Archaeological Projects. *Journal of Field Archaeology* 15/4: 431–440.

Kansa, E. C., and A. Bissell

2010 Web Syndication Approaches for Sharing Primary Data in “Small Science” Domains. *Data Science Journal* 9: 42–53.

Kintigh, K.

2006 The Promise and Challenge of Archaeological Data Integration. *American Antiquity* 71/3: 567–578.

Schloen, J. D.

2001 Archaeological Data Models and Web Publication Using XML. *Computers and the Humanities* 35: 123–152.

Snow, D. R., M. Gahegan, C L. Giles, K. G. Hirth, G. R. Milner, P Mitra, and J. Z. Wang

2006 Cybertools and Archaeology. *Science* 311/5763: 958–959.





## CHAPTER 7

# USER-GENERATED CONTENT IN ZOOARCHAEOLOGY: EXPLORING THE “MIDDLE SPACE” OF SCHOLARLY COMMUNICATION

*Sarah Whitcher Kansa and Francis Deblauwe*

## INTRODUCTION

The term “social media” describes “a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, which allows the creation and exchange of user-generated content” (Kaplan and Haenlein 2010) and gives people the ability to interact and socialize on the “Social Web.” Unlike traditional media (such as newspapers and television), social media are accessible and inexpensive: anyone can produce and distribute information, whether a single comment or a complete thesis, simply by using the Web. Publicly available information published by end users is called “user-generated content,” and it has emerged as a new way of communicating archaeology. Its forms include blogging, podcasting, wikis, image-sharing, databases, and “tweeting,” and the low barrier-to-entry for most of these technologies means that many archaeologists are experimenting in new and open forms of communication.

Content produced by end users is a hallmark of Web 2.0. Rather than being limited to the more “authoritative” voices of traditional media, social media users now have tools for global, “bottom-up” forms of dissemination. With the removal of the “gate-keeper” (that is, editorial oversight of print publication), any and all content can be shared. Thus, user-generated content runs the gamut from idle gossip to research, and can be perceived as a mere distraction or as reliable, citable content. The distinction between the two

largely depends on the perspective of the audience: one person's "noise" will be another's data. This chapter explores some uses of social media in archaeology and discusses how the democratization of information-sharing in this way impacts archaeological research and communication.

## THE SCOPE OF USER-GENERATED CONTENT IN ARCHAEOLOGY

Many archaeologists participate in online forum discussions or closed email lists, and many have some form of Web presence, whether their own blog or simply an online curriculum vita. User-generated content is a good option when scholars wish to rapidly and easily share information with a broad audience (see also Harley et al. 2010). Sharing research results quickly and widely, even if such practice does not necessarily benefit from peer review before posting, has the potential to communicate research faster. While sharing can be fast, its impact can be hard to gauge. Audience perception of the quality, reliability and legitimacy of such communications may vary widely. For these reasons, peer-review publication remains firmly entrenched, and social media pose little threat to this traditional approach to vetting quality. More often, social media are used to *complement*, rather than replace, peer-review publication.

As discussed below, social media include a diverse array of email lists, social networking platforms, blogging and micro-blogging platforms, and discipline-specific "portals." For example, a search for "archaeology" in the photo-sharing site **Flickr** returns over 81,000 reusable images.<sup>1</sup> Another simple service is WordPress, which offers easy-to-use blogging software and is the platform of choice for many excavators who want to report from the field. Other services try to provide more "depth" around a topic or project. Examples include the **Pylos Regional Archaeological Project: Internet Edition**,<sup>2</sup> which offers a site gazetteer and numerous finds databases, including images; and the **Digital Archaeological Atlas of the Holy Land (DAAHL)**,<sup>3</sup> a digital atlas that includes site maps, photographs, and artifact descriptions from what the developers hope will eventually be tens of thousands of sites.

---

<sup>1</sup> <http://www.flickr.com/search/?q=archaeology&l=cc&ss=0&ct=0&mt=all&w=all&adv=1>

<sup>2</sup> <http://classics.uc.edu/prap/>

<sup>3</sup> <http://daahl.ucsd.edu/DAAHL/>

Content circulating through these various systems includes musings, early-stage research ideas, questions, news, career advice, conference presentations, and preprints of formal publications. Some research shared this way can see lively debate and evaluation within the research community once posted. Advocates for archaeological use of social media view such informal online discussions as very valuable to informing and improving on subsequent, more formalized, peer-review publications.

All of this activity involving social media represents critical experimentation in how to best use the Web to aid archaeological communication. This experimentation is ongoing, and even more “casual” tools such as blogs and wikis are useful examples of how researchers are attempting to use the Web to improve scholarly communication and supplement traditional print publication.

## CASE STUDIES FROM ZOOARCHAEOLOGY

This chapter draws on case studies from zooarchaeology to help illuminate how and why social media are playing an increasing role in professional communications. Zooarchaeology serves as a useful example for a number of reasons. It has a highly active research community and has a 700-member international organization, the International Council for Archaeozoology (ICAZ), whose members meet and communicate regularly. The materials (animal remains) and methodologies are global in scope; scholars working in Europe, for example, may struggle with similar questions and methodological challenges as their colleagues in South America. Finally, zooarchaeology overlaps with other disciplines such as ecology, geology, paleontology, and biodiversity, so it benefits from tools that facilitate scholarly exchange and discovery of research content beyond its disciplinary boundaries. The following discussion highlights some of the various social media the zooarchaeological community employs, the needs that these tools are meeting, and the impact they are having on scholarly communication within the field of zooarchaeology and beyond.

### The ZOOARCH Email List

**ZOOARCH**<sup>4</sup> is an email list dedicated to zooarchaeology-related discussion that currently reaches over 1,000 subscribers worldwide. ZOOARCH is an

---

<sup>4</sup> <https://www.jiscmail.ac.uk/cgi-bin/webadmin?A0=zooarch>

active venue for conversation. Because messages are delivered immediately over email, posts are “pushed” to the user and have a far greater chance of being noticed than purely Web-based forums, which have a “pull” interaction model (requiring the user to actively visit a website). ZOOARCH posts that provide a link to an item in BoneCommons (discussed below) see over 100 views in the first few hours after the email message arrives in subscribers’ inboxes (observed using BoneCommons’s view counter).

Questions, which come to the list almost daily, involve a wide diversity of needs among the zooarchaeological community. These can range from requests for publications to identification of “mystery bones,” to people searching for lost email addresses. With approximately 300 more members than ICAZ, the list is reaching beyond the traditional zooarchaeology research community to include others interested in the subject (such as contract archaeologists, museum professionals, and hobbyists). Nearly all questions that come to the ZOOARCH list are answered, usually by more than one person. This type of service is particularly beneficial for addressing “long tail” questions—that is, idiosyncratic queries that do not easily fall into a general category. A researcher can pose a question to a whole community of people most likely to have the answer, and with just a few responses, has made some helpful steps forward in his or her research. This is a much faster way of getting information (and, most importantly, *current* information) than trawling publications for answers. Some may argue that “trawling publications for answers” is beneficial to research, giving researchers time to ponder their steps, and to make serendipitous findings that lead in different directions that may not have been considered at the outset. But the same applies to large user communities like ZOOARCH, where a question posed to a wide group of colleagues may elicit unexpected answers that lead in new research directions. These new forms of communication are extremely important in helping further research. Like conferences, they provide access to a community of scholars who might have insights—only now that community is much larger and is accessible immediately (one need not wait until the next conference).

Though demographic statistics are not available on those who post to ZOOARCH, our experience with the list suggests that junior people are doing more of the questioning, while senior researchers are providing more of the authoritative answers. As with other “knowledge management systems,” it seems that the list enables junior researchers to benefit from the advice of

more experienced scholars. Senior scholars have less need for ZOOARCH-type tools because they have less to learn from them (see related discussion in Haas et al. 2007). They also tend to have different channels for finding answers to questions, relying on more direct and targeted communications with other senior colleagues who are also domain experts (Markey 2007), rather than broadcasting a question to a wide community. Though senior scholars may find questions in ZOOARCH at times distracting, their participation in some discussions suggests that they see some value in this form of communication, even though participation can sometimes be a time sink (Harley 2010: 71–72).

## BoneCommons

The Web plays a limited direct role with the ZOOARCH list, which remains rooted within the world of email communications. Nevertheless, the ZOOARCH list helps to illustrate how multiple communication channels and technologies can complement one another. **BoneCommons**<sup>5</sup> was developed in 2006 by the **Alexandria Archive Institute**<sup>6</sup> as an open access Web-based system to complement the ZOOARCH list and help advance communication within the global zooarchaeological community. It was conceived as a community hub, where people would gather and relevant content could find them, instead of expecting them to go out and find it (Figure 7.1). The original site was a forum aimed at facilitating discussion and contact among zooarchaeologists worldwide by offering a place to post papers, ideas, images, questions, and comments.

Soon after its inception, the site became associated with the ZOOARCH email list in order to provide a place where people who wanted to post materials to ZOOARCH could upload attachments, which are not allowed on ZOOARCH. This relationship proved extremely helpful to the zooarchaeological community by facilitating discussions around an image. The flow of information between BoneCommons and the ZOOARCH list is shown in Figure 7.2: here a researcher posted a photo on BoneCommons and then, in an email to the ZOOARCH list, placed a link to that photo. The researcher's question was discussed, and another researcher pointed her to an "expert" colleague best qualified to answer the question.

---

<sup>5</sup> <http://alexandriaarchive.org/bonecommons/>

<sup>6</sup> <http://alexandriaarchive.org/>

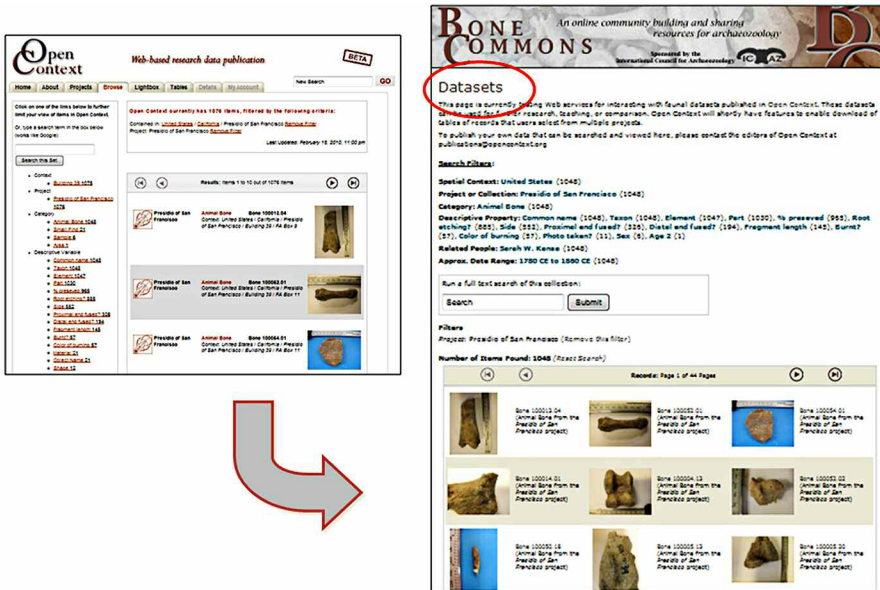


Figure 7.1. Custom feeds of zooarchaeology-related data flowing from Open Context into BoneCommons.

Though useful for some years, the BoneCommons forum became fraught with spam and security issues, making it too costly to maintain. In 2009, the site was remodeled with new software that offered better functionality, design, and security. The new BoneCommons is built with free, open-source content management software offered by **Omeka**.<sup>7</sup> All items in BoneCommons have clear citation and standard Dublin Core library metadata and protocols so that they can be archived appropriately. While adding archiving and citation, which make the new BoneCommons system a more permanent scholarly resource, site developers also wanted to continue offering tools for users to share content easily. Fortunately, Omeka offers a suite of plug-ins to allow different levels of user-generated content. BoneCommons employs the “Contribution” plug-in, which provides a simple form that users can fill out to upload content.

Before its remodel, BoneCommons was predominantly used to post images linked to ZOOARCH questions. In the year following the remodel, it has now become more a place for scholars to share their publications and

<sup>4</sup> <http://omeka.org/>

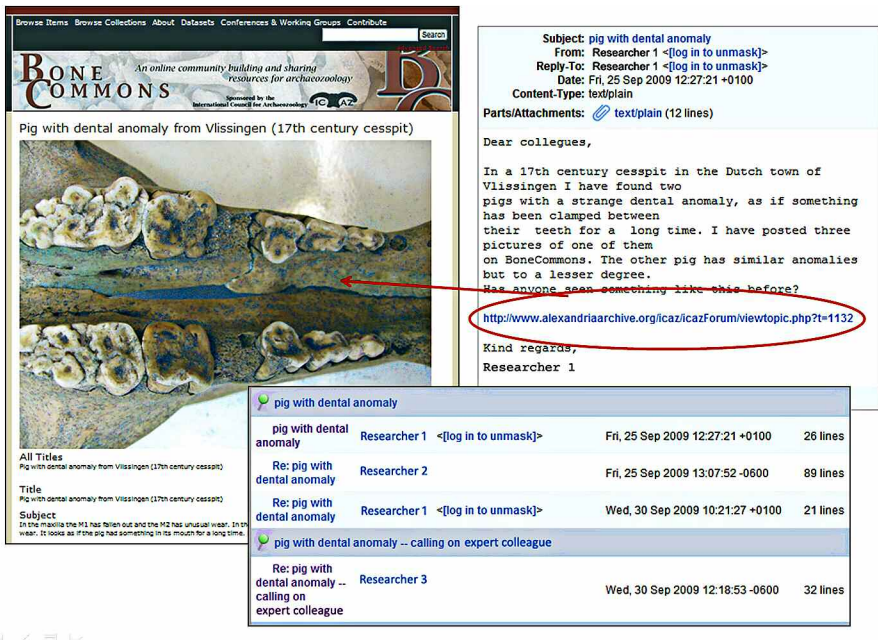


Figure 7.2. Diagram showing the flow of information between the ZOOARCH email list and BoneCommons. *Clockwise from upper left:* Researcher 1 posts an image and description on BoneCommons and then posts a question to the ZOOARCH list with a link to the image in BoneCommons. A discussion ensues on the ZOOARCH list between three researchers. In the end, Researcher 1 is pointed to a specific colleague who is an expert in the domain for further advice on the matter.

presentations. The more “professional” appearance and tools offered by the new site may account for this change; people feel more comfortable sharing their work in a more scholarly-looking site with clear licensing, citation, and archiving, as opposed to a more casual forum setting. However, the community has also seen a recent shift in the practice of posting conference presentations online. After the 2006 ICAZ international conference, titles, authors, and abstracts of conference papers and poster presentations were made available on BoneCommons. All presenters were given the option of posting their communications on BoneCommons, but only a handful chose to share in this way, opting instead to wait years for print publication. The administrator of BoneCommons (Sarah Whitcher Kansa, a co-author of this chapter) reported receiving requests over the years from users worldwide who were seeking presentations and full papers associated with the abstracts and other

posts listed in BoneCommons. These requests help illustrate a demand for certain forms of content and show that researchers were using the Web to locate relevant resources. But until recently, when researchers began more actively to share presentations and papers with BoneCommons, that demand languished unmet.

By the next ICAZ international conference, in August of 2010, the community's perception of sharing research openly on the Web had clearly changed. Of more than 700 oral and poster communications, over 100 were posted to **BoneCommons** in the four weeks following the conference.<sup>8</sup> This enthusiastic response has to do in part with the conference organizers themselves requesting that presenters share their work on BoneCommons. It also has to do with the medium of communication. Poster presenters, in particular, have readily agreed to share PDF files of their posters, largely because they may see that genre as a smaller, completed project and, if they do publish it in print, it will take the form of a more detailed publication (75 percent of the shared work on BoneCommons from the ICAZ 2010 conference are poster presentations). Furthermore, those sharing their work readily agreed to a Creative Commons Attribution-Share Alike license, which was set as the default for sharing conference communications on BoneCommons. Thus, it seems that, as in print publication, people tend to accept the default copyright policy proposed by the publisher.

Posting behavior since the 2010 conference also highlights that, in the four years since the 2006 conference, there has been a shift in researcher perception of using the Web for research. Many now embrace the medium as a way to communicate their work in one way or another, whether in the form of sharing PDFs of publications, commenting on another researcher's work, or responding to a question. The change in BoneCommons use took some years to occur, but it is important to note because it demonstrates that the perceived barriers to adopting Web technologies in the "static" world of academia are, in fact, not insurmountable.

## The Zooarchaeology Social Network

Launched in late 2009 by James Morris, the **Zooarchaeology Social Network**<sup>9</sup> (Figure 7.3) provides a "private" space for communications that are

---

<sup>8</sup> <http://alexandriaarchive.org/bonecommons/items/browse/tag/presentation+shared>

<sup>9</sup> <http://zooarchaeology.ning.com/>



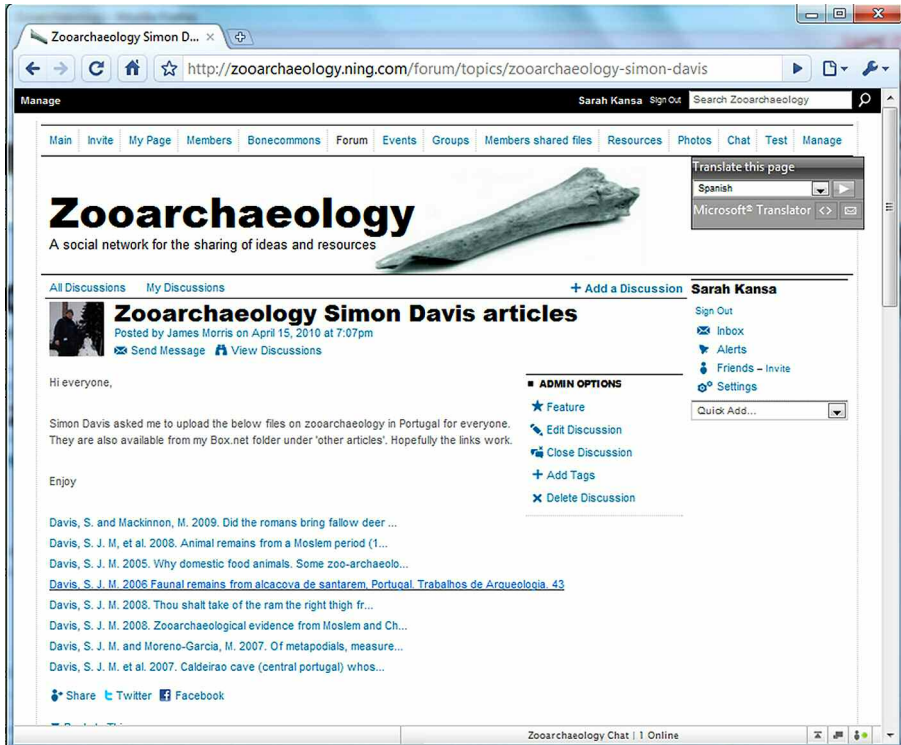


Figure 7.3. Screen-shot of the Zooarchaeology Social Network.

less well accommodated by the ZOOARCH email list or the completely open access BoneCommons site. At the time of this writing, the network has 701 members, drawing from the same community as ICAZ and the ZOOARCH email list. However, though the Zooarchaeology Social Network has features that mirror some of the other current web resources for zooarchaeology, it offers enough new tools that address some of the outstanding needs of the user community. Specifically, Zoobook has proven to be an ideal place for members to post publications that others cannot access or share publicly due to copyright restrictions. This is not something that could be done on an open access site like BoneCommons, so there is a clear need for closed networks at certain times. The site also provides a social networking component that some people enjoy, offering tools to help users build their personal profiles, start interest groups, and post comments.

## Other Media: Facebook and Twitter

Mention Facebook or Twitter at a conference and many researchers will roll their eyes. A few of those same people will turn around shortly thereafter and post an update to their Facebook page or “tweet” about the session they are attending. Social media like Facebook and Twitter may be extremely distracting and, along with blogs, email, and other messaging tools, take precious time away from our research. However, they also can be important broadcasting tools that can fill a niche in scholarly communication, sharing items of interest via links and brief comments. Because of the ease of posting and viewing tweets, Twitter is for many people the best way to keep abreast of the latest news as well as maintain weak, but important social bonds with distant colleagues.

In spite of its casual appearance, Twitter is a primary means of communication for the most up-to-date information and announcements (job postings, publications, and the like). Even major grant foundations are using Twitter and Facebook for reaching out to the broadest audience possible. Individual scholars (particularly the younger ones), academic institutions, and museums are also jumping on the Social Web bandwagon to extend the reach of their scholarship:

Ten years ago only birds tweeted . . . now even museum staff are doing it. Embracing new forms of communication will be vital if we are to keep collections alive. I think we’re entering a new phase of collaboration, with user-generated content increasingly shaping interpretation. The fact that repatriation and disposal are so contentious means real objects and specimens still matter, but it’s symbolism and stories that fire people up. With sustainability as our watchword, collections will have to pay their way. The best museums of the future will share their stuff or dispose of it (Peter Brown on the [Heritage Key website](#)).<sup>10</sup>

Comments on Twitter or discussion threads on an email list like ZOOARCH are ephemera; there is little expectation for these conversations to be cited. Other venues, like BoneCommons, provide basic bibliographic metadata for their content to facilitate citation. One of the key distinctions

---

<sup>10</sup> <http://heritage-key.com/world/future-archaeology-12-expert-predictions-decade-ahead>

between ephemeral conversations and semi-formal posts on BoneCommons is that the former provides *comments* on a topic, while the latter makes more of a public *statement* (taking a formal position on a topic).

Interestingly, digital preservation efforts are starting to capture more and more ephemera on the Web. Starting in late 2010, the Library of Congress now archives every Twitter tweet since Twitter's inception in 2006 (see the [LOC blog post](#) on this topic).<sup>11</sup> "The Twitter archive . . . will be mind-numbingly complete. Everything from reactions to the uprising in Iran . . . to your roommate's two-sentence analysis of [the 2009 film] *Hot Tub Time Machine* will be saved for posterity. Which is, from a historian's perspective, historic" (Beam 2010). This comes back again to the long tail of scholarly research: what one person calls "noise" is exactly the information another person is seeking.

## ARCHAEOLOGICAL BLOGGING

"Blogging: Never before have so many people with so little to say said so much to so few." ([Despair.com](#))<sup>12</sup>

Blogs are a different case than the tools discussed above because, rather than being an open space for participation by many, blogs are channels that people own for broadcasting to a targeted audience, and they are primarily solo endeavors. The more influential blogs are typically aligned around a subject or theme. Being a more mainstream and "newsy" medium for communicating about archaeology, blogs can serve as a useful means for bloggers to reach out to a broader community beyond their specific discipline. Blogging is very common today, particularly with the rise of Twitter, which can be seen as a form of micro-blogging. In fact, blogging and tweeting have provided tools for increased participation by enthusiasts in public conversations about archaeology. While our discussion focuses on professional blogging, we must stress that, using the current tools available for measuring trends it can be difficult to distinguish between professionals and enthusiasts, so participation by "nonprofessionals" might account for some of the changes in blogging that we see over time.

---

<sup>11</sup> <http://blogs.loc.gov/loc/2010/04/how-tweet-it-is-library-acquires-entire-twitter-archive/>

<sup>12</sup> <http://www.despair.com/blogging.html>

There is no complete master list of archaeological blogs.<sup>13</sup> People tend to discover them through browsing and reading around the Web, although word of mouth also plays a big role. Archaeology is a popular, if often misunderstood, presence online. If the Web as a whole reflects society's interests and obsessions, archaeology appears to be losing a bit of ground: **Google Insights for Search**<sup>14</sup> (Figure 7.4) shows a steady decline in the number of searches for “arch(a)eology.”<sup>15</sup> Rather than a decline in interest in archaeology, this trend may simply show that other topics are on the rise.

Evidence suggests that non-academic domains for archaeology blogs are on the rise. The clustered metasearch site **Yippy**<sup>16</sup> (formerly called Clusty), allows for a rough check (Figure 7.5).<sup>17</sup> The commercial top domain has the most archaeology-related content and has even increased its share in one year and a half at the expense of the academic Web.

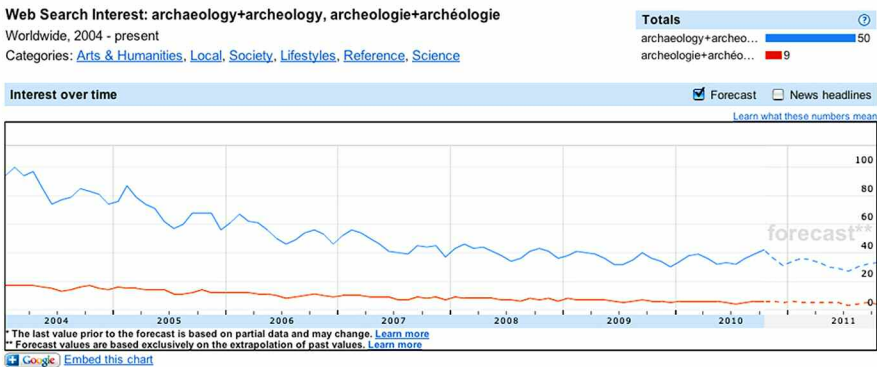


Figure 7.4. A search in Google Insights for Search shows that archaeology-related domains make up a smaller proportion of the Web over time.

<sup>13</sup> One of the more successful attempts can be found at About.com ([http://archaeology.about.com/od/blogs/Archaeology\\_Blogs.htm](http://archaeology.about.com/od/blogs/Archaeology_Blogs.htm)).

<sup>14</sup> <http://www.google.com/insights/search/>

<sup>15</sup> We also included “archeologie/archéologie,” the term used in French and Dutch to establish that this is not an Anglo-Saxon-only trend.

<sup>16</sup> <http://search.yippy.com/>

<sup>17</sup> These are rough estimates of the presence of the term “arch(a)eology,” not exact measurements of predominantly archaeological websites (March 2009 data from Deblauwe 2009; current data obtained online [<http://search.yippy.com/>] on September 13, 2010).

Blog-related web statistics can be obtained using [Google Blogs Search](#).<sup>18</sup> This graph (Figure 7.6), too, is an indication of the relative “strength” of different domains rather than an absolute measurement. Even more so than for the Web in general, blogs mentioning archaeology use the “.com” root domain.

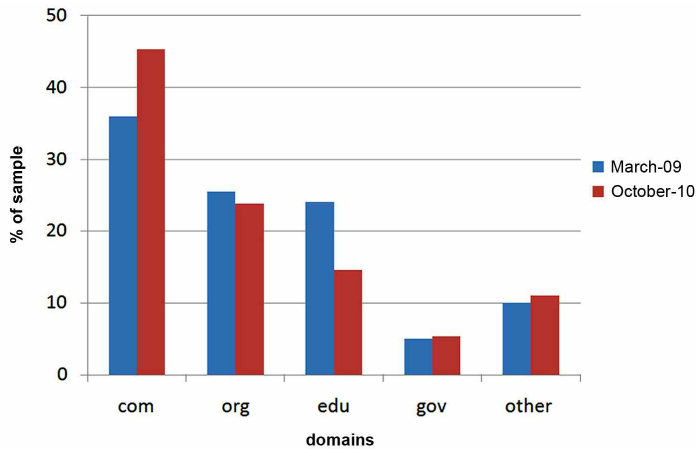


Figure 7.5. A Yippee (formerly Clusty) search for domain names shows an increase in .com at the expense of .edu.

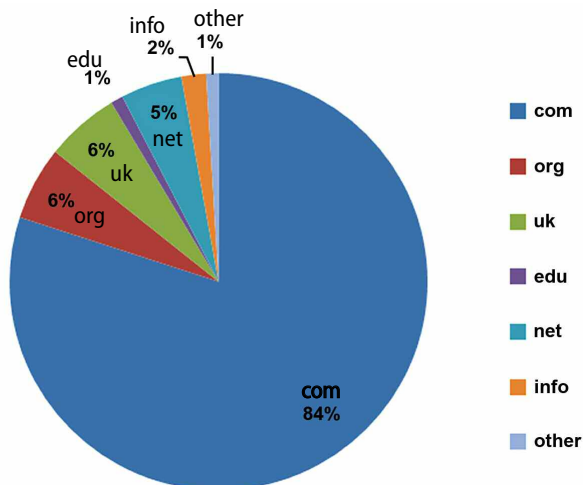


Figure 7.6. A Google Blogs search for “archaeology,” showing the relative strength of different domains.

<sup>18</sup> <http://blogsearch.google.com/?tab=wb>

Technorati's 2009 *State of the Blogosphere*<sup>19</sup> report found that 72 percent of bloggers worldwide blog as a hobby. Likewise, archaeologists tend to blog on their own time, outside their institutional setting. The shift to the commercial domain in archaeology could be linked to the low "acceptance" of blogging in academia. More likely, the availability of free, easy-to-use blogging platforms such as **Blogger**<sup>20</sup> and **WordPress**<sup>21</sup> may very well be the determining factor in this.<sup>22</sup> After all, one may find it easier to start a blog using a commercial service than to try to do so through one's home academic institution, where technical support and server space may be more bureaucratically entangled and difficult to arrange. Furthermore, if a researcher—say, a graduate student or an adjunct professor—is not permanently attached to an academic institution, then a commercial blogging service would offer more continuity than an ephemeral academic affiliation. Finally, commercial blogging services provide a place outside of their home institutions where bloggers can speak their minds.

Blog content, even that of narrowly defined, predominantly archaeological blogs, varies significantly. It can be centered on an excavation, a site, a type of artifact, a geographic area, a time period, a culture or civilization, or a theme that cuts across time periods and geography. In short, as a personal expression it offers a mirror of the many ways archaeologists specialize. Archaeologists also often have divided loyalties. For example, the authors of this chapter have interests in Mesopotamia, the ancient Near East, architecture, computer and Internet applications, spatial analysis, and zooarchaeology. This is in varying ways expressed in the posts on the Alexandria Archive Institute's blogs (**Digging Digitally**<sup>23</sup> and **Heritage Bytes**<sup>24</sup>). The more "institutional" a blog is, the more it tends to be focused on a specific topic or set of topics, with little personal insight. Personal, non-institutional blogs are more likely to touch on many different topics. The author of **The Archaeology of the Mediterranean World blog**<sup>25</sup> sees his blogging as "a sometimes

<sup>19</sup> <http://technorati.com/blogging/article/state-of-the-blogosphere-2009-introduction>

<sup>20</sup> <https://www.blogger.com/start>

<sup>21</sup> <http://wordpress.com/>

<sup>22</sup> About 60 percent of all archaeology-related blogs in the .com top domain use the free Blogger servers (.blogspot.com), and close to 30 percent use the WordPress ones (.wordpress.com), according to Google Blogs Search (accessed September 13, 2010).

<sup>23</sup> <http://www.alexandriaarchive.org/blog/>

<sup>24</sup> <http://ux.opencontext.org/blog/>

<sup>25</sup> <http://mediterraneanworld.wordpress.com/>

bizarre blend of academic and popular. This uneven character of blogs is what distinguishes them from more formal academic writing, but is also what makes them such a compelling medium” (Carraher 2008).

Evidence suggests that especially younger scholars (more familiar and at ease with social media) have more influence online than in the brick-and-mortar world, suggesting that the existing “hierarchy” of academia is not replicated in blogs. One could opine that blogging thus serves a democratizing and reinvigorating function in academia. In a nod to the growing relevance of blogging, a session dedicated to archaeological blogging took place at the 2011 Society for American Archaeology Annual Meeting in Sacramento,<sup>26</sup> constituting archaeological blogging’s “coming out” of sorts in professional circles. Archaeological bloggers now are more likely to identify themselves openly than in earlier years, when anonymous expressions of grievances with academic institutions and practices were a frequent motive for blogging. Advocacy of all kinds is still a factor, but it is more often practiced in the open.

Archaeology-specific blogs commonly cross-reference one another, providing “memes” that get picked up by others. Examples of trendsetting blogs are *Rogue Classicism*<sup>27</sup> and the *Ancient World Bloggers Group*.<sup>28</sup> There is no evidence for extensive discussions ensuing in comment sections of archaeological blogs. As discussed above, email lists are more conducive to ongoing discussions. Archaeological blogs as a subset of academic blogs matter greatly because they “are content-rich, and tend to focus on very specific areas. We create an enormous signal in the chaos of the internet. . . . Google controls how we find information; but often, academic blogging tells Google what’s important” (Graham 2010).

This last point helps illustrate one of the key roles played by blogging in archaeological communications. Blogs tend to share and comment on news, often relating to new funding opportunities, calls for papers, or announcements of new projects and collections. Many blogs also comment on policy and access concerns and developments. In that sense, blogs act as “attention focusers,” where, through their commentary and links, they help readers identify other significant and noteworthy content on the Web. As such, a

---

<sup>26</sup> See Morgan 2010.

<sup>27</sup> <http://rogueclassicisism.com/>

<sup>28</sup> <http://ancientworldbloggers.blogspot.com/>

blogger's notion of significance is tied to his or her motivation for blogging. In some cases, bloggers use the medium to promote their own research and projects, or to promote the work of allies and collaborators. In other cases, bloggers use the medium to advance policy agendas. Many blogs, including the *Stoa Consortium*,<sup>29</sup> the Ancient World Bloggers Group, and *Savage Minds* (an anthropology group blog),<sup>30</sup> frequently advocate for policies in favor of open access publishing.

In general, blogs tend not to share original research; yet, some exceptions exist. Blogs that are centered on a specific excavation or survey project often discuss research questions, methods, and progress. Sometimes the intent is instructional, since such blogs may give students a more intimate understanding of how fieldwork unfolds. In other cases, researchers may use a blog as a personal diary that just happens to be, almost incidentally, available for the world to inspect. Discussions of archaeology on the Web represent another exception where research is discussed in blog venues. For discussions on how to best use the Web for archaeology, blogs authored by Sebastian Heath, Tom Elliot, Leif Isaksen, and the Ancient World Blogging Group do convey original research, even if that research is presented somewhat casually and in "draft" form. At times, such blog posts provoke some comment and response from other blogs. However, because researchers have so many communication channels available, including email, telephone, and Twitter, responses provoked from research shared in a blog may come via a different medium.

## DISCUSSION: A MIDDLE SPACE FOR COMMUNICATIONS

For the time being at least, there seems to be little indication that user-generated content would ever replace more traditional forms of communication such as publication in refereed journals. While peer-reviewed publication is still the gold standard in archaeological communication, BoneCommons demonstrates that more casual Web communications have an increasingly recognized place in the profession. User-generated content in archaeology tends to both complement and supplement traditional publication; that is, it fills a "middle space" between casual conversation and formalized publica-

---

<sup>29</sup> <http://www.stoa.org/>

<sup>30</sup> <http://savageminds.org/>



tion. Web 2.0 offers tools for pre-publication discussions to take place, often in disciplinary portals like BoneCommons and the Zooarchaeology Social Network. Thus, user-generated content can help shape research outcomes by broadening the initial discussions and musings of researchers in a less formal setting.

Participation in the Social Web is both goal-directed and impacted by the expectations of others. In BoneCommons, participation is largely about self-promotion (that is, sharing one's research) or about requests for advice and help (such as sharing images so that others can help with the identification of a specimen). Social media sites often take years of development and buy-in from the respected leaders of a discipline. BoneCommons now holds a substantial corpus of presentations largely because BoneCommons hosted abstracts from the ICAZ 2010 conference, and the conference organizers specifically requested that people share their presentations there. The conference organizers (recognized authorities in the field) communicated an expectation for the posting of presentations and participants responded to it. This indicates that there is more participation in social media if such practice is promoted by established leaders in a field.

Social media also tend to succeed better when they are more conversational and reciprocal in their communications. In other words, bloggers who only talk about themselves and their own projects see less response than those who also highlight the work of others. Similarly, BoneCommons has more active participation because of its links with active conversations on the ZOOARCH list. When postings come to users by email, they are more directly tied into everyday workflows. Furthermore, the BoneCommons example shown in Figure 7.2 demonstrates how scholarly expectations can draw people into the use of social media (where the response of the third researcher to "ask the expert" drew that person into the conversation with the expectation that he would reply out of a perception of scholarly obligation).

The Social Web promises to help fill a hole that exists in scholarly communication. Traditionally, scholars communicate through print publication, a process that can take years, from completion of the research to its scholarly "debut." The other common form of communication has been presentations and casual discussions at conferences which, while beneficial, often occur just once a year. New forms of communication through social media help to fill the "middle space" between formal publication and casual discussion, and do so in a much faster and more efficient way that can involve more players.

Besides broadening the audience and speeding up the process of communication, social media can also address another need: that of finding the “needle in the haystack.” The distribution of BoneCommons searches over three months (from April through June 2010) resulted in a classic power-law graph (Figure 7.7), where a few topics dominate and the rest fall into a “long tail” of topics that cannot be categorized. Nearly half the searches were for people and publications. In that sense, BoneCommons acts as a useful community-curated index of zooarchaeological research. The site is a finding tool, open to Google searches, that connects users to publications and conference presentations. The remaining 51 percent of searches were for an assortment of things that could not be easily grouped into categories (hence, the long tail). This “everything else” category parallels the wide variety of search terms used to connect users to content in Open Context (see discussion in Chapter 2, and Figure 2.3). The long tail of search terms, each of which occurs only rarely, highlights the great breadth of user interests.

Finally, the comparison with Open Context reveals some interesting issues in the uptake of social media in research settings. Open Context made some early experiments with user-generated tagging. However, users made very few tags. BoneCommons also attempted to solicit user contributions

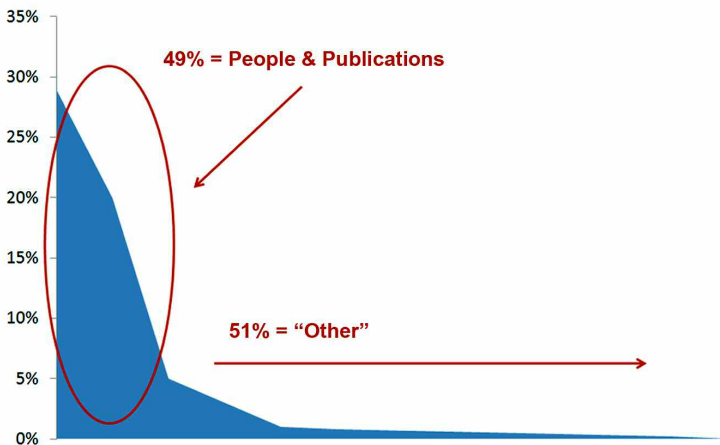


Figure 7.7. The “long tail” of 1,132 searches leading to BoneCommons (searches from April to June, 2010).

and, in contrast, it has been much more successful in this respect. Explaining these differences requires looking at the specifics of these systems. Open Context and BoneCommons serve very different purposes and host different sorts of content. Open Context experimented with tagging to supplement metadata on published data sets. However, since these data sets already had rich metadata, additional tags were less needed. More generally, publication in Open Context is much more involved and editorially guided, akin to publishing in a journal. The rate at which new data sets get published is also more journal-like. In contrast, posting in BoneCommons is rapid and represents very little investment in effort. The less formal nature of BoneCommons therefore leads to greater uptake and more frequent participation than Open Context. In the end, these experiences show that adoption rates for user participation in “scholarly media” will vary tremendously and depend on the specifics of the systems and the needs and interactions of the users involved.

## CONCLUSION

This volume explores many projects that aim to increase access to information. However, sharing and preserving archaeological content without the means to make sense of it is a pointless undertaking. In this context, one would expect user-generated content to exacerbate this problem; if we are drowning in information, adding more information would seem to make matters worse. However, user-generated content often works in somewhat counterintuitive ways. In some cases, user-generated content helps to prioritize and make sense of data. Blogs and Twitter help identify noteworthy content on the Web and thus can be seen as an important (if not very specific) source of metadata regarding the significance of content on the Web. This impact is felt even if few people actually read a blog post, as seems typical for archaeological blogs (see Harley et al. 2010: 97–100). Such links can elevate search engine rankings and thereby indirectly impact the discoverability and salience of noteworthy content.

In a somewhat different way, social media systems may help make sense of archaeological observations by enabling researchers to efficiently obtain expert advice. As discussed with examples from BoneCommons, the spare time of networked zooarchaeologists around the world can be an important resource to help an analyst identify a puzzling bone specimen or explore a

concept. The expertise obtained from such online collaboration can thus improve the quality and specificity of data produced in field or collections research.

This last point touches on an important issue. Social media sites like BoneCommons do not benefit all members of the research community equally. The clearest beneficiaries are likely to be junior scholars, since they tend to need expert advice on identifications as well as pointers to key background research (see also Harley et al. 2010: 74–79). BoneCommons can also provide junior researchers with models of presentations, posters, and research topics that they can emulate. More senior researchers do not have identical needs. Senior scholars typically review and edit papers routinely and supervise student research. These activities help them keep up-to-date with their field. Senior researchers also already sit within well-established professional networks, where advice, news, and gossip circulate. Thus, social media platforms often provide a less welcome and often distracting duplication of information flows for more senior professionals.

It will be interesting to watch how research-oriented social media develop along with more formal and conceptually sophisticated forms of cyberinfrastructure. There is some potential for tension. Since senior researchers control funding streams and promotion processes, the potential utility of social media may be underappreciated in “cyberinfrastructure” policy making. In that sense, BoneCommons is rather unique in having the endorsement of a professional organization and its leadership. At the same time, social media are clearly not at the cutting edge of informatics research. Blogs and the content management system behind BoneCommons now represent very standard and prosaic features of the landscape of the contemporary Web. Because social media systems are so easy to use and deploy, we expect to see continued experimentation in this space. Similarly, because social media systems are already so well established, there is little prospect for these systems to attract much in the way of cyberinfrastructure grant support.

While sites that support user-generated content fall outside the usual bounds of cyberinfrastructure, designers of archaeological cyberinfrastructure need to recognize the important role played by social media systems and user-generated content. Social media systems work by leveraging the simple power of the hyperlink. By sharing links, users share what is noteworthy. Unfortunately, some collections and cyberinfrastructure, though highly technically sophisticated, are not designed to support linking to a specific resource.

Often such collections, especially archives of publications, reports, or other specialized collections, are locked behind a login barrier. In other cases, as is sometimes seen in government information systems, the design of a system breaks good practice in web design. For example, one cannot link to a search result from the National Parks Service's National Archaeological Database (NADB). While this critique of the NADB may seem trivial, hyperlinks are a fundamental aspect of web interoperability (Isaksen 2008). These barriers inhibit others using social media platforms from creating links that could highlight potentially significant content. Such barriers also inhibit conversations about that content. From the perspective of social media systems, collections that do not support stable hyperlinks to their content are "silos" lacking in a simple but critical aspect of interoperability.

Thus, even though the Social Web lacks the prestige and conceptual sophistication inherent in many e-science or cyberinfrastructure efforts, it should not be ignored. Since the Social Web plays an important but poorly acknowledged role in resource discovery on the Web, designers of archaeological cyberinfrastructure should pay careful attention to the requirements of the Social Web if they want to maximize engagement with their users.

## REFERENCES CITED

- Beam, C.  
2010 #Posterity How Future Historians Will Use the Twitter Archives. *Slate Magazine* website, April 20, 2010. Retrieved from <http://www.slate.com/id/2251429/> (accessed August 29, 2010).
- Carraher, W.R.  
2008 Blogging Archaeology and the Archaeology of Blogging. *Archaeology* (website), January 17, 2008. Retrieved from <http://www.archaeology.org/online/features/blogs/> (accessed August 29, 2010).
- Deblauwe, F.  
2009 Archaeological Communities Online (Part 1). *Heritage Bytes* (blog), March 11, 2009. Retrieved from <http://ux.opencontext.org/blog/2009/03/archaeological-communities-online-part-1/> (accessed August 29, 2010).
- Graham, S.  
2010 Why Academic Blogging Matters. *Electric Archaeology: Digital Media for Learning and Research* (blog), March 2, 2010. Retrieved from <http://electricarchaeologist.wordpress.com/2010/03/02/why-academic-blogging-matters/> (accessed August 29, 2010).

Haas, M. R., and M. T. Hansen

- 2007 Different Knowledge, Different Benefits: Toward a Productivity Perspective on Knowledge Sharing in Organizations. *Strategic Management Journal* 28: 1133–1153.

Harley, D., S. K. Acord, S. Earl-Novell, S. Lawrence, and C. Judson King

- 2010 Assessing the Future Landscape of Scholarly Communication: An Exploration of Faculty Values and Needs in Seven Disciplines. UC Berkeley, Center for Studies in Higher Education. Retrieved from [http://escholarship.org/uc/cshe\\_fsc](http://escholarship.org/uc/cshe_fsc) (accessed June 22, 2010).

Kaplan, A. M., and M. Haenlein

- 2010 Users of the World, Unite! The Challenges and Opportunities of Social Media. *Business Horizons* 53/1: 59–68.

Isaksen, L.

- 2008 Pandora's Box: The Future of Cultural Heritage on the World Wide Web (slideshow as PDF). Invited keynote paper published in *Digital Heritage in the New Knowledge Environment: Shared Spaces and Open Paths to Cultural Content (Proceedings)*, ed. M. Tsipopoulou. Athens.

Markey, K.

- 2007 "The Online Library Catalog." *D-Lib Magazine* 13. Retrieved from <http://www.dlib.org/dlib/january07/markey/01markey.html> (accessed August 29, 2010).

Morgan, C.

- 2010 Blogging Archaeology 2011—The Abstracts. *Middle Savagery* (blog), September 17, 2010. Retrieved from <http://middlesavagery.wordpress.com/2010/09/27/blogging-archaeology-2011-the-abstracts/> (accessed September 30, 2010).

#### SECTION IV

### SUSTAINABILITY, QUALITY, AND ACCESS

Researchers today have a variety of publishing options beyond the traditional ones where a central authority of some sort plays ultimate arbiter in the decision to distribute their work. Now, with the click of a button (relatively speaking), everyone can be both author and publisher. While in many cases this is gloriously liberating, many in the academy, including the archaeological community, distrust the unfettered multivocality of easy Web-based self-publishing.

As demonstrated in this section (and in this volume in general), many scholars have explored the Web (and Web 2.0 models) as a platform to evolve and supplement traditional models of communicating and publishing archaeology. To many archaeologists, a Web 2.0 model may offer obvious, tangible, and immediate benefits over traditional venues of scholarly communication. This interest is at least partially a result of the economics of the scholarly publishing industry. Archaeologists have particularly demanding requirements for publication. They routinely need to communicate rich media, such as maps, illustrations, diagrams, photographs, and even data tables. These can be expensive to reproduce through traditional book and journal venues. Universities have difficulty absorbing such expenses. As university libraries see their budgets shrink (as a result of the economic cuts currently impacting many universities around the world to one degree or another), they make hard choices as to which journals they can realistically afford to purchase. Because, unlike biotechnology or nanotechnology, archaeological research offers little prospect for spurring great technical advancement and economic growth, archaeological publications can be vulnerable to budget cuts.

In many ways, open access and online publications mitigate this economic problem as well as provide a host of other benefits. Publishing materials online means that the “time to publication” is greatly reduced, and publications make their way into the broader scholarly ecosystem much faster than they would traditionally. The economics of publishing materials online

is also quite different. When publishing online in an open access framework, there is less concern with the cost per unit. In a traditional publishing model, the cost of printing and distribution can be so high that the publisher can afford to produce only a very limited number of units. Delivering materials online (especially in a Web 2.0 framework) negates the cost-per-unit equation, allowing the work to reach far more scholars than it might have.

Broader access also means that there is a far greater potential for pursuing previously unexplored research paths and interdisciplinary collaboration. Digital publication in archaeology also forces us to rethink the nature of what exactly constitutes a “publication.” In the case of Willeke Wendrich’s chapter on the UCLA Encyclopedia of Egyptology (Chapter 8), there is a clear sense of “publication as platform.” The venue she develops can deliver rich media, maps, data sets, monographs, short encyclopedia entries, video, and more. This diversity of content goes far beyond the boundaries of traditional print. At the same time, adherence to recognized professional standards of editorial control means that this diverse digital content will (hopefully) exhibit the same quality as more conventional print publication. Moreover, the idea of “publication as platform” can mean that content will be interoperable with other knowledge systems, thus enabling scholars to combine and recombine information in innovative ways to facilitate new interpretations and understandings.

Despite these benefits, an open access model is hardly without its challenges. Wendrich herself is skeptical of the long-term financial sustainability of open access forms of dissemination. Quality must be maintained and, in order to do so, editorial and review processes need backing from some sort of revenue stream, as Harrison Eiteljorg discusses in Chapter 10 in regard to the challenges of holding digital content to rigorous academic standards. The quality and sustainability concerns raised by Wendrich and Eiteljorg cannot be glossed over. At the same time, open access advocates point to a number of successes in advancing sustainable (so far) and high-quality models of freely accessible, high-quality, high-impact scholarly publication. For instance, the Public Library of Science’s flagship open access journals boast impact scores that rival *Nature* or *Science*. It remains to be seen whether or not these successes can be replicated in archaeology, and this open access volume represents an early experiment in the model for this discipline.

Quality and sustainability concerns represent only two of the issues that complicate archaeological adoption of Web-based publishing. One also must



consider how open access or any other digital dissemination fits with models of tenure and promotion. Traditionally, publication in peer-reviewed journals has been a vital (if not central) part of the tenure and promotion machine. The reality is that many in the scholarly community do not view open access or other digital publications as having the same value as work published in more traditional, paper-based publications. This concern clearly speaks to the aforementioned issue of quality and assessment. As many of the authors in this section suggest, there is no real reason that, given acceptable standards for assessment, open access scholarship cannot be counted toward tenure and promotion in the same manner as its print counterparts. Wendrich very rightly states that archaeology must work to evolve the concept of authorship to include the online publication of data. The idea behind this is that many archaeologists are still reluctant to publish raw data online for fear of falling prey to “data thieves” who publish broad overviews based on the painstaking work of others. If there existed a collective sense that scholars receive citation when their raw data are used in other publications, there might be a far greater incentive for raw-data publication.

Stewardship and stability, also of great concern in this discussion, are issues that all of the authors in this section address to one degree or another. If a journal or book publisher disappears, the journals and books themselves still exist. However, if a purely digital publication (whether a publication or digital repository) ceases to exist, there is a far greater possibility that its contents will disappear forever. As a result, there is an increased need for sustainability and preservation among digital publications.

In many ways, Jingfeng Xia’s discussion (Chapter 9) of disciplinary archives for archaeological publications may represent the model for digital dissemination that best accommodates career pressures faced by academic archaeologists. With such disciplinary archives, archaeologists continue to rely upon traditional publication venues for vetting and marking contributions as “quality.” At the same time, the archive can promote wider accessibility and publication speed by disseminating preprint versions of accepted, peer-reviewed publications. The question of whether scholars place their work in open access archives or publish in open access platforms is becoming more an issue of institutional imperative than of personal choice. An increasing number of universities, cultural institutions, and funding agencies are requiring that content be openly published online. Thus, while wholly digital or wholly open access forms of dissemination remain in their early,

experimental stages in archaeology, online dissemination of versions of printed media seems now firmly entrenched in the field. Only time will tell how these various forms of scholarly communication will co-evolve with one another as well as with changing perceptions of quality, career needs, and other social expectations in the profession.

## UCLA ENCYCLOPEDIA OF EGYPTOLOGY, ARCHAEOLOGICAL DATA, AND WEB 2.0<sup>1</sup>

*Willeke Wendrich*

True encyclopedic knowledge in this day and age means not just the ability to retrieve information on a wide range of subjects, something the Internet offers *in extenso*, but to have access to the relevant sources, of guaranteed quality, with clear guidelines on how to search for the desired material, together with data about when, how, and why that material has been collected. With the increased availability of information on the Internet, the emphasis of our research is changing from data collection (finding that rare manuscript in an obscure archive due to sheer tenacity, or spending years in the field excavating and recording), to data selection, data reduction, and data interpretation. This change increases exponentially the potential to open unexplored research paths and interdisciplinary work, but also raises the problem of the quality of data (Borgman 2007).

In the context of research and education, the quality of academic information is a fundamental issue. Students being trained in a certain discipline need to learn how to critically assess their sources. This is perhaps the main task of an academic education, and traditionally has depended strongly on the reputation of individual scholars and institutions. For students who are

---

<sup>1</sup> I would like to thank several of my cooperators in the development of the UEE: Jacco Dieleman, Zoe Borovsky, Yoh Kawano, Christina Patterson, Andrew Tai, Lisa MacAulay, Stephen Davison, Elizabeth Waraksa, and Hans Barnard. Acknowledgments are also due to the National Endowment of the Humanities, which has funded the UEE with two sequential grants. In addition, I would like to thank Eric Kansa and Sarah Whitcher Kansa for their thought-provoking remarks during the editing process.

learning how to assess quality, there are a few helpful pointers, such as class reading lists and the holdings of the university library. At the same time, accredited members of academic communities do not always agree: academic debates, sometimes mellow, sometimes vicious, demonstrate that there are serious differences of opinion on what entails “quality.” Therefore, even if a book is part of the holdings of the university library, there still might be objections against its contents from a sector of the profession. Other books have been purchased specifically to illustrate problematic approaches or methods. Books that are published by “outsiders,” such as science journalists and non-professional enthusiasts, enlarge the scope of the literature, but also complicate its assessment. Mostly, though, a rule of thumb for beginning students could be “if it is in the university library, it has a role in the ongoing discussion of the discipline.”

When we include the Internet, the quantity of information, contributed by a wide variety of individuals and institutions, is bewilderingly vast. Much of this information reflects recent developments and is often more up-to-date than that in printed volumes or journals. The publication process takes time, while the Internet is considered by many academic authors as a less official medium with easy, broad, informal access. It is relatively easy to publish reports, usually preliminary or with only a summary of the results, on a project website or in a blog. These are generally not considered to be scholarly or scientific publications, but merely informal accounts or announcements. The “real” publication is often a traditional printed article or book. The division between print and online publication is, however, fading rapidly. Traditional print journals are increasingly also published or accessible online, and there is no difference in quoting the print form or the downloaded PDF. So how can students, and the general public for that matter, assess the quality of information when it is not part of a corpus of traditional printed sources that have been vetted and published in the traditional way, whether made available on the Web or not?

There are two answers to this question, based on a variety of premises. One is that the tried and tested system of academic validation should stay intact, which could mean three things: that information on the Internet should be ignored; that only information that mirrors existing publications (PDF copies) should be considered valid; or that publications prepared and vetted in the same way as printed publications (peer-reviewed and stable) are suitable or authoritative.

The other answer engages with the great unknown and embraces the development of the Web, the involvement of new communities, the consideration that multivocality requires new assessment criteria and mechanisms, and the exciting but also frightening prospect that the development of the Internet is quite unpredictable. Some projections, based on an extrapolation of the present situation, maintain that information will be accessed everywhere, at any time, and that it will be highly personalized, location based, and relentlessly shared (Perry 2008). Web 2.0, also known as the “Social Web,” represents the opposite of traditional publishing in several fundamental aspects. The dichotomy between author and reader has disappeared: everybody is an author, and all opinions are published. Also, the time between publication (in the broadest sense) and reaction has been diminished dramatically: reactions can be posted instantaneously. The authority of writer and critic is based on frequency and intensity of “Web presence” rather than on professional career or scholarly reputation.

Although not immediately obvious, these two answers, which can be characterized hyperbolically as “securely isolated” or “out-of-control explosive,” are not mutually exclusive. Web 2.0 increasingly intersects with the traditional academic and publishing world. There are several good examples of this, such as Wikipedia and the commercial bookselling website [Amazon.com](http://amazon.com).<sup>2</sup> Wikipedia is designed to be written by its users, based on the premise that the ultimate information resource arises from a broad sharing of knowledge. Authors and editors need not be hired and paid, while users will donate their expertise and their time for a wide range of reasons, one of which is the sense of being part of a communal effort to improve the standing of knowledge in the world (Ciffolilli 2003). This is effective to a certain extent, but knowledge is neither value-free nor objective, and understanding the contributing community, as well as the programming, is of utmost importance in assessing the information quality (Druck et al. 2008; Voss 2005). Amazon.com and other online commercial endeavors provide the user with targeted suggestions, based on an analysis of previous purchases and search history. The website offers the possibility to publicly share one’s opinion by writing reviews and using tags to classify and group certain products for one’s own use, or for the world to peruse. The website has a large user community which spends much time and energy in authoring often lengthy and

---

<sup>2</sup> <http://amazon.com/>

opinionated reviews and extensive reading lists. As an overcommitted academic, my suspicion is that these are written by people who have either a very specific agenda or nothing better to do with their time. Without false modesty, I have to concur that the consequence is that my well-founded expert opinion is never heard on the Amazon.com website, even though I buy many of my books for academic use online. A case in point is set of reviews for a book with the fascinating title *Shamanic Wisdom in the Pyramid Texts* (Naydler 2004). When I accessed the website on May 3, 2011, the book had received five stars based on four raving reviews, one of which was titled “Hopefully Naydler has hit the <reset> button of Egyptology” (Amazon.com 2011). In the eyes of the reviewer, Egyptology apparently belongs to the “securely isolated” rather than the “out-of-control explorative” camp, unwilling to be swayed by alternative theories brought up by those from outside the discipline.

So what happens if one of my undergraduate students, who has not yet acquired a firm background in humanistic scholarship, sets out to write a paper on the ancient Egyptian Pyramid Texts? Why would using a book with the spectacular title *The Cannibal Hymn* (Eyre 2002) be considered laudable, while quoting *Shamanic Wisdom in the Pyramid Texts* would likely result in a lower grade unless the text were given a critical discussion? As stated earlier, the scholarly community relies on prior knowledge, peer review, and the reputation of the researcher for its quality assessments. A student, lacking specialized knowledge, finds helpful indications of the reputation of authors by checking university libraries (some of which do an excellent job in vetting online resources, such as <http://guides.library.ucla.edu/content.php?pid=21445&sid=152354>) or the name of the publisher (in the example above, these are, respectively, Liverpool University Press and Inner Traditions). When information is gained from websites, there are similar hints—for instance, whether a website is related to a university (in the United States, recognizable by the extension “.edu”), a reputable printing house (e.g., [Wiley Interscience](#)),<sup>3</sup> or scholarly collections of previously printed publications ([JSTOR](#))<sup>4</sup> or other media such as images ([ARTstor](#)).<sup>5</sup> Few websites, however, provide a good indication of authorship, purpose, policy, and financial backing. To avoid potential bewilderment and frustration, students need to be given guidelines on how to recognize scholarly or scientific reasoning, which

---

<sup>3</sup> <http://www.interscience.wiley.com/>

<sup>4</sup> <http://www.jstor.org/>

<sup>5</sup> <http://www.artstor.org/>

means that arguments, rather than statements, are presented and that the full range of evidence is put forward and considered critically, rather than only evidence that supports one specific argument while ignoring others. This requires a change in emphasis within academic education, as students need to be even better trained in critical reading and thinking than before. In archaeology, this means that the data need to be made available, with an explanation on how they were gathered and how they relate to the overall argument.

These issues are at the heart of the vision for the **UCLA Encyclopedia of Egyptology (UEE)**,<sup>6</sup> which consists of three parts: UEE Open Version, UEE Full Version, and UEE Data Access Level.

The **UEE Open Version**<sup>7</sup> provides articles on a wide range of subjects related to ancient Egypt by well-established authors whose contributions are peer reviewed. These articles are published online and can be downloaded freely (Figure 8.1). Egyptology has as its object of study the history, practices, and conceptual categories of a culture that was remarkably prolific in terms of written texts, art, architecture, and other forms of material culture. The richness of this culture, of which much has been preserved, allows us to reconstruct religious thinking, economic systems, intimate details of daily life, as well as ancient pathology, to name just a few aspects. For 30 years the *Lexikon der Ägyptologie* (*LÄ*; seven volumes edited by Wolfgang Helck, Eberhard Otto, and Wolfhart Westendorf), published between 1975 and 1992, has been the standard reference work in Egyptology. This great body of knowledge is still extremely useful for professionals in the field, but it begins to show signs of age. It obviously does not incorporate recent archaeological discoveries in Egypt, nor new insights or changed views that are at the core of the discipline as it evolves. The development of research and scholarly discourse makes revision of the range and configuration of entries of the *LÄ* urgent, but to publish a revised edition in print is prohibitively expensive. For American undergraduate students and for large numbers of the general public interested in ancient Egypt, the *LÄ* poses several problems. It is an expensive series, only available at specialized libraries, and most of the texts and all entry titles are in German. Despite the many articles that are published in English or French, and notwithstanding the English and French indices to the article titles, in practice the German language creates an insurmountable

---

<sup>6</sup> A multipage PDF with annotated screen shots of the UEE Open Version and UEE Full Version are on view at <http://www.escholarship.org/uc/cioa>.

<sup>7</sup> <http://repositories.cdlib.org/nelc/uee/>

The screenshot displays the eScholarship website interface. At the top, the eScholarship logo (University of California) is on the left, and navigation links (Home, About, Browse, Publish, Help, My Items (0)) are in the center. A search bar is on the right. Below the navigation bar, there are social media links (Facebook, Twitter) and the UCLA Encyclopedia of Egyptology logo. The main content area is divided into two columns. The left column contains a 'General Information' sidebar with links to the encyclopedia's website, library of Congress number, ISBN, and various administrative links. The right column features an 'RSS' feed and a 'Recent Work' section listing 80 publications from 2008 to 2011, including titles like 'Mud-Brick Architecture', 'Taxation', 'Birth House (Mammisi)', 'Throne', 'Amarna Art', 'Cosmogony', 'Village', 'Glass Working', 'Sex and Gender', 'Foreign Deities in Egypt', 'Quarrying and Mining (Stone)', 'Painted Funerary Portraits', 'Reuse and Restoration', 'Usurpation of Monuments', and 'Child Deities'. A footer at the bottom contains links for 'My Account', 'Contact eScholarship', and a statement about the site being powered by the California Digital Library with The Berkeley Electronic Press.

Figure 8.1. The interface of the UEE Open Version in eScholarship ([http://escholarship.org/uc/nelc\\_uee](http://escholarship.org/uc/nelc_uee)).

barrier for many. By contrast, the UEE is published in English, with Arabic, German, and French titles, and abstracts in Arabic for every article. The target audience of the UEE is both the scholarly community and the general public with an interest in ancient Egypt.

The advantages of online publishing are obvious at all phases of the production and use of the online encyclopedia. In the writing phase, all tasks are done through the same online system: the invitation to authors, tracking whether authors are keeping to their deadline, submission of the manuscript, and peer review. In the publication phase, articles are published whenever the peer review and copy editing have been finalized. The UEE can afford to grant its overcommitted authors very flexible deadlines because the process is not held up by one tardy author. In the use phase, the “traditional” aspects of the UEE are apparent. The UEE has strict version control, which means



that the text of an article is stable, and when changes are made, the new version of the article, as well as the original text, will be available. Thus scholars can refer to “article editions,” which can be quoted as regular printed publications. It is a solution for the opposite notions of (1) being able to update a text whenever necessary and (2) the academic tradition of quoting a specific text to support or refute an argument. This is only possible if the quoted text is stable, while revisions are found under the same title but as a different edition. The fine line between a journal article published both on paper and online versus a journal article accessed only online is slowly blurring. It is not the publication medium, but the processes before publication—peer review and the editor’s assessment—that are becoming the most important reasons to trust a publication. This is of great importance in the present period, in which academic promotion and tenure are still mostly based on paper publications (Borgman 2007). That said, though supportive of academic publishing’s mission, this firm grounding in the traditional values of textual stability and the reputations of the publishers and researchers is not particularly conducive to innovation or paradigm shifts.

The UEE Full Version (in development)<sup>8</sup> pushes the innovative envelope slightly further (Figure 8.2). By using the peer-reviewed scholarly article texts as assets equal to photographs, drawings, videos, and 3D virtual reality (VR) models, which are pulled together on the fly through dynamic searches, the traditional encyclopedic entry becomes something more than just a published article. Searches are expanded to include not only the metadata of the visual materials, but also the bibliography. Articles that quote the same publications (suggesting that the topics are related) can be linked and listed. Thus a combination of algorithms and intense markup brings unexpected underlying relations to the fore. All UEE texts are marked up in XML using an international open standard for online publication, the **Text Encoding Initiative (TEI)**.<sup>9</sup> These XML master texts and all UEE images are hosted by the UCLA Digital Library and are described with metadata that adhere to national digital library standards: the **Metadata Object Description Schema (MODS)**<sup>10</sup> and the **Metadata Encoding and Transmission Standard (METS)**.<sup>11</sup> Use of such standards ensures that UEE content will

---

<sup>8</sup> <http://www.uee.ucla.edu/>

<sup>9</sup> <http://www.tei-c.org/index.xml>

<sup>10</sup> <http://www.loc.gov/standards/mods/>

<sup>11</sup> <http://www.loc.gov/standards/mets/>

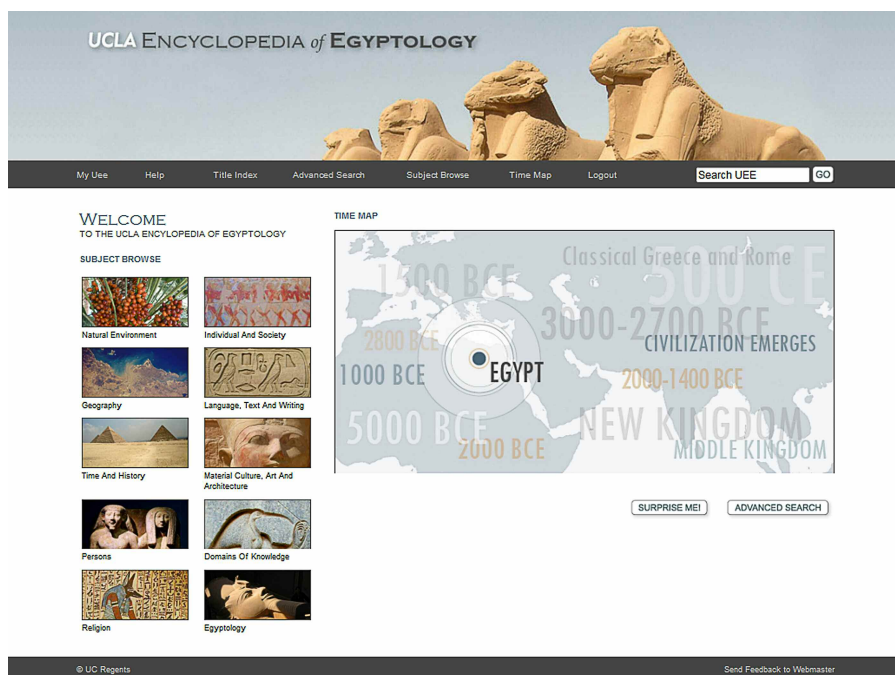


Figure 8.2. The interface of the UEE Full Version (not publicly accessible yet)

be interoperable and exchangeable with other information systems, thus enabling users to access the information in new ways and facilitate the creation of new knowledge and insights. This reliance on standards and the collaboration with the UCLA Library ensures the stewardship of the UEE content over time and enables it to be reusable in as-of-yet unknown applications. For the images, the UEE, through its archive in the UCLA Digital Library, is compliant with the light version of the Getty Institute's **Categories for the Description of Works of Art (CDWA Lite)**.<sup>12</sup> For international Egyptological terminology, the UEE makes use of the **Multilingual Egyptological Thesaurus (MET)**<sup>13</sup> for the spelling of place-names, but has developed a new

<sup>12</sup> [http://www.getty.edu/research/conducting\\_research/standards/cdwa/cdwalite.html](http://www.getty.edu/research/conducting_research/standards/cdwa/cdwalite.html)

<sup>13</sup> <http://www.ccer.nl/apps/thesaurus/index.html>

standard for the rendering of time periods. Adherence to, and augmentation of, existing standards is part of the development of a best practice in Egyptology, in order to promote data interoperability. The many online initiatives on ancient Egypt can only benefit from a common terminology, even if at this moment it is not entirely clear how different initiatives will interact. The UEE, as a standard work for the field, has an obligation to be as open and accessible as possible, to link not only to Egyptological initiatives, but also to more general (art) historical, geographical, or archaeological ones.

The Full Version makes use of the search functionality and multimedia features of the Internet and, at the same time, stays close to the concept of a traditional encyclopedia. The study of all the different elements and features of ancient Egypt is truly multidisciplinary, involving Egyptologists, archaeologists, linguists, geologists, and many other professionals involved in research in Egypt. The power of digital information, textual and pictorial, lies in the superior search and retrieval capabilities of digital tools for access and analysis. These not only provide the user with access to very specific content, but will also, as the UEE develops additional search capabilities, generate new research paths. By the same token, the digital encyclopedia still offers old-fashioned browsing, similar to a printed encyclopedia, where interesting entries are found serendipitously, on the opposite page or a few pages away from the actual subject under study. Unobtrusively highlighting subjects in the text is an expedient way to provide access to related entries, and for the benefit of nonprofessional users, convenient in-text explanations can be activated to explain Egyptological terminology.

An important interface is the UEE time map, which offers different views to provide temporal and spatial contexts for articles, photographs, plans, drawings, videos, and VR models (Figure 8.3). This time map illustrates the difference between the moderated, controlled interface of the UEE and the Web 2.0 mapping features. In Google MyMaps and in Google Earth, any user can add content to locations: for instance, photographs from a trip can be linked to the place where they were taken, or pages from a diary describing a building can be linked to its location. Google enables the use of a time slider, a balk with a time division which appears on the map page or satellite photograph, with pointers that can be manipulated. Thus a trip can be mapped out not only in space, but also in time. The information, however, is not checked in a systematic fashion for accuracy, consistency, or necessity. The UEE time map makes use of the Google application programming

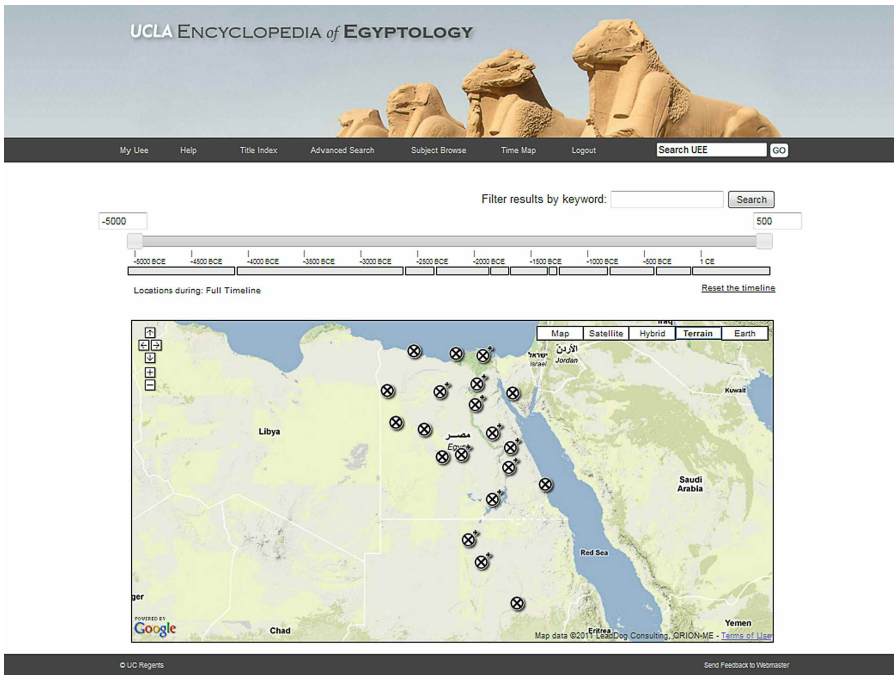


Figure 8.3. The UEE time map, overview.

interface (API) and links this with highly structured, consistent, and vetted data. The database, which sits behind the interface, contains different levels of nested information:

- ▶ *Regions*: Areas of Egypt, which function as zoom levels to go to more detailed maps
- ▶ *Governorates*: The official administrative boundaries of the modern Egyptian state, which are also the organizational divisions of the Ministry of Antiquities, the organization responsible for the monuments, museums, and archaeological sites in Egypt
- ▶ *Nomes*: The ancient Egyptian administrative units, of which the exact boundaries are not known, but for which textual evidence provides the information needed to determine to which nome a particular town belonged
- ▶ *Sites*: Areas with ancient monuments; these can contain several features

- *Features*: A classification of functional units, such as settlements, cemeteries, or temples, but also linear features such as roads, rivers, and canals, which on their own or in combination with others comprise a site (Table 8.1)

Table 8.1. Feature Types for the UEE Time Map

CATEGORY	FEATURE TYPE	CATEGORY	FEATURE TYPE
<i>Necropoleis</i>	Burial chamber	<i>Domestic Architecture</i>	(Boundary) stele
	Catacombs		Cave
	Cemetery		House
	Cenotaph		Hut
	Forecourt		Palace
	Funerary enclosure		Scatter
	Mastaba		Settlement
	Pyramid	<i>Industrial Architecture</i>	Harbor
	Rock-cut tomb		Nilometer
	Polytaph		Quarry/mine
	Sacred animal burial		Quay
	Serdab		Stable
	Shaft tomb		Storage facility/granary
	Tumulus		Workshop
<i>Religious Architecture</i>	Bark station	<i>Military Installations</i>	Fortified well
	Birth-house		Fortress
	Court		Road station
	Chapel/shrine/sanctuary		Watch post
	Church	<i>Linear Features</i>	Boundary
	Colonnade		Canal
	Divine temple		River/branch
	Gate		Road/route
	Hypostyle hall	<i>Natural features</i>	Gezira / turtle back
	Memorial temple		Natural lake
	Monastery		Wadi
	Naos		
	Podium	<i>Religious Architecture, continued</i>	
	Portico		
	Pronaos		
	Pylon		
	Rock art location	Sun temple	Temenos wall
	Sanatorium	Sacred lake	Window of appearances
	Solar court		

The regions and governorates are modern and do not have a particular time stamp. The nome system changed slightly over time and was fundamentally reorganized in the third century BCE. These changes are incorporated in the map and can be made visible by indicating a particular period with a time slider. Sites are dated by their features, because over time a site may have expanded or contracted, changed its function, fallen out of use, been reused, or abandoned forever.

The time map makes a distinction between the period of construction and the period of occupation. Features, such as specific chapels, pylons, or boundary stelae, were built during the rule of a particular king or pharaoh. An individual period of a king's rule is the smallest time span used in the diachronic overview and search capability offered by the map. The next step comprises the dynasties, spanning a number of rulers. This is the traditional Egyptological time division based directly on ancient Egyptian sources, which date events according to the number of years a king has been in power, and refer to history by dynasty, defined as the rule of a family, or more often a power block linked to a particular geographical region. The temporal units above that mark the historical periods—the Old Kingdom, Middle Kingdom, New Kingdom, and the Intermediate Periods—were devised by Egyptologists. All of these time segments are linked to years BCE and adhere to the UEE standard, which is based on a somewhat arbitrary decision in a field full of debates on the length of rule, co-rulership, and even the number of rulers at many points in Egyptian history. The existence of such debates is the subject of separate articles. At present, the only way users can contribute their opinion is by contacting the author or the editor. One of the items on our wishlist is to add an interactive discussion platform for user contribution. This will have to wait, however, until the major part of the content has been written, copy edited, marked up, and made available with online functionality.

Based on the time and date stamp of each item in the UEE, an article is accompanied by dynamically generated maps showing each place mentioned in the text or illustrated in the images. However, the map can also be used as a search tool. Selecting a period shows all sites that were either built or active during that particular time. Zooming in to a region (for instance, the Eastern Nile Delta) brings up individual sites (Tanis). Clicking on the site reveals the exact location of the various features (Table 8.1), such as a cemetery, settlement, and temple. That feature—for example, the Temple of Mut—can be shown in Google Earth, and visuals such as plans or photographs can be

used to overlaid the map at the appropriate location (Figure 8.4). The functionality of the time map and the main search functionality of the UEE can be gleaned in the [PDF of the online version of this chapter](#)<sup>14</sup> (arrow keys can be used to navigate through the different screens in the PDF).

The UEE Open Version is presently online and is gradually growing in size. The Full Version is at present available to UCLA students, and depending on the development of a model to make the encyclopedia financially sustainable, will be made public either for free or based on a subscription.

The third component of the UEE, the Data Access Level, is in its early stages of development. It was born from three related concerns: openness, accountability, and student learning. In the first place, as indicated above, perhaps the most important skill students should acquire during their university education is the ability to critically appraise arguments. To enable students and others to do so, the line of reasoning should be presented in an accessible, controllable fashion, which opens it up for critique. Because

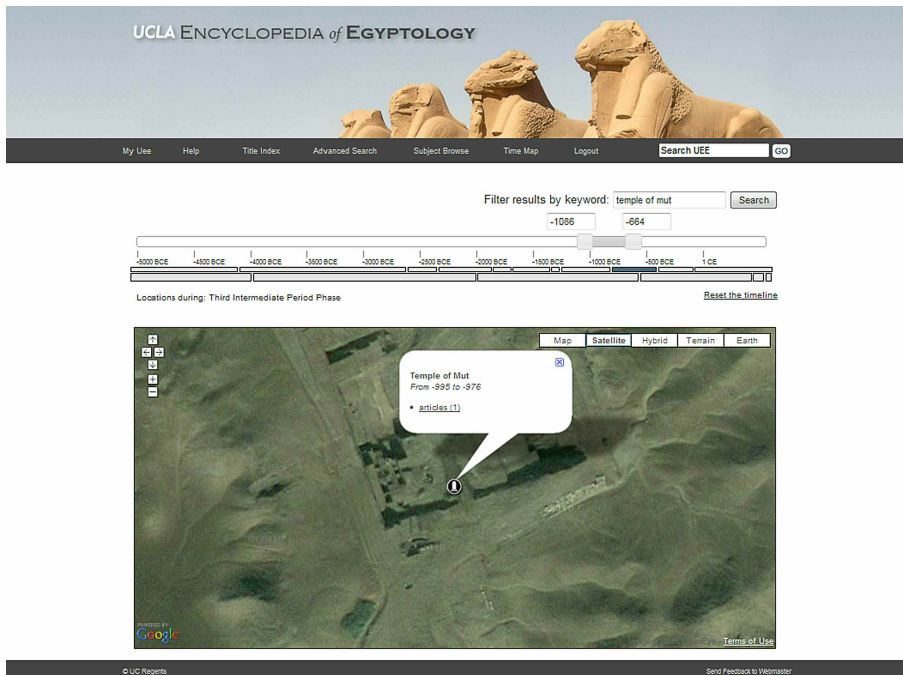


Figure 8.4. The UEE time map, zoomed in to the site of Tanis in the Eastern Nile Delta.

<sup>14</sup> <http://www.escholarship.org/uc/cioa/>

archaeological excavations destroy their most important object of study—the archaeological context—a research endeavor can never be repeated. An exception is the study of objects found during excavation and archived in storage rooms, as represented by the online **Digital Archaeological Archive of Comparative Slavery (DAACS)**.<sup>15</sup> For most archaeological studies, however, openness to falsification can only be realized to a certain extent by accounting for and publishing the data. In traditional archaeological publications, one rarely sees original data tables made available together with the analysis and interpretation. Diagrams, or statistical accounts, are often published, but they represent summaries or abstractions of the original data. Publishers consider it too costly, and authors assume that there is no interest in pages-long lists of numbers, top plans, and photographs representing the raw data of an excavation. Furthermore, archaeological data hardly ever provide clear-cut conclusions. Online publication enables the posting of databases, with which at least the quantification part of an interpretation can be evaluated, reproduced, and used for the same or alternative research questions. It is as close as archaeology will ever come to a laboratory situation in the sciences, where reproduction of results is one of the key factors in research assessment.

The same issues we saw in relation to the publication of texts are also at play where the publication of audiovisual assets and databases is concerned. There is no system in place for the (peer) review of archaeological data; nor is there standardization of either data gathering or metadata, the information that describes how and why particular data were gathered. Furthermore, there is no ready solution for the archiving of data. The UEE Data Access Level will contribute to attempts to offer original data, from survey, excavation, philological, or art historical research, and will be a test-bed for how a user of general articles can drill down to data that are linked to specific research questions.

While the prices for data storage are decreasing, the costs for migrating data to new versions of programs and updated carriers (from tape to hard drive to flash memory) continuously increase. Archiving archaeological data is an expensive endeavor; so, while excavators may be well disposed to storing their data in a trusted repository, for universities, such repositories are difficult to sustain. The Data Access Level does not set out to collect all the data ever gathered, but will try to be a cross between an archive and a portal

---

<sup>15</sup> <http://www.daacs.org/>



that points the way to original information on the archaeology of Egypt stored elsewhere, and possibly also to other Egyptological information. Where the encyclopedia article offers a quick introduction and recent literature, the Data Access Level enables users to appreciate the information that underlies the summarizing interpretations of the UEE.

Because of the interdisciplinary nature of archaeology, the research of a single archaeological project results in a number of different databases, rooted in the various disciplines. The organization of and access to these data are not simple matters. We are dealing with the conundrum that the wealth of information, which is the result of these studies, is often difficult to integrate into one system or even to systematize in any way, because the terminology and types of data vary considerably. Expanding the effort to encompass different research groups complicates the problem considerably. As is clear from other contributions to this volume, there is a great resistance among archaeologists against standardized data structures, not only because of the various disciplinary and national backgrounds of researchers, but also because each archaeological site is different, and the questions posed by the research groups have direct influence on the type, granularity, and quantity of information considered relevant (Baines and Brophy 2006).

Furthermore, if we maintain that theory and method are closely linked, then it could be argued that we hold back the development of the profession by too rigid a standardization of the way we record excavations, finds and visual, audiovisual, and spatial information.

Yet we make use of standardized data collection for quite a large portion of our data. We utilize only a limited number of geographic coordinate systems, we usually give our measurements in metric values, and stratigraphic excavation is at present the standard method, which results in data that are sufficiently uniform to allow comparison. Initiatives such as [Open Context](http://opencontext.org/)<sup>16</sup> set out to find a happy medium between standardization and variation. The things that stand in the way of further standardization are directly related to:

- ▶ Differences in terminology
- ▶ Levels of granularity
- ▶ Methods

---

<sup>16</sup> <http://opencontext.org/>

- Flexibility versus rigidity
- Acceptance by the discipline(s)

Differences in terminology have been approached by providing a thesaurus that “translates” terminologies of different initiatives into a common denominator. Such a generalized term sometimes covers the exact same feature which simply has a different name (for example, excavation unit = trench; locus = unit); more often it covers approximately the same concept, with very particular differences in nuance, strongly adhered to by the researcher. The level of granularity of these “translations” is the smallest common denominator, but is also directly related to the research need of a project and the excavator’s decisions on how to spend precious recording time. Archaeologists often find themselves in strained situations because of limited funding or because of an immediate threat of destruction or need for intervention (rescue archaeology). Conscious decisions about how much should be recorded and how detailed the records should be are made on an ongoing basis and are established for each unique situation.

Even if two researchers use the same terminology and record the same things, the exact method might vary. “How-to” books therefore specify exact recording methods, in order to enable consistency in the record (e.g., MoLAS 1994; Orton et al. 1993; Rice 1987; Wendrich 1994). As DAACS demonstrates, comparison is most successful if there is consistency in the record, and this requires consistency in the method.

As outlined above, consistency in terminology and methodology also creates a level of rigidity, which might turn out to be detrimental to the development of the profession. By forcing a methodology onto research questions or project circumstances for which they are not suitable, by making method into law, we would be blocking new research avenues and other types of important innovation—which is quite unacceptable for our anarchic, freedom-loving community of practice. It should be possible to find a balance between flexibility and consistency by carefully thinking through the structure and types of data that are generated. Such can be achieved by collecting or routing data collections through central nodes and providing them with extensive metadata, which specify the terminology and method in great detail, and by using smart screen designs, which present all the data in the same format while indicating where the record differs from a previously defined standard or from the most common practice.

For the UEE Data Access Level, this practically results in the following steps. Because preservation of archaeological data is such a critical issue, it is important to keep the threshold low and to ingest as much material as possible to prevent irretrievable data loss. This means that in practice there will be three data quality levels: peer-reviewed data for which extensive metadata clarify the recording method, granularity, and extent to which the information is standards compliant; peer-reviewed data that does not pass the review criteria but may be the only information available on particular excavations; and data that have been submitted but have not (yet) been peer-reviewed. UEE will work with data owners to improve accessibility through the path outlined above if they are impaired by the differences in data. By providing clear indications of where data are compatible and where they are not, and by providing on the output site a seamless communication with analytical programs and a publishing platform, many of the present obstructions can be cleared away or circumvented. Editorial processes in data assessment will make the publication of data more streamlined and closer to the publication of field reports. These very processes may help promote the scholarly legitimacy of data publication. They may also help overcome another enormous obstacle: the overprotection of data ownership. At present, the lack of copyrights on data makes many researchers hesitant to publish their information in an open and accessible way, even if there are well-established social norms among scholars that require citation when using someone else's data set. Publishing a database, however, is not much different from publishing printed data overviews in traditional journals, which are likewise not covered by copyrights but are protected by social norms.

It is foreseen that all users of the UEE Full Version will have access to the peer-reviewed data along with the full reports. Many archaeologists have a keen sense of ownership of their data and are hesitant to make them available, let alone to have them reviewed. An opinion often heard is that the publication of data opens one's work to "data thieves" who publish grand overviews based on the painstaking work of others, reaping the fruits of the work without ever having organized a large a research project or having raised the considerable funds required to do the research. Such resistance can be partly overcome by working toward an expansion of the concept of authorship to include the (online) publication of data, such that scholars are quoted when their data are being used in comparative research, and by allowing authors to determine when their data can be published. It seems

likely that, over time, the concept of data publication will gradually become more accepted, eventually becoming a recognized academic endeavor (Borgman 2007). By embedding this shift in an integrated presentation of archaeological data and interpretation, it will be easier to steer professional practice toward granting tenure and promotion based on results and activities other than publishing a monograph or a series of articles in high-level journals.

Archaeology, in most regions of the world, is prone to use, or abuse, for political purposes (e.g., Boytner et al. 2010; Caton-Thompson 1931; Meskell 1998). Archaeologists cannot shield their work from being enlisted for purposes not their own. This brings us back to the discussion on Web 2.0 as a medium in which the distinction between author and public has almost disappeared. Should our archaeological data be available in an open system, accessible for everybody, or should it be password protected and made available only to members of the guild? If multivocality, the interpretation of our archaeological heritage and history by those who are not part of the established profession but who have their own interests in using the information, is one of the potential interests of archaeological knowledge dispersal, then the Internet, and in particular Web 2.0, provides the ideal vehicle and process. So far, UEE has chosen a hybrid system, with access to all articles freely available through the Open Version, while the full data are made available on a subscription basis. This choice was not instigated by concerns of abuse, but solely based on a need to generate a sustainable income to create and maintain a long-term, high-quality resource that can be updated regularly. It has been agreed that all users with an IP address in Egypt have free access to the UEE Full Version, so that every Internet café in Egypt provides a gateway to gain deeper knowledge about the Egyptian cultural heritage. A stumbling block is the language. At present, each article has an abstract in Arabic, and funds are being sought for translating the article texts into the language of modern Egypt.

Although the content of the UEE is heavily mediated, some aspects of the Full Version, such as MyUEE, which will enable users to annotate and tag UEE content, are distinctly inspired by Web 2.0 and the great potential of building communities by sharing ideas and information.

As outlined above, the sustainability model that is in development for the UEE is a hybrid between open access and a subscription-based version. Both versions will have the same high-quality content, but the functionality will

differ, with the UEE Full Version providing searchable maps, and links between terminology, references, and keywords which will enhance the heuristic capabilities of the encyclopedia by offering surprising underlying structures rather than purposely built-in connections. This hybrid form of open/closed access (also known as a “freemium” model) will be augmented by donations, by grants for extra functionality and continuing innovation, and, ideally, by an endowment which will ultimately enable the Full Version to be made available for free. Intensive contact with several academic publishers to develop a financially sustainable model has brought to the fore how many models are in development in 2010. Because of the long-term, continuous development and update (over decades rather than years) and the strict scholarly nature of the UEE, the editorial board determined that inclusion of advertisements or other commercial links would be less desirable than a subscription model. The fickleness of commercial interests may not be the ideal support for an independent, well-established, yet flexible and responsive scholarly endeavor as the UEE is designed to be.

## CONCLUSION

Online media provide opportunities that are not available in print publications. Innovative search functions and data mining that make use of textual, numerical, pictorial, geospatial and audiovisual formats have an important heuristic function that by far surpasses that of searches in printed volumes. The vision for the UEE is that, in addition to presenting a conveniently accessible body of excellent content, it will ultimately open up new research methods, through combining and interweaving the knowledge of eminent scholars with original data in completely new ways. Archaeology is well positioned to be at the forefront of developments in data publication and digital scholarship, because it is inherently multidisciplinary and collaborative, and even its traditional publications include many photographs, plans, drawings, diagrams, and tables. A certain tension exists between the protection of data and the ideal of the Internet to create a universal resource. The friction between the profession as a closed guild of approved members and an open-for-all Internet community, between “securely isolated” or “out-of-control explorative” is probably a temporary one. Social networking allows for groups of persons with similar interests to find one another, send links to relevant online information, discuss subjects, and quickly disperse newly

gathered information. Such social networks range from virtually anonymous and completely open, to closely mediated or restricted; but even the most accessible ones result in self-selective communities. If anything, the archaeological community will have to try hard to get their interpretations read, their data used, and their information heard. The development to which the UEE contributes is one of greater openness, more intense cooperation, and the provision of credit for authors of both data and interpretation. An open and accessible presentation greatly increases the possibilities for interdisciplinary scholarship. Giving researchers from other fields access to accurate introductions and summaries will enable them to make comparisons across disciplines.

Web 2.0 audiences are accustomed to having a voice, a means of expression, which seems contrary to the concept of a highly mediated professional channel such as the UEE. At the same time, there is a clear need for an authoritative guide to negotiate fields where there is too much (mis)information. In scholarly communication, it is accepted that “good” information is a matter of disciplinary intersubjectivity, an agreement established by peer review, while at the same time acknowledging that strong and valid differences of opinion exist and should be expressed. The concept of multivocality has been introduced in archaeology to express that there are different groups, with different interests and voices, that have a stake in the same subject area. And yet, some of these voices are considered “scholarly valid” while others are not. By qualifying the audience and making explicit the scholarly preconceptions, it is clear which voices are being heard and which need to find an alternative outlet. Although Web 2.0 enables scholars to be engaged in the world and avoid the ivory tower trap, scholarly communication requires a fine balance between providing and exchanging qualified information and being open to the world beyond academia.

## REFERENCES CITED

Amazon.com.

- 2011 Amazon.com reviews of Naydler (2004), accessed May 3, 2011. Retrieved from [http://www.amazon.com/Shamanic-Wisdom-Pyramid-Texts-Tradition/dp/0892817550/ref=sr\\_1\\_1?ie=UTF8&cs=books&qid=1234221091&sr=8-1](http://www.amazon.com/Shamanic-Wisdom-Pyramid-Texts-Tradition/dp/0892817550/ref=sr_1_1?ie=UTF8&cs=books&qid=1234221091&sr=8-1).

- Baines, A., and K. Brophy
- 2006 What's Another Word for Thesaurus? Data Standards and Classifying the Past. In *Digital Archaeology. Bridging Method and Theory*, ed. T. L. Evans and P. Daly. London/New York: Routledge.
- Borgman, C. L.
- 2007 *Scholarship in the Digital Age. Information, Infrastructure, and the Internet*. Cambridge, MA/London: MIT Press.
- Boytner, R., L. Schwarz-Dodd, and B. J. Parker (eds.)
- 2010 *Controlling the Past, Owning the Future: The Political Uses of Archaeology in the Middle East*. Tucson: University of Arizona Press.
- Caton-Thompson, G.
- 1931 *The Zimbabwe Culture: Ruins and Reactions*. Oxford: Clarendon Press.
- Cifollilli, A.
- 2003 Phantom Authority, Self-Selective Recruitment and Retention of Members in Virtual Communities: The Case of Wikipedia. *First Monday* 8/12.
- Druck, G., G. Miklau, and A. McCallum
- 2008 Learning to Predict the Quality of Contributions to Wikipedia. In *Proceedings of the Wikipedia and Artificial Intelligence: An Evolving Synergy Workshop at AAAI-08, Stanford University, California*. AAAI Press. Available at <http://www.aaai.org/Papers/Workshops/2008/WS-08-15/WS08-15-002.pdf>.
- Eyre, C.
- 2002 *The Cannibal Hymn: A Cultural and Literary Study*. Liverpool: Liverpool University Press.
- Helck, W., E. Otto, and W. Westendorf (eds.)
- 1975–1992 *Lexikon der Ägyptologie*. 7 vols. Wiesbaden: Otto Harrassowitz.
- Meskell, L. (ed.)
- 1998 *Archaeology under Fire*. London/New York: Routledge.
- MoLAS
- 1994 *Museum of London Archaeological Service (MoLAS) Archaeological Site Manual*. London: Museum of London.
- Naydler, J.
- 2004 *Shamanic Wisdom in the Pyramid Texts: The Mystical Tradition of Ancient Egypt*. Rochester, VT: Inner Traditions.
- Orton, C., P. Tyers, and A. G. Vince
- 1993 *Pottery in Archaeology*. Cambridge Manuals in Archaeology. Cambridge/New York: Cambridge University Press.
- Perry, M.
- 2008 The Appliance of Science: Web 2.0. *Information World Review* 252 (03 Dec 2008).
- Rice, P. M.
- 1987 *Pottery Analysis: A Sourcebook*. Chicago: University of Chicago Press.

Voss, J.

- 2005 Measuring Wikipedia. In *International Conference of the International Society for Scientometrics and Informetrics: 10th Meeting 24–28 July 2005, Stockholm (Sweden)*.

Wendrich, W.

- 1994 *Who is Afraid of Basketry: A Guide to Recording Basketry and Cordage for Archaeologists and Ethnographers*. Leiden: Centre for Non-Western Studies, Leiden University.



## OPEN ACCESS FOR ARCHAEOLOGICAL LITERATURE: A MANAGER'S PERSPECTIVE

*Jingfeng Xia*

It has been more than several years since I first proposed a digital repository for archaeological literature at the Annual Meeting of the Society for American Archaeology in Vancouver in 2008. Over the past several years, the open access movement has been energized by a series of efforts and policy changes. One of the most important efforts was the Obama administration's delivery of a strong voice in the open access debate in late 2009. Then, after Harvard University's Law School and the faculty of Arts and Sciences voted (in early 2008) for a mandate policy to make scholarly articles of their faculty available free online through an institutional repository (Cohen 2008; Suber 2008), many more higher educational institutions, such as Brigham Young University and Stanford University, have also participated in the movement and set their own mandate policies on self-archiving, not only in the United States, but also all over the world (Suber 2010). These actions tell us that the authorities are willing to push for development of new means of scholarly communication, inasmuch as the open access movement, though decades old, has been slow to catch on in academia, owing both to scholars' reluctance to contribute to online resources and to the publishing industry's increasing anxiety about, and resistance to, open access. At the same time, they symbolize a new trend of repository management that makes self-archiving an obligation of, rather than a plea to, individual scholars.

In archaeology, in tandem with the development of many open access initiatives over the past several years, it has become increasingly clear that a digital repository of scientific research is critically important. Various open access models have been introduced and optimized to promote information

sharing within the process of wider scholarly communication. And many archaeologists now accept the idea that the online preservation and dissemination of archaeological data should become a major priority of the discipline, as exemplified by ongoing projects carried out by the [Alexandria Archive Institute](http://www.alexandriaarchive.org),<sup>1</sup> the [West Bank and East Jerusalem Archaeological Database Project](http://digitallibrary.usc.edu/wbarc/),<sup>2</sup> and the University of Southern California's Digital Library, recent winners of the 2009 [Open Archaeology Prize](http://www.alexandriaarchive.org/openup.php).<sup>3</sup> Another area of new open access achievements is the formal publication of peer-reviewed electronic journals that are free to all, such as *Museum Anthropology Review* sponsored by Indiana University's Wells Library. Scholars have worked collaboratively with information professionals and librarians on pioneering the digital enterprise to transform critical research mechanisms in archaeology.

I have previously pointed out the necessity of developing an open access repository that can collect, preserve, and disseminate digital archaeological resources (Xia 2006). This necessity is due to the inherent nature of archaeological data. That is, because archaeological data vary considerably in their type and format (making it difficult to standardize the description of objects excavated from diverse archaeological sites), because archaeological data are collected in great quantities (such that a large proportion of archaeological material is always excluded from formal publications), and because such data take so much time to process, data sharing has long been problematic for those who would seek comprehensive analyses and interpretations. Another major factor contributing to our need for a repository is the traditional practice of archaeological publishing, with its broad scope, slow speed, limited priority, and selected distribution. Many archaeologically relevant publications are scattered among science, humanities, and history publications, making information search a challenging job. Traditionally, scholars have relied on print journals in the library of their institution or in their personal collections to access archaeological literature (Robinson and Posten 2005). But with the ever-increasing price of journal subscriptions, archaeologists' access to journal articles is further restricted, especially in the context of their already constrained resources. This does not even take into consideration the archaeological gray literature and field data which inherently have a significantly limited public access.

---

<sup>1</sup> <http://www.alexandriaarchive.org>

<sup>2</sup> <http://digitallibrary.usc.edu/wbarc/>

<sup>3</sup> <http://www.alexandriaarchive.org/openup.php>

Open access now provides an innovative way of distributing scientific data and research within the archaeological community. In the past decade, substantial efforts have been undertaken to make these resources available online at little or no cost. The advantages of digitized archaeological data and literature are multifaceted: electronic materials can be preserved long-term, research outcomes can be circulated quickly and inexpensively, and data and literature can be made available in a centralized location. Also, studies have demonstrated that open access can have a positive impact on research quality. For example, it has been shown that after a publication has been posted online with open access, its chances of being cited are greatly increased (Antelman 2004; Gargouri et al. 2010). In anthropology, open access articles have been cited more often than their non-open-access counterparts (Xia and Nakanishi 2012).

Different open access models have their own characteristics to benefit archaeological research. While a free online database can maintain raw data for scholars to download for various analyses, a peer-reviewed electronic journal can maximize the exposure and visibility of scholarly publications. At the same time, a digital repository that focuses on collecting research literature in the form of e-prints can serve as a centralized platform through which a body of scattered archaeological materials can become accessible to everyone in the world.

How to successfully manage an open access digital repository has been a hot topic for repository advocates and managers (Davis and Connolly 2007; Swan and Brown 2005). The style of repository management varies, which can be roughly categorized into those that serve an organization (institutional repositories) and those that work for a scientific field (subject repositories). However, one of the management tasks that is common to all digital repositories is to enhance user-generated content and promote the participation of scholars in self-archiving their research articles, which is an essential component of Web 2.0 (O'Reilly 2008). This article proposes the development of an open access repository for archaeological literature and discusses the foreseeable problems and solutions pertaining to its management style.

## AN OPEN ACCESS REPOSITORY FOR ARCHAEOLOGY

It is generally believed the open access movement began in the early 1990s by [arXiv](http://arxiv.org),<sup>4</sup> a subject-based repository in physics (Moed 2007). Since then,

---

<sup>4</sup> <http://arxiv.org>

repositories (also called “e-prints” or “archives”) have developed into an interactive web platform that provides users the ability to submit contributions and access research results. The advances of information technologies have made the online dissemination of free articles possible, and even easy. A repository usually has the following major characteristics:

1. *Free*: Both information acquisition and retrieval are controlled by users whose aim is to broaden information sharing. Commercial activities are restricted to a minimum.
2. *Online*: Everyone in the world can access it as long as an Internet connection is available.
3. *Scholarly*: Repository content includes research data, ideas, and formal or informal publications, mostly research-related in nature.
4. *Perpetual*: Materials, deposited in a database, may be preserved in better condition and potentially persist longer than many traditional forms of publication such as paper and microfilm, if the technology can be developed for the purpose of preservation.

Repositories are divided into institutional and subject-based, which have different sponsoring agencies and target audiences and, thus, dissimilar management styles. Subject-based repositories appeared first and have been operated in several academic disciplines, including economics, computer science, philosophy, and physics. Most of them have achieved great success. Some subject repositories have included archaeology in their content coverage; and most institutional repositories have allowed, or even required, archaeologists who work in the institution to make their research publicly available, for example, in the **Social Science Research Network**.<sup>5</sup> However, there is no single, free repository specific to archaeology and accepted by archaeologists at the national, much less the international, level.

Technologically, it is relatively straightforward to plan and implement a digital repository. Computer hardware has become cheap but powerful enough to store and maintain huge amounts of data. At the same time, computer software has become mature enough to handle data acquisition, preservation, and dissemination. Some open source applications are particularly designed to meet the requirements of digital repositories; examples include

---

<sup>5</sup> <http://www.ssrn.com>

**EPrints** by the University of Southampton,<sup>6</sup> **Fedora** by the University of Virginia,<sup>7</sup> and **DSpace** by MIT<sup>8</sup> and Hewlett-Packard (Lynch 2003). These software applications are free to download, easy to set up, and flexible enough to be customized to fit the needs of a particular environment.

Like many other subject repositories, this proposed archaeological literature repository may contain two major technological components: an online database to store and index articles, and a web interface where users can both deposit and retrieve content. To deposit, any contributor would utilize the interface, with or without login authentication, to upload a file and provide some metadata elements to the database, such as article title, author name, subject keywords, journal name and date, among others. To retrieve, a reader would be able to search the content in the database by keywords, article title, author name, and journal name, or browse by author affiliation, journal, subject, geographic location, and the like. A fancy graphical interface is not a necessity; it is more important that the database be configured to accommodate standard file formats, such as PDF (at the time of this writing). Figure 9.1 shows a sample web interface from arXiv.

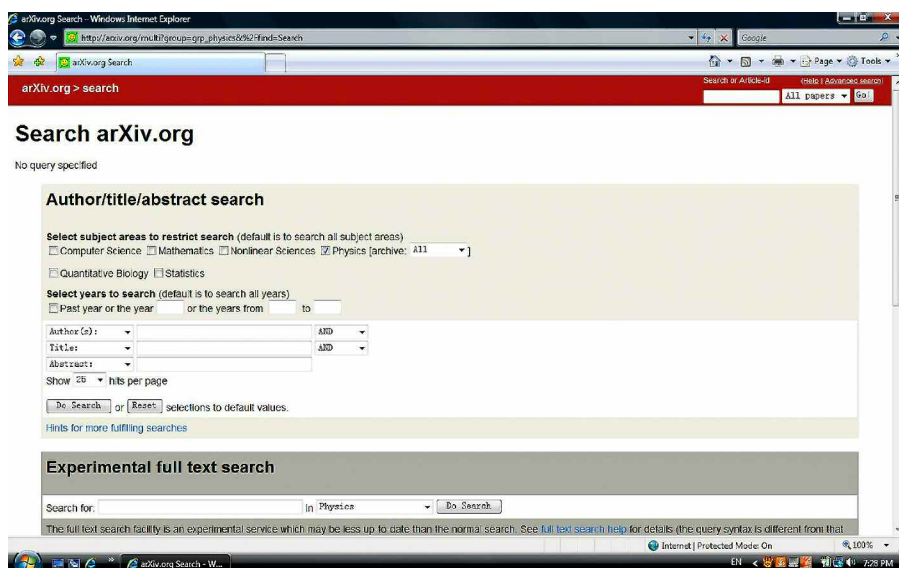


Figure 9.1. A sample web interface from arXiv.

<sup>6</sup> <http://www.eprints.org>

<sup>7</sup> <http://fedoraproject.org>

<sup>8</sup> <http://www.dspace.org>

From my experience, the style of repository management is the key to success. Because the accumulation of content materials in a repository depends on scholars who are both contributors to and readers of the repository content, it is vital to find ways to encourage scholarly self-archiving. To do this entails understanding the attitude of scholars toward making contributions to a repository. Repository advocates and managers have worked out many strategies to promote self-archiving practices among individual scholars, arguing that participation in repositories supports scholarly research. In a general sense, the more content a repository holds, the more likely it is that scholars will rely on it for their research. For an archaeological literature repository, the major concerns with regard to management may include the following.

## MANAGEMENT ISSUES

### Scholars' Awareness

It is a core task of the repository manager to let everyone know that such a repository exists, how it works, and what the benefits to individual scholars are. The repository literature (e.g., Swan and Brown 2005) has widely recognized that it is important for scholars to be familiar with a repository in order for them to participate in digital material contributions. It is believed that the traditions of information sharing within the discipline play a significant role in familiarizing scholars with the new means of scholarly communication (King et al. 2006; Lawal 2002), because “cultural norms were regularly brought up as justification for behavior” (Davis and Connolly 2007).

In the practices of repository management, several strategies have been adopted to draw the attention of scholars to self-archiving their research results. For example, some institutional repositories in Canada initiated a pilot project that focused on working with selected disciplines to collect digital content and then marketed this experience to other disciplines (Shearer 2006). In the Netherlands, repository managers targeted some of the most prominent scholars and convinced them to cooperate with repository management, hoping their acquaintance with the operation of digital repositories would influence the behavior of their junior colleagues (Feijen and van der Kuil 2005). A repository in economics—**RePec**<sup>9</sup>—has partnered with indi-

---

<sup>9</sup> <http://repec.org>

vidual departments and institutions who assist with reminding their faculty to make their new publications available in the repository (Cruz et al. 2000). A library and information science repository—**E-LIS**<sup>10</sup>—has created an editorial system that uses scholar volunteers scattered in different locations and countries to cover as wide a geographical area as possible. The archaeological literature repository would do well to adopt these as the operational models. Such strategies have proved effective and efficient in the management of digital repositories (e.g., Cruz et al. 2000; Joint 2006).

Increasing scholars' awareness of archiving can also be accomplished through other means, including publications, conference presentations, and personal communications. Given that archaeologists are accustomed to communicating with one another through personal networks, informal communication may also be an efficient way of spreading the idea of a repository's value. Obviously, the more people know of its existence and value, the more likely they are to deposit their own work there. As increasing numbers of articles become available in the repository, more people will use it to find articles and make further contributions.

Self-interest should motivate archaeologists to contribute voluntarily to a repository. Several studies have found that when an article becomes available online, its chance of being downloaded and cited by other researchers is dramatically increased over those same odds should it remain only in print (Antelman 2004; Lawrence 2003). Such findings, if widely broadcast, especially in regard to digital repositories, should encourage archaeologists to take advantage of information technology and become repository activists. Like other scholars, archaeologists also care about the visibility and impact of their research among peers, not only for the sake of their career development, but also for their self-satisfaction. There are numerous archaeological and archaeologically related journals around the world, and very few of them offer an online version free of charge. Therefore, the potential for the success of the archaeological repository, if managed properly, is out there.

A new trend of repository management is to mandate self-archiving activities, requiring scholars to make their research accessible through a repository. Mandated policies, like those mentioned above, have been implemented in some institutional repositories globally and in a subject-based repository (NIH-supported **PubMed Central**<sup>11</sup>) in the United States. Unfortunately,

---

<sup>10</sup> <http://eprints.rclis.org>

<sup>11</sup> <http://www.pubmedcentral.nih.gov>

this can hardly be a management practice for archaeology, at least in the early stage of repository development, because of the traditions within archaeology. It is known that the traditions of information sharing in a scientific discipline have a great influence over how digital repositories are developed in a discipline (Lynch 2003). Archaeological projects have diverse sponsorship, making a mandatory policy hard to implement.

## Concern over Copyrights

Scholars' resistance to self-archiving usually arises over copyright concerns: who owns a published article? Yet academic authors tend not to be especially knowledgeable about copyrights: many fail to carefully read the copyright agreement with a publisher before an article or book is accepted and published, being unaware that copyright agreements can differ substantially among publishers or even among journals published by the same organization. Some authors may simply be too cautious to post their own copyrighted articles; but some may have signed away their rights and don't even realize it. This has been an issue in encouraging scholars to contribute their articles to repositories (Gadd 2003a, 2003b).

Repository advocates and managers recognize this obstacle and have made great efforts to collaborate on solutions with the publishing industry, library administrators, as well as individual scholars. The open access movement has delivered promising results in the past years, so copyright issues are not as big an obstacle as previously thought. Now, more publishers support open access and allow proper use of their publications in varying degrees. Many of them no longer consider open access detrimental to their commercial profits, understanding that the free availability of journal articles may actually increase the "findability" of their materials and thus generate subscriptions to their journals. Publishing industry agreement on authors' self-archiving makes the accumulation of repository content simpler than ever before.

The **Romeo-Sherpa**<sup>12</sup> project in Britain documents the policies of many major publishers and categorizes them into different groupings, color-coded according to their copyright agreements. A green publisher allows the archiving of pre-prints (articles before peer-review) and post-prints (articles

---

<sup>12</sup> <http://www.sherpa.ac.uk/romeo>



after peer-review but before being formatted by the publisher). The American Anthropology Association is on the green list. Most other publishers in archaeology are also open access supporters. In other categories, the blue publishers agree to the archiving of post-prints, the yellow publishers tolerate pre-prints, while the white publishers still maintain exclusive copyrights.

In practice, repository managers can help scholars identify any copyright violation to avoid possible legal problems. For example, a repository may consider incorporating the Romeo-Sherpa list into its own authentication system to automatically verify the level of copyright agreements of an archived article. It is much easier for a subject repository (versus an institutional repository) to carry out the verification because only a set number of professional publishers need to be dealt with. This process, however, should be undertaken on the repository management side rather than on the contributors' side, since individual authors tend not to have time (or, perhaps, interest) to ascertain the level of copyright agreements of any particular journal in order to self-archive (Mackie 2004).

Self-archiving can be relatively straightforward if archaeological authors post their articles in pre-print form (except for articles by the white publishers who, according to the current Romeo-Sherpa list, are mostly clinical medical publishers). The key is to educate individual archaeologists about the difference and make them feel comfortable making contributions to an archaeological literature repository. Digital repositories have been around for almost 20 years; and we have not heard of any single instance of copyright infringement.

Archaeological literature may be in a unique position with regard to copyright issues. Privately held "gray literature" dominates archaeological publications in the form of excavation reports, fieldwork notes, working papers, and the like. Some of this material, formal or informal, may be owned by a sponsoring agency, a museum or a consulting company, rather than by an author or registered publisher. These types of output pose additional challenges with regard to making them available online free of charge. Projects sponsored by governments, Native American groups, or individuals will make this situation even more complicated. Although the general rule is to handle the copyrights of these publications case by case, it is a topic for further exploration.

## Mediated Archiving

Scholars often claim to be too busy to do self-archiving (Mackie 2004). This is more an issue for institutional repositories than for subject repositories, in that scholars are obligated to work with the former but are interested in working with the latter. In any particular discipline, scholars share the same research interests and like to share research information and results. Because of this, subject repositories, such as arXiv for physicists and PubMed Central for health scientists, have seen more success than institutional repositories (Xia 2008). An archaeological literature repository, being a subject repository, should not suffer the same usage problems that institutional repositories tend to see.

Many institutional repositories have introduced a practice known as “self-archiving mediation” or “mediated archiving,” with positive results. It began with the involvement of librarian liaisons (institutional repositories were typically operated by university libraries) who took on the task of uploading articles to repository databases on behalf of their faculty. Some institutions assigned department administrative staff to undertake archiving for their faculty. Later, students were hired to do the work. A recent survey of institutional repositories in Australia and Europe found that, in some cases, “mediated archiving” made up as much as, or even more than, 95 percent of the entire repository content (Xia 2007). Analysis of the names of article depositors in the metadata fields revealed that a very small number of people were responsible for the majority of the depositions, and these people were not article authors themselves.

This practice has changed the definition of self-archiving, as archiving is no longer by the “self.” Its pros and cons are equally obvious. On the positive side, there is a remarkable increase in the number of articles in many repositories; on the negative side, the quality of the metadata has decreased, being produced by people other than authors. Without reading an article, which is certainly common in mediated archiving, non-author depositors often have difficulty selecting the correct terms to describe the article when filling in the metadata fields. As a result, repository users have a harder time finding such an article in later data retrieval. Fortunately, this problem can be easily solved: after the initial stage, during which time repository managers face great pressures to acquire large amounts of content, they can slow down the pace of archiving activities and concentrate on improving the quality of metadata.

Subject repositories have different management styles than institutional repositories. But this does not prevent the former from adopting strategies proven useful and applied by the latter. Mediated archiving is an example. Actually, some subject repositories have already integrated third-party mediation into their practices; E-LIS is an example. It seems that scholars like this system and prefer to grant the repository managers the right to mediate their archiving rather than archive the articles themselves. The major remaining question, then, about mediated archiving in subject repositories is financial: who is going to pay for those who carry out article archiving? Unlike institutional repositories, where supporting personnel are easily available, subject repositories, because of their independent status, often survive on a shoestring. The possible answer may be volunteers who are scholars themselves and are willing to contribute their time to support the new means of information sharing. They may work for a repository for personal interests, as do those who work as journal editors and reviewers. This, of course, is easier said than done. This proposed archaeological literature repository needs a solid sustainability plan to support its implementation, maintenance, and development.

## Sustainability

Websites come and go. But we need a repository that can stay long and survive rapidly changing technologies. Both hardware and software change quickly. Today's technology for data storage may become obsolete tomorrow; and today's standards may no longer be used in a couple of years. This is an important issue for repository management and should be given careful consideration in the stages of planning and implementation. A repository is ideally housed and supported by a well-known professional organization, such as a university or a research institute, so that its sustainability in terms of financial support, personnel, and technology can be guaranteed. It should provide the necessary security and credibility to archaeologists and compel them to build a long-term relationship with the repository. Eventually (and ideally), a repository is a collaborative effort and one that benefits everyone in the field.

To measure the sustainability of a subject repository, technology, personnel, and finance, must be considered, for both the initialization of a project and the long-term maintenance of a repository. Unlike institutional

repositories which reside in organizations, subject repositories are supported by interests, requiring an entity that is both willing to commit to its long-term support and capable of keeping up the expenses for the necessary infrastructure. A mobile repository with changing hosts will easily lose the trust of scholars even if its transitions are transparent to end users.

Most subject repositories were born from grant projects, particularly with support from government agencies like the National Science Foundation (NSF), the Mellon Foundation, and the National Endowment for the Humanities (NEH). After the initial funding, those able to adopt a durable business model have survived to play a continuing and important role in scholarly communication. For example, one of the most successful repositories, arXiv, moved its residence from Los Alamos National Laboratory to Cornell University, where it will benefit from a healthier infrastructure. At the same time, many other repositories retreated from the landscape due to the lack of continuous support. Hence, the key for a subject repository to survive is the long-term commitment of an organization, one preferably rooted in its initial planning. The involvement of organizations, with their established infrastructure, can highlight the sustainability of a repository.

Another successful business model of sustaining a subject repository is operated by E-LIS. Its approach “is based in the voluntary work of an international team of information management specialists” (Rclis.org 2008). Taking advantage of the Internet, E-LIS organizers invited individuals and institutions from all over the world to join their efforts. For example, its collaborations come from national repositories on different continents and local staff repositories in many countries (De Robbio and Coll 2005). Most countries have their own local editor(s) so that the supervision of its self-archiving can be both cost-effective and management proficient. This management style provides another good example for the archaeological literature repository to follow.

Professional associations might make a perfect candidate for managing a subject repository, given their active role in facilitating professional activities and scholarly resources. Many associations have their own publishing outlets and produce influential manuscripts and leading journals in an academic field, and thus have an insight into the challenges and potentials of scholarly communication. But for unknown reasons, very few, if any, associations have become involved in the development and administration of digital repositories. Repository managers should pay attention to any possibilities of advo-

cating for open access and try to spark the interest of association leaders, most of whom are scholars, in building a subject-specific, open access database.

## Full Text Availability

Users of a literature repository are more interested in reading the full text of an article than in skimming an abstract. This has been a problem for some institutional repositories that have had to accumulate e-content material quickly but lacked the necessary time and personnel to archive all of the full articles. Their concentration on content quantity resulted in some quick actions to collect only the abstracts of articles, thus saving time and energy (and likely getting around copyright issues). This shortsightedness brings prosperity to a repository, but actually harms the development of the open access movement. It is more difficult to convince contributors who have already been disappointed to keep a long-term relationship with a repository than to educate those who are new to the practice of self-archiving and free access repositories.

An existing remedy to balance full texts and abstracts in an institutional repository is to add an external link to the outside website where the full text of the deposited item is available. However, will the full-text link lead to another free online source? And why would users bother to continue exploring a repository full of abstracts? In most cases, an external link is a dead link because either the linked website has changed its address or it requires login authentication. That kind of full-text link is worse than nothing. Studies have found that in some repositories, the percentage of abstract-only content can be greater than 90 percent (Xia and Sun 2007), and dead external links are much more common than successful links (54 percent; see Xia 2008).

On the other hand, full-text articles are the mainstream of content material in subject repositories, as the full text is generally required to make a deposit there. When scholars contribute to a subject repository, they have to be willing to make the entirety of an article available to others. Hence, were such a literature repository to become available to archaeologists, it may not suffer from this problem of full-text access. The archaeological repository, in fact, may need more than the accessibility of full texts. For example, this repository may encourage the contribution of high-quality images so that archaeological data may be used and reanalyzed by others read-

ing and downloading articles from the database. This will be one of the technical challenges to the repository developer and manager.

## Other Issues

Many other issues have emerged in previous repository management, which may or may not be applicable to subject repositories but should be mentioned. Such issues include users' concerns about plagiarism, accuracy and currency of information retrieved, versioning of articles deposited, and the complexity of the process of making self-contributions (Davis and Connolly 2007). Each scientific discipline has its own tradition of information exchange and may cope with the open access movement differently; this will create special requirements for the management of its repository. Archaeology shares some characteristics of information handling found in other disciplines, such as history and area studies in the humanities; anthropology, sociology, and geography in the social sciences; geology, chemistry, and animal studies in the sciences; as well as fields in agriculture and medical sciences. At the same time, archaeology has its own way of managing information and sharing ideas. It will be a challenging but exciting task to construct and maintain a subject repository for archaeological literature.

## CONCLUSION

Since 2004, Web 2.0 has become a new concept in online communication. Digital repositories, through their practices of self-archiving and free access, emphasize user-to-user and user-to-system interactions and exploit the real characteristics of Web 2.0. Their reliance on user-generated content will provide valuable experience for digital libraries and other types of online applications to improve their services, leading to revolutionary forms of scholarly communication.

This article has reviewed the successes and challenges of repository management in the past decade, with an eye to applying these lessons to the implementation, maintenance, and development of an archaeological repository for research literature. It has focused on such issues as making a repository known, clarifying copyright misunderstandings, maneuvering through self-archiving mediation, proposing sustainable business models for the longevity of a repository, and improving the usability of repository content, with an at-

tempt to discover the positives and negatives of the management experiences of other repositories. Following a brief analysis of these important management factors, this article went on to discuss the viability of the archaeological repository, and emphasized its significance to the advancement of archaeological information sharing and its benefits to archaeological research.

This research has its limitations. Because technologies change very rapidly, today's issues on repository development and management may no longer be significant tomorrow. Also, archaeology is different from other scientific fields in information exchange; and thus its repository may be unique in its management needs. Nonetheless, I believe an understanding of the general concerns of other repositories offers useful lessons that are applicable to archaeological digitization.

## ACKNOWLEDGMENTS

The author thanks Eric Kansa and Sarah Whitcher Kansa for the invitation to the wonderful discussion at the SAA Annual Meeting in Vancouver, Canada, and for providing constructive comments on the early drafts. The revision of this paper also benefited from a casual talk with Fred Limp of the University of Arkansas during the meeting, for which the author is grateful.

## REFERENCES CITED

- Antelman, K.  
 2004 Do Open-Access Articles Have a Greater Research Impact? *College & Research Libraries* 65/5: 372–382.
- Cohen, P.  
 2008 At Harvard, a Proposal to Publish Free on Web. *The New York Times*, February 12, 2008. Retrieved from [http://www.nytimes.com/2008/02/12/books/12publ.html?\\_r=1&oref=slogin](http://www.nytimes.com/2008/02/12/books/12publ.html?_r=1&oref=slogin) (accessed July 16, 2010).
- Cruz, J. M. B., M. J. R. Klink, and T. Kriche  
 2000 Personal Data in a Large Digital Library. *Proceedings of the 4th European Conference on Research and Advanced Technology for Digital Libraries*. Retrieved from <http://openlib.org/home/krichel/phoenix.a4.pdf> (accessed July 16, 2010).
- Davis, P. M., and M. J. L. Connolly  
 2007 Institutional Repositories: Evaluating the Reasons for Non-Use of Cornell University's Installation of Dspace. *D-Lib Magazine* 13, no. 3/4. Retrieved from <http://www.dlib.org/dlib/march07/davis/03davis.html> (accessed July 16, 2010).

De Robbio, A., and I. S. Coll

- 2005 E-LIS: An International Open Archive towards Building Open Digital Libraries. *HEP Libraries Webzine* Issue 11. Retrieved from <http://library.web.cern.ch/library/Webzine/11/papers/1/> (accessed May 3, 2011).

EPrints.org

- 2008 ROARMAP (Registry of Open Access Repository Material Archiving Policies). Electronic document. Retrieved from <http://www.eprints.org/openaccess/policysignup/> (accessed July 16, 2010).

Feijen, M., and A. van der Kuil

- 2005 A Recipe for Cream of Science: Special Content Recruitment for Dutch Institutional Repositories. *Ariadne* Vol. 45. Retrieved from <http://www.ariadne.ac.uk/issue45/vanderkuil/> (accessed July 16, 2010).

Gadd, E., C. Oppenheim, and S. Proberts

- 2003a RoMEO Studies 1: The Impact of Copyright Ownership on Academic Author Self-archiving. *Journal of Documentation* 50/3: 243–277.  
2003b RoMEO Studies 4: An Analysis of Journal Publishers' Copyright Agreements. *Learned Publishing* 16/4: 293–308.

Gargouri, Y., C. Hajjem, V. Larivière, Y. Gingras, L. Carr, T. Brody, and S. Harnad

- 2010 Self-Selected or Mandated, Open Access Increases Citation Impact for Higher Quality Research. Retrieved from <http://arxiv.org/abs/1001.0361> (accessed July 16, 2010).

Joint, N.

- 2006 Institutional Repositories, Self-Archiving and the Role of the Library. *Library Review* 55/2: 81–84.

King, C. J., D. Harley, S. Earl-Novell, J. Arter, S. Lawrence, and L. Perciali

- 2006 Scholarly Communication: Academic Values and Sustainable Models. Report, Center for Studies in Higher Education, University of California, Berkeley. Retrieved from [http://cshe.berkeley.edu/publications/docs/scholarly\\_comm\\_report.pdf](http://cshe.berkeley.edu/publications/docs/scholarly_comm_report.pdf) (accessed July 16, 2010).

Lawal, I.

- 2002 Scholarly Communication: The Use and Non-Use of E-Print Archives for the Dissemination of Scientific Information. *Issues in Science and Technology Librarianship*, Fall. Retrieved from <http://www.istl.org/02-fall/article3.html> (accessed July 16, 2010).

Lawrence, S.

- 2003 Online or Invisible? *Nature* 411/6837: 521.

Lynch, C.A.

- 2003 Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age. *Portal: Libraries and the Academy* 3/2: 327–336.

Mackie, M.

- 2004 Filling Institutional Repositories: Practical Strategies from the DAEDALUS Project. *Ariadne* Vol. 39. Retrieved from <http://www.ariadne.ac.uk/issue39/mackie/> (accessed July 16, 2010).



Moed, H. F.

- 2007 The Effect of “Open Access” on Citation Impact: An Analysis of ArXiv’s Condensed Matter Section. *Journal of the American Society for Information Science and Technology* 58/13: 2047–2054.

O’Reilly, T.

- 2008 What is Web 2.0? Retrieved from <http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html> (accessed July 16, 2010).

Rclis.org

- 2008 *Who We Are*. Retrieved from <http://rclis.org/about.html> (accessed May 3, 2011).

Robinson, W. C., and P. E. Posten

- 2005 Literature Use of Scholars Publishing in Leading Anthropology Periodicals. *Behavioral & Social Sciences Librarian* 23/2: 1–17.

Shearer, K.

- 2006 The CARL Institutional Repositories Project: A Collaborative Approach to Addressing the Challenges of IRs in Canada. *Library Hi Tech* 24/2: 165–172.

Suber, P.

- 2008 *Open Access News*. Retrieved from <http://www.earlham.edu/~peters/fos/fosblog.html> (accessed July 16, 2010).  
2010 *Open Access News*. Retrieved from <http://www.earlham.edu/~peters/fos/fosblog.html> (accessed July 16, 2010).

Swan, A., and S. Brown

- 2005 Open Access Self-Archiving: An Author Study. Report, Key Perspectives Limited. Retrieved from <http://eprints.ecs.soton.ac.uk/10999> (accessed July 16, 2010).

Xia, J.

- 2006 Electronic Publishing in Archaeology. *Journal of Scholarly Publishing* 37/4: 270–287.  
2007 Assessment of Self-Archiving in Institutional Repositories: Across Disciplines. *Journal of Academic Librarianship* 33/6: 647–654.  
2008 A Comparison of Subject and Institutional Repositories in Self-Archiving Practices. *Journal of Academic Librarianship* 34/6: 489–495.

Xia, J., and L. Sun

- 2007 Assessment of Self-Archiving in Institutional Repositories: Depositorship and Full-Text Availability. *Serials Review* 33/1: 14–21.

Xia, J., and K. Nakanishi

- 2012 Self-Selection and the Citation Advantage of Open Access Articles. *Online Information Review* (forthcoming)



## WHAT ARE OUR CRITICAL DATA-PRESERVATION NEEDS?

*Harrison Eiteljorg*

The excavation or survey is over. The information has been properly entered into a database; the CAD models and/or GIS data sets have been accurately completed; the scientific studies of samples have been completed. Now what?

Now, of course, comes the hard part—the long, slow, painstaking process of analysis. This includes analyzing each of the artifact categories, the architecture, the stratigraphy, and any samples subjected to some form of scientific study. Most important of all, though, is the analysis of the site or project as a whole. Those analyses will have been going on for some time at various levels, in both formal and informal ways. However, they will now have to be both more formal and more comprehensive.

As challenging as the analyses are, getting the combination of information and analysis into proper publication form, whether paper or electronic, is even more difficult, more time-consuming, and more complex, because it requires managing the most unmanageable of archaeological assets: the personnel who have done all the hard work over the years but want to put pen to paper no more eagerly than the director.

### DATA PRESERVATION: THE LAST STEP

The work at issue in this article comes *after* the analysis and publication; it is the process of dealing with all the data files at the very end of this process, after the publication has been finished. The foregoing short and terribly oversimplified list of things that must be done after the physical work of the project has been completed was used as the starting point here to remind the reader that the focus on the data files comes at the end of a long, arduous

road. The planning has been completed; the physical work of survey and/or excavation has been completed; the data recording has been completed; the object conservation has been completed; cleaning up and otherwise preparing the data files for use in analysis and publication have been completed; the analytic processes have been completed; the publications have been completed, or at least are near enough to completion that the project director(s) can move on to other matters. All this work precedes the final disposition of the data files, and publication is not only the last of those jobs leading up to data disposition but the most difficult and often the least rewarding of them all.<sup>1</sup> (I understand that publication may include public presentation of the data files and will address that possibility.)

What does the disposition of the data files involve? That is a harder question to answer than it may seem, because a good, full answer presumes an understanding of the requirements of any individual repository where files are to be sent.<sup>2</sup> While there are individual differences, there are really two approaches to data disposition that are important to this discussion.

## DATA DEPOSIT IN ARCHIVAL REPOSITORIES

An archival repository is intended to hold original project data files, as complete files, and then to migrate them from one file format to another as required so that their utility remains intact when software requirements change. The repository will take the files, copy them onto whatever media are used internally, presumably announce their presence and availability, index them in some fashion, and present them to the public as complete digital files according to some established guidelines. Such a repository will also

---

<sup>1</sup> By the time the data have been analyzed, there may be conflicts concerning intellectual property rights to the data, with project director and specialists contending as to whose property the files truly are. These and other issues of intellectual property rights will not be considered here. See Kansa, Bissell, and Schultz (2005) for a discussion of the complications of multiple stakeholders and intellectual property.

<sup>2</sup> The issue of repository demands also raises the added question of file documentation: has the documentation of the files been completed? If so, that would mean that descriptions of the file structures, storage systems, version controls, and the like, have been kept current throughout the life of the project. Unfortunately, that level of project documentation is rare unless the project in question is large enough to support a permanent computing staff. For most projects, then, documentation will be required prior to deposit. (Documentation need never be done if the files are simply discarded or orphaned, a possibility that, sadly, a realist must include in the realm of likely outcomes but one that will not be discussed here.)

migrate any proprietary files into nonproprietary formats unless that is impossible (e.g., CAD files can be virtually impossible to migrate to a nonproprietary format without a loss of information). That initial data migration is probably simple enough that it can be completely automated by the project team; the repository could also automate the process if it has the requisite software.<sup>3</sup>

This final process of data migration to nonproprietary formats may require some additional documentation, but that should be easily completed since it should involve no more than adding to the extant documentation the facts of the migration processes. The repository may also choose to migrate some files immediately for the sake of easier access.<sup>4</sup>

The end result is similar to a traditional paper-based archive. The material is available for inspection, but it is up to the individual researcher who comes to the archives to know how to gather information, understand it, and perform some kind of synthesis.

## DATA DEPOSIT IN WEB 2.0 REPOSITORIES

Web 2.0 repositories, the second of our two general repository types, aim to share individual items of data, rather than whole files, with scholars over the Web. These repositories have long been the pot of gold at the end of the digital rainbow: the easily accessible storehouse for archaeological data that makes it possible for virtually anyone to locate any item of data easily, reliably, and quickly.

A Web 2.0 repository must create an organizational schema for all data contained in the repository, so that access systems can function effectively.<sup>5</sup>

---

<sup>3</sup> Having just completed the preparation of archival materials from the CSA Propylaea Project, I must acknowledge that the work was more time-consuming and more intellectually demanding than I expected. Much of that was because some data files were still incomplete and I had not been diligent about preparing documentation as the project moved along, but the job was harder than expected. It took a few months to complete rather than the expected few weeks.

<sup>4</sup> For instance, the Archaeological Research Institute at Arizona State University, when preparing photographs from the author's older propylon work for presentation, migrated the images from uncompressed TIFF format into individual PDF files so that they could easily be magnified online and would not require additional programs for manipulation.

<sup>5</sup> Of course, that does not mean that the organizational schema must be unique. A repository might—and arguably should—agree to use a standard created via cooperative processes.

To be accessible, the data items in the repository must match the file formats and organizational schemata required by the repository's internal software. Today that means that the supplier of the data must move the individual data items from their original files and organizational schemata into new forms, both new file formats and new organizational schemata. To take the simplest of examples, that might mean that all photographs need to be converted to PDF or JPEG format<sup>6</sup> so that they can be used on the Web; data about them would also need to be put into forms expected by the repository. The data accompanying any set of images in a repository may or may not match the data recorded by the project, of course. A project team may have documented the name of the photographer or camera or even the lens used, but it is hard to imagine such information being of interest to a Web 2.0 data repository. (That does not mean that the information could not be included. However, the depositor would be required either to create a system tailor-made for submitting data not included in the depository's schema—a new XML schema, for instance—or to simply dump the data into a catchall miscellaneous category. The first solution adds yet more work to the depositor's schedule; the latter makes the information harder to find and use.)

Conversion to JPEG or PDF format may or may not already have been done for internal reasons, but—a very important but—a proper archival process for a project should retain all photographic files in their original formats to prevent loss of information, especially if those original formats are TIFF or one of the proprietary camera formats that permit particularly effective manipulation of the images with modern photo software. In either case, one can confidently expect that better versions of the images and better enlargements of portions of them will be possible in the future if the original image files have been retained. Whether or not the images should be culled so that duplicates or low-quality images are discarded is a practical issue; whether original formats should be retained is not. It is a theoretical judgment that, in my view, requires the preservation of original image files, not just JPEG derivatives. Web 2.0 repositories, being focused on access to data, may or may not be prepared to accept the original image files as well as JPEG derivatives, just as they may or may not include all the data categories attached to the images in the project data set.

---

<sup>6</sup> The extent of file compression is an issue that may be ignored for the purposes of this discussion.

It may also be argued that transferring data items into the data-organization schema of a repository—when that schema is different from the one used in the project database—is similarly a critical theoretical issue, not just a practical question. Indeed, that is an argument I would make. If a project team has carefully and methodically created a data-organization schema that reflects accurately the processes used in the field, then requiring the data to be put into some new organizational schema is not just a simple, practical choice; it is a theoretical one that violates or at best wastes the organizational work of the project digital data team. To give one concrete example, a project database may (in my view, should) have been organized to provide explicit ways to honor scholarly disagreements among project participants—e.g., room A in structure B may be treated as belonging to the first phase of the structure by one scholar and as belonging to the second phase by another. Choosing to organize and preserve information in such a fashion will be extremely valuable to anyone accessing the data years after the completion of the project (especially since it is unlikely that publications will discuss such issues), but having such a complex data-organization schema in a Web 2.0 resource will surely not be universal.

I would also make the argument that anyone from outside a project must understand the excavation or survey processes to understand the data fully. Therefore, having the data removed from the organizational schema that expressed the work processes is not a neutral act but a disabling one; it makes certain kinds of understanding more difficult, if not impossible. At the least, any process that involves moving data from one schema to another should be thoroughly documented, and the documentation should then become a part of the repository so that any user has access to it. Accessing data without documentation—as, for instance, the “DT Death Pit Human Bones” table in *Open Context*<sup>7</sup>—provides the user with no information about the available data-entry terms, limits, defaults, and so on. These are not trivial omissions but, on the contrary, inform users of any data set in critical ways.

These problems with organizational schemata and individual data items may be far less significant if only the data about *artifacts* are to be put into Web-accessible form. If, for instance, pottery files are to be translated into some new form for web access, it is certainly possible to imagine a file structure

---

<sup>7</sup> <http://opencontext.org/tables/c019a92edd3bb2f4e400c3d7df8f4663>

that would not present a particular problem for data access (though the issue of scholarly disagreement would often remain a problem). Information about trenches, cuts, loci, and lots might then be left out of the process altogether. Approaching the data in this way, however, creates other problems. Either a portion of the project data—the information about trenches, cuts, loci, and so on—is to be ignored altogether, which I take to be an unacceptable alternative, or the artifact data would need to be deposited twice, once as individual data items translated into the new schema of the web resource and a second time as part of the original, complete data set, including all the non-artifact data. In the latter case, the artifact data would be available through the Web 2.0 repository *and* as part of a coherent whole available for downloading and ultimate reuse as a unified database.

The need to preserve files in two different ways is not, however, the only problem with separating data access from data organization. There is the critical problem of controlled vocabulary. How useful is it to have a huge ceramics data set if the individual items have not all been described with the same controlled vocabulary? Many people believe this problem can be overcome without great difficulties. I am not one of them.

There are certainly areas—chipped-stone tools, perhaps—that have and use well-controlled vocabularies. However, there are many more areas for which the vocabulary is controlled only by what might be called consensus: general agreement as to terms and definitions but no specific, documented source for the terms and definitions, and the consensus, being informal, is not consistent from time to time and place to place. Definitions and usages change and reflect idiosyncratic local uses. The problems with terminological consistency, of course, do not include the problems introduced by multiple languages. An example of this problem, without the intrusion of language issues, is the term *amphoriskos*. An *amphoriskos* is a miniature amphora, and that definition is fairly standard, with no attempt to specify a boundary between full-sized and miniature (the definition provided at the [Louvre's web-site](http://www.louvre.fr/llv/glossaire/detail_glossaire.jsp?CONTENT%3C%3Ecnt_id=10134198673228616&FOLDER%3C%3Efolder_id=9852723696500935&bmLocale=en)<sup>8</sup> is: “A miniature amphora with two side handles, used for storing perfumed oil.”). The [ArchaeoWiki site](http://www.archaeowiki.org/Amphoriskos),<sup>9</sup> however, calls the *amphoriskos* “short to very short” and then defines *short* as 15–25 cm and *very short* as under 15

<sup>8</sup> [http://www.louvre.fr/llv/glossaire/detail\\_glossaire.jsp?CONTENT%3C%3Ecnt\\_id=10134198673228616&FOLDER%3C%3Efolder\\_id=9852723696500935&bmLocale=en](http://www.louvre.fr/llv/glossaire/detail_glossaire.jsp?CONTENT%3C%3Ecnt_id=10134198673228616&FOLDER%3C%3Efolder_id=9852723696500935&bmLocale=en)

<sup>9</sup> <http://www.archaeowiki.org/Amphoriskos>



cm on [another page](#).<sup>10</sup> These definitions are not contradictory, but neither are they equivalent.

If and when such aggregated data resources exist—however the vocabulary problems have been handled—there will be many separate aggregations of data. That is, it is simply not within the realm of the possible, much less the probable, that all archaeological data relevant to a specific time/place/culture will be in a single repository. So all the problems encountered will have to be solved on the individual repository level and then solved again on some über-level; only then will users be able to access data from multiple repositories transparently. Yet, as difficult as it may be to imagine data from a vast number of multiple sources in a single, easily accessed form, it is even more difficult to imagine using data from multiple aggregated data sources and expecting some sort of terminological consistency. Indeed, it seems far more reasonable to expect a terminological nightmare.

If data tables (or data that could be in tabular form) were the only issue, one might press on, assuming that the vocabulary issues could be addressed in some acceptable way via look-up tables of equivalencies or some other system. But what does one do with CAD models and GIS data sets, not to mention those image files already mentioned? CAD models cannot be effectively used on the Web (or even at one's own computer without the right software), but they can be shared via file transfer (for use only with the right software). GIS data are more readily shared, but there are serious questions about suitability of individual maps, for instance, for purposes other than those for which the particular scale and inclusions were chosen.

The picture of a “Web 2.0” resource as I’ve described it is an idiosyncratic one in the sense that I am far more pessimistic than many regarding the possibilities for achieving some form of vocabulary control after the fact. I am also more pessimistic about the usefulness of resources that are inherently dissimilar because they are derived from data collected in different ways by different people at different times and with different purposes.

## PRESERVATION IN A WEB 2.0 REPOSITORY

There has been an unstated assumption in the foregoing that simple preservation of the original data is not at issue in a Web 2.0 repository. That is not

<sup>10</sup> [http://www.archaeowiki.org/Jar\\_size](http://www.archaeowiki.org/Jar_size)

a warranted assumption. Many Web 2.0 repositories are more focused on sharing data than preserving data.<sup>11</sup> That means that they may not accept as their responsibility either the archiving of original files or the separate archiving of data contributed by each individual scholar as those data are brought into the corpus. (For instance, the Mission page of the [Alexandria Archive Institute](#)<sup>12</sup> included, on August 3, 2010, many statements about access to information, but none about preservation of that information. The Vision and Values pages similarly discuss access but not preservation.)

Some have argued that making data available on the Web is itself an act of preservation by virtue of its being copied and kept by many people in many places. This idea even has its own acronym, **LOCKSS** (for Lots of Copies Keep Stuff Safe),<sup>13</sup> though the more general term is *distributed preservation* or *distributed archiving*. On the strength of this idea, some have been motivated to create standards, including standards for long-term preservation (via file migration). However, the goal seems to be to develop automated approaches, and that, in turn, means that there will be a significant time lag between the desire and the achievement.<sup>14</sup> While a proper implementation of the LOCKSS idea includes real steps to preserve the utility of files, it rests on an optimistic assumption that automated ways to accomplish preservation/migration will be developed.<sup>15</sup> Until those automated preservation schemes

---

<sup>11</sup> To the best of my knowledge, there is no promise of true archival preservation of the data at all in Web 2.0 repositories, simply presentation plus backup and redundancy to prevent accidental loss via some computer malfunction and reliance on distributed preservation. I do not know what measures, if any, are taken by Web 2.0 repositories to preserve data in the event of the complete failure of the operation. That is, what would happen to the data in a Web 2.0 repository if the whole endeavor were to come to a halt? A traditional archival repository, on the other hand, has a legal responsibility to retain the material in perpetuity, if necessary by transferring ownership.

<sup>12</sup> <http://alexandriaarchive.org>

<sup>13</sup> <http://lockss.stanford.edu/lockss/Home>

<sup>14</sup> See, for instance, Abby Smith's (2006) discussion of the National Digital Information Infrastructure and Preservation Program (Distributed Preservation in a National Context, *D-Lib Magazine* 12/6 [June 2006] at <http://www.dlib.org/dlib/june06/smith/06smith.html>).

<sup>15</sup> The LOCKSS project should be distinguished [<http://www.lockss.org/lockss/Home>]. This is a more formal approach but is concerned with library resources, not born-digital scholarship.

have been developed for all the file types of concern to archaeology, from simple photographs to complex CAD and GIS data sets, the LOCKSS notion is not a safe one for the discipline.<sup>16</sup>

There is also a serious problem with distributed archiving if it involves multiple copies of critical files. In a distributed system, whether or not it works as expected, there is no longer such a thing as version control. That is, the data obtained has no guarantor, no seal of authenticity indicating where it was produced, by whom, when, for what purposes, or under what conditions. When I think of the documents I have on my own computer that represent steps in a long process ending in a given document but with many intermediary versions leading up to that final stage, I cannot but recoil at the thought of relying on files saved by chance and by users who have no reason to care about—much less time or effort to spend—accurately preserving the information as gathered at the point of origin. I do not want to be in a position of relying on a transmission system that starts with data from a project, passes the data through an unlimited, undefined, and unmanaged series of users, and then sends the data to me for further analysis. Do you? Indeed, I see this as a disabling problem for those Web 2.0 repositories that lack an explicit and well-designed system for preserving data as originally placed in the system, contributor by contributor. Scholarship depends upon a chain of evidence that makes clear the origins of, and modifications to, data as well as the names of those who have been involved in all processes; distributed archiving threatens that chain.

Let us assume for a moment that I am what I have been taken to be on occasion: a curmudgeon who refuses to move into the next phase of the computer revolution. Let us further assume that the problems discussed so far are either solved or never really existed or were simply figments of an overactive imagination. In that case, . . .

---

<sup>16</sup> Obviously, distributed archiving without the preservation work is unacceptable, but the Internet Archive [[www.archive.org](http://www.archive.org)] offers just such an approach via the so-called Wayback Machine [[www.archive.org/web/web.php](http://www.archive.org/web/web.php)] which purports to supply archival preservation simply by keeping copies of materials that were on the Web. That is not archiving to satisfy the needs of the discipline of archaeology. A file from 1995 may or may not be useful to someone needing the information 20, 50, or 100 years later..

## REWARDS AND PUNISHMENTS

If the optimists are correct, and good, useful, robust, properly archived Web 2.0 resources are realistic possibilities, there is another critical question. Will scholars contribute their data? Indeed, this discussion began with the litany of jobs awaiting the scholar who has completed a project. Preserving the data was the last of those jobs, but we moved to the problems and difficulties of preserving rather than going on to consider this equally important question: What are the rewards and the penalties at stake for spending the time, effort, and money required to make certain that data have been preserved?

Let us begin with the easier category, rewards. There are only psychic rewards unless one works in an environment that requires data preservation. No colleague will praise you for archiving your data; no appointments committee will advance you to the next level in your university because of it; no professional organization will celebrate your achievement; no funding agency will give you your next grant because you archived the data from the last one. Those are broad and unqualified statements that, because they are unqualified, cannot be completely correct. In general, however, I think they are accurate today; I can only hope that the situation will change in the long run.

Penalties are easier to find—but not the kinds of penalties one imagines in the carrot/stick approach. There are few penalties for those who do not preserve their data. This is changing but not fast.

There are, though, many penalties for those who choose to preserve data by moving them to a Web 2.0 repository. First, there is the loss of time. It will take time, a good deal of time, to move data from the extant format(s) to those required by a Web 2.0 repository, even if tools to assist are available. Second, there is money. Not only is time money in the usual sense, but the odds are slim that the scholar in charge will have the technical expertise to make the translations required to move data from one system to another, in which case expert help—someone else's time—will be required. Expert help must be paid for. Third, there is the seemingly simple problem of finding personnel to assist who have both the necessary technical skills and the familiarity with archaeology required for them to understand the translation problems and the necessity of certain kinds of solutions. It seems a simple problem but is not. (For instance, a page accessed from [Open Context](http://opencontext.org/about/services/)<sup>17</sup> on

<sup>17</sup> <http://opencontext.org/about/services/>

August 3, 2010, contains the following computing acronyms and suggests that the use of each is understood by a potential user: XHTML, KML, GeoRSS, JSON.) Fourth, and probably most important, is the difficulty of maintaining intellectual focus. At the end of a long, complex process, it is increasingly difficult to continue to keep the kind of intellectual focus required to push on to completion. As the work nears its end and the final goal is in sight, one is more and more ready to move on to something else. The intellectual effort required, however, to translate the data will be significant, even with expert technical help, and that intellectual effort will be hard to summon up at the last stages of a project. Publication, the prior stage, will also have sapped much of the energy that might otherwise be available.

The first preservation possibility mentioned above—archiving original project data sets without providing access to individual data items on the Web—does not involve the difficulties discussed in connection with the Web 2.0 world. That is, providing data files and migrating them to nonproprietary formats, assuming properly documented files, should not be so time-consuming, expensive, or demanding of intellectual focus as to render the job too difficult. The rewards, on the other hand, are the same. That is, the psychic rewards are the same if the project director takes some pride in the fact that the files have been preserved, and the external rewards, when they come, will likely be equivalent. Funding agencies, appointments committees, and permit-granting entities are likely to be as impressed by one's placing files in a relatively passive archival setting as by putting the data into a Web 2.0 resource. At present, neither process will be credited by anyone, and in the future I suspect that they will retain similar status, although it is not unreasonable to expect that some people or committees may be more impressed with preservation that includes easier systems of access.

## PRESERVING DATA WHILE PUBLISHING

If data preservation in a Web 2.0 repository were accomplished in tandem with the publication process, there would be some advantages and disadvantages. One advantage arises from simply moving the processes forward in time. That would prevent the loss of focus and time at the end of the project. In addition, the time required would be less concentrated and less noticeable. On the other hand, preserving the data while preparing individual publications would mean working only on small groups of files at one time,

not the entire data set. This is a significant problem, especially if the data organization is really tight; the tighter the organization, the more critical the problems. In addition, errors are likely to be found as data files are related to other data files in the final analytic processes. If individual files are preserved as the publication process goes forward, it is likely that more errors will exist in those files—errors requiring correction later—than would have been the case with a preservation process at the end of the analytic process.

## CONCLUSION

To recapitulate, then, I believe an argument can be made that, despite their potential desirability, Web 2.0-style resources are inherently problematic for archaeological data for the reasons laid out above. I believe also that the rewards for the scholar who makes individual data items available on the Web are few and far between while the costs are very high. Finally, I believe that the rewards are as high and the costs much lower if files are simply archived and made available only as downloadable data sets. In addition, as noted above, such downloaded data sets do not demand that scholars who have carefully and methodically created well-organized and integrated data sets discard that organizational work in order to make the data publicly available.

A related point must be made. In 1994, a group of scholars initiated the Archaeological Data Archive Project, which I directed, for the purpose of providing an archival repository for archaeological data files. The project continued until 2002 when it was discontinued because it had failed to attract a single completed data set other than the one taken from a CD already published by the scholars involved. Not one. We did not even have a serious inquiry from a completed project other than those with whom we initiated contact. It may well be that we approached the task poorly, and I would not defend some of my own choices in retrospect. It may also be that the archival project was ahead of its time, an argument that has some validity. But we asked almost nothing of data contributors; we tried to make it clear that we would do all the “heavy lifting” if necessary. On the other hand, Web 2.0 projects will require an enormous investment of time, talent, money, and energy. It seems to me unrealistic to expect that a more complex and costly approach offering no better rewards will succeed where a simpler and much less costly one failed.

Whether or not Web 2.0 projects are launched and succeed in attracting data, one thing seems clear to me at this point. Data files are being lost with some regularity; the resulting loss of information for which we have paid dearly and for which there is no replacement is unconscionable.

Given the urgency of the problem, I see a need for a two-track approach. Let the Web 2.0 projects accept original project data sets—with many different kinds of data files—for the simple, relatively passive archival preservation process defined above (with a legal requirement for preservation in perpetuity, as with paper-based archives). Let them continue to work on the processes, technologies, and intellectual problems related to vocabulary and access; on access issues surrounding CAD models, image files, and other, similar problems; and on the archival processes required to maintain version control. Then, when the time has come, let the data in their possession be translated into the necessary schemata to be accessible as individual data items (while held in the original forms as well). This seems to me to provide the best of both worlds, by making preservation easier and less costly until such time as the benefits of the more difficult and costly approach can be seen, measured, compared, and accepted by the archaeological community. It should help overcome the costs and problems generated by Web 2.0 deposit requirements and, one hopes, get more data files into good repositories before they are irretrievably lost. If the most critical problem for archaeological data is their potential loss, this two-track approach seems both a realistic solution and a forward-looking one.

The two-track approach has a clear downside: it adds to the costs by making archival preservation a two-step process. Who will pay? Had I an easy answer to that question, I would be restarting the Archaeological Data Archive Project, not sitting at a computer writing this paper. However, if there are organizations willing to fund Web 2.0 projects, those same organizations should be willing to undertake preservation now and access when it is more realistic.

## REFERENCES CITED

- Kansa, E. C., J. Schultz, and A. N. Bissell  
2005 Protecting Traditional Knowledge and Expanding Access to Scientific Data: Juxtaposing Intellectual Property Agendas via a “Some Rights Reserved” Model. *International Journal of Cultural Property* 12: 285–314.

Smith, A.

- 2006 Distributed Preservation in a National Context. *D-Lib Magazine* 12/6 (June),  
retrieved from <http://www.dlib.org/dlib/june06/smith/06smith.html>.



## CONCLUSION

# WEB 2.0 AND BEYOND, OR ON THE WEB NOBODY KNOWS YOU'RE AN ARCHAEOLOGIST

*W. Fredrick Limp*

The papers in this volume address, in a variety of ways, how Web 2.0 technologies may be of use to archaeologists. The specifics of that mandate need to be kept clear. We are not simply looking at interesting ways in which Web 2.0 technologies and methods can be used by archaeologists, but how may they be *of use*. The usefulness of any technique or method to archaeology, or to any field, is not a simple assessment of the value of that method (or technology) but an assessment of it in the context of archaeology and the benefits that derive from its use. Also of consequence is the social setting in which this process takes place. This latter factor can't be overemphasized. As much as we might like to have the "best technology" win, the history of technology adoption makes it quite clear that there are many complex relationships of power, socialization, and "network externalities" (in economic terms, cf. Katz and Shapiro 1986) that strongly affect what technology is finally adopted.

Hopefully, it goes without saying that we must not become enamored by technology for technology's sake—or at least not too enamored; after all, there is some value in bright, shiny things. But at the end of the day, the question is how the archaeological enterprise is improved. Can we do things faster or cheaper, can we answer questions we could not answer before, or can we empower scholars who were previously marginalized?

In the following I want to first consider the following key topics:

- ▶ What is Web 2.0?
- ▶ How can/will it affect archaeology?
- ▶ As archaeologists, why should we care? Will it affect us? If so, how?

## WHAT IS WEB 2.0?

It seems useful to begin our discussion by focusing first on what it is we're talking about. The papers in the volume have addressed the definition of Web 2.0 in a number of ways, but it seems useful to again consider defining it. As we've have seen, the term was (apparently) first applied in 2004 by Tim O'Reilley (2006) when he defined it as a "business revolution in the computer industry caused by a move to the internet as a platform." Since then, people such as (Sir) Tim Berners-Lee (known as the inventor of the World Wide Web) have argued that term is meaningless, since the technology components have (essentially) been in place since the early days of the Web. There are many other perspectives, and there is even some discussion as to whether the term reflects real technology development or is just market hype.

I personally believe that there is something that is happening that is real, not just hype, but I believe it is essentially more real as a *social* and *institutional* process and not (so much) as a new technology. It seems to me that the key elements are:

- ▶ Separation of content from its representation and use/reuse
- ▶ Fast interactive pages
- ▶ Architecture of participation
- ▶ Rich user experience

The last three properties seem fairly clear, but it may be useful to spend a bit of time explaining the first, separation of content, because it is a fundamental property of all Web 2.0 efforts. In the old days of computing, a vertical solution would be developed by writing a specific body of code to access a defined data set, perform a specified set of operations, and present the result to the user. This solution was (usually) fast and could limit any unintended damage that inexperienced or malicious users could wreak—in short, it gave one control. It also meant that each data set often had a different interface, with different operations and tasks to learn, and it was very difficult to move data from one analytical environment to another. Data was stovepiped, and comparative studies or even simple integration was at best tedious and at worst impossible.

Web 2.0's property of separating content from its representation means that different processes can access a range of data sets or multiple ones at the same time. It is not enough, however, to separate these; it is also necessary to provide defined specifications on how they can be accessed, queried, and so on. In a simple example, using the well-known spreadsheet paradigm, a Web 2.0 application would need to tell potential users what each column contains, its properties, and the like. "Telling" in the Web 2.0 sense does not mean simply publishing a detailed description of the coding book, but involves digitally exposing the rules and structures in such a way that other programs can automatically access and "understand" these data. When this is done, users can develop programs that do interesting, often unexpected things with this data. Data from multiple, previously unconnected sources can be accessed and manipulated. Generally the term used for the programs that do this sort of thing is "web services." In addition to publishing the specifications of the data, web services developers publish their input-output specifics—that is, the external connections to their programs (frequently called APIs). It is not necessary that they make public the interior properties, only what they accomplish; thus it is possible for a developer to keep her or his intellectual property while making the process public. Finally, the presentation or the "results" are also separated, and different ways to present the results can be developed by different people. Again, the input-output specifications must be published so that other programmers can "access" the results. Perhaps the output of the statistical program in one presentation service might be a simple graph, but in another it might involve integration into a map.

The formal separation of content from the ways in which it is utilized has major implications for archaeology. Very specific techniques have been developed to accomplish these goals, but I believe that the *goals*, rather than the techniques, are the essential elements for archaeologists to focus on. By taking this tack, I assume that there is a sizable pool of innovation and that the "success" of an innovation is set by its social/economic value. This perspective is very different from one that sees innovation as essentially limited and would therefore privilege innovation as "making change possible." I think archaeologists can make this argument for many early societies, but in my view there is ample evidence in today's high-tech world that innovation is massive and that it is the process of innovation *uptake*, not its creation, that is the critical constraint—at least in archaeology. A further corollary is the assumption

that, in general, our information technology (IT) tools (and specifically those for Web 2.0) are useful and applicable to archaeological data. This is not a given, but it does seem to be the case. By way of a counter example, we can consider some IT developments that have not been as useful to archaeology as they were to the larger community. A simple example might be the relational database model. Though clearly not lacking in great value, this data model has conceptual limits for archaeological data. This highly structured way of organizing information has many advantages, but it is not well suited to representing complex hierarchies or relatively unstructured data. Archaeology and philology, in particular, are rife with complicated spatial, temporal, and logical hierarchies and loosely structured text. Cultural heritage researchers need a data model that conforms to their research, rather than being forced to squeeze their data into what may be an inappropriate relational mold (Schloen 2009).

To achieve the Web 2.0 objectives (and not just adopt its technologies), we need to see a number of common threads. First, it is essential to decouple traditional system design and modularize all elements so that data and applications can be repurposed. For example, a single database “back-end” can serve multiple web services with data. Modularization of processes into web services allows best-of-breed application selection and mash-ups. In this setting, developers are not stovepiped into a single structure but can rapidly use and reuse others’ work, and new development becomes faster.

So far, this sounds like any stock IT sales pitch. But I believe that there are a couple of hidden elements in this structure that have considerable bearing on whether Web 2.0 technologies become widespread in archaeological settings. One key element in the rapid and ubiquitous adoption (outside archaeology) of (first) Web 1.0 and (now) 2.0 technologies was the already extensive digital infrastructure available for these technologies to build upon. Companies had massive databases, and there were substantial bodies of technology (hardware, software, and staff) already in place. Second, there was substantial money to be made though disintermediation (that is, the elimination of the middleman), and the web technologies excelled at that. Disintermediation released huge amounts of money that could be swept up by any company choosing to adopt web technologies. When you can eliminate steps in a sales chain, or in the way management reports data to corporate leadership, you can reduce the employee count and save money. Of course, it is also the case that you can increase the speed at which information is aggregated,

providing competitive advantages. Business leaders call this the “value proposition.” Finally, there was (and is) considerable benefit to many, but not all, companies through the development of interoperability. By “interoperability” I mean the creation of standards or specifications (de jure and/or de facto) that permit data and applications to interact. Through newly possible interactions between companies or even within a single company’s previously isolated data silos, we can see the value of disintermediation. Interestingly, the development of new interoperability capabilities typically disadvantages the existing market leaders and empowers the smaller ones. It is interesting to consider if there is a parallel effect in scholarship generally and in archaeology specifically. I might mischievously suggest that it seems likely that there is. For example, one of the reasons that bright students and faculty choose to be a part of distinguished university X is the quality of its library and museum collections. If these assets were as easily accessed over the Web by faculty and students at little school Y, then at least some of the competitive advantage of distinguished university X would be diminished.

My characterization of the value of disintermediation for existing data has one weakness: there are clear examples where Web 2.0 tools were created absent an existing body of data but where the new tools now, in turn, have made it possible to create and aggregate new data. Some examples include social media (such as [Facebook](http://www.facebook.com/))<sup>1</sup> and crowd sourcing (for example, [Wikipedia](http://www.wikipedia.org/)).<sup>2</sup> This suggests that there is a second dynamic underway other than the “simple” economic disintermediation as I have characterized. Perhaps we can consider this as “social disintermediation.” In this case, the Web 2.0 tools have allowed individuals to connect in easier and more scalable ways and have allowed for the advantages of this interaction to be captured by the community. Clearly, Wikipedia removed the barriers that existed in traditional encyclopedias for creating and, much more importantly, disseminating information. And Facebook reduced the “social costs” of interaction between individuals, diminishing or eliminating the cost that face-to-face communication or even distant synchronous communication requires, while making it possible to engage many more participants with the same effort.

If my first characterization is correct, then in order for Web 2.0 to be important in archaeology we need to ask: (1) Do we have massive existing

---

<sup>1</sup> <http://www.facebook.com/>

<sup>2</sup> <http://www.wikipedia.org/>

digital assets “waiting to be free”? (2) What benefits accrue (and to whom) from disintermediation? (3) Is there a movement toward standards and specifications that empowers disintermediation? In the archaeological and general scholarly context, I think it is also critical to add a final factor—namely, (4) sustainability. By sustainability, I mean the continuing availability of data and tools into the future. Unlike businesses, where creative destruction is the norm, archaeology has a responsibility to ensure that data continues. This final requirement places particular responsibilities on Web 2.0 efforts.

Alternatively, if archaeologists are to use new Web 2.0 tools to actually create content rather than to improve accessibility to existing content, then we need to consider the nature of that content. At a very basic, perhaps cartoonish level, archaeology can be seen as a field almost uniquely focused on things (typically artifacts but any material remains), the ways in which those things were found or recovered, and, finally, the ideas, thoughts, opinions, and theories about how those things relate to one another and to the contexts (physical and ideational) in which they once existed and now do so in the present. If this simplified view is at least partially correct, then the question of content creation takes on a different guise. When we describe or characterize actual material remains and their contexts in a web setting, we are not creating content (at least in one sense) but, more accurately, we are repurposing it—or perhaps, more precisely, converting its media form. We go from a physical object to a digital representation of one. Obvious examples are digital site files, object inventories, and the like. These are digital representations of physical materials. However, we can also imagine other tools that allow the true “creation” of content—that is, the creation of relationships, ideas, opinions, or theories about these objects. Viewed from this perspective, the development of new tool kits that allow archaeologists to create either form of content provides a different structure of incentives. We can, perhaps, characterize the first as top-down, where large institutions with large data sets make them accessible—for example, a major museum or university providing easy Web-enabled access to data on sites or objects in its collection. The other is bottom-up, where individuals provide the content—for example, where users create linkages of many forms (classificatory, analytical, and the like) between these and other objects and identify previously unrecognized relationships. While at one level the two are quite different, the value of each effort is multiplied by the presence of the other. If this view is correct, then we need to encourage a simultaneous push for the adoption

of Web 2.0 strategies by large institutions as well as by individuals and should look for an architecture that does not privilege one at the expense of the other.

If the idea of social disintermediation is correct, then the question is also whether the development of new tools that allow new content creation will empower a new cycle of community development by archaeologists. Of course, it need not be an either/or situation, and both may be operating. In fact, it would seem likely that there may be interesting combinations of both. The archeological community is—by Web standards, anyway—quite small and the number of Web-active archaeologists some fraction of the whole. I will return to this possibility later, but examples that highlight the value of socially “disintermediative” tools in combination with archaeological efforts might be the integration of the [World Heritage database with Google Earth](#).<sup>3</sup>

## CONTENT IS KING

It is essential to realize that Web 1.0, Web 2.0, or any future Web iteration is ultimately built on content, and this is especially true for archaeology. In many Web 1.0 and 2.0 applications, there was already considerable digital content when the first iterations appeared, so the task then was to “release” it from its current data silo and “make it free”—though in some situations, new Web 2.0 tools also made possible the creation of new content. We need to ask, first, if archaeology has such existing digital data that can be released by web tools and applications and/or whether there are specific new tools that would encourage the development of new content. Do we have digital archaeological content yearning to be free? From my limited (primarily Americanist) perspective, the short answer is “not much,” but there are exceptions, and these need to be carefully considered to see how they may serve as a guide for the future. Nick Eiteljorg’s commentary in this volume (Chapter 10) about the lack of success of the Archaeological Data Archive is not a criticism of his efforts but reflects the state of the art. Existing Web-accessible archaeological digital data sets are few. In the current volume, we have the example of the [Archaeology Data Service](#),<sup>4</sup> operational and growing be-

---

<sup>3</sup> <http://whc.unesco.org/en/news/570>

<sup>4</sup> <http://ads.ahds.ac.uk/>

yond its first decade, as well as the **Integrated Archaeological Database (IADB)**,<sup>5</sup> also now almost a decade old (as well as other useful British initiatives from **English Heritage**<sup>6</sup> and others). There are also other examples, such as the **Digital Archaeological Archive of Comparative Slavery**,<sup>7</sup> now entering its ninth year, the **Perseus Project**,<sup>8</sup> and particularly valuable efforts around Scandinavia, to note a few of the more prominent ones.

In the United States, many states (for example, **New Mexico**,<sup>9</sup> **Arizona**,<sup>10</sup> and **Arkansas**<sup>11</sup> have site-management-oriented systems that have a long life, and the National Park Service has its **National Register**<sup>12</sup> **Automated National Catalogue System (plus)**<sup>13</sup> and other databases. While these state and national systems are undoubtedly valuable, they all have management as their primary (and sometimes only) goal. Furthermore, throughout the United States, site location information is tightly held and, as a result, general Web access to these site management systems is very restricted. Such restrictions make the use of Web 2.0 approaches largely untenable. Interestingly enough, this is not the case in Britain, where site location data is public.

So, reconsidering our first criterion, “do we currently have data yearning to be free,” the answer would clearly be no, but with some exceptions. Some of the current volume’s authors suggest that this limitation can be addressed by improving the ease by which data can be made digital and by recognizing that much data is already born digital. This is true but the value of any single digital record can be seen as a function of the total number of other possible linked or relevant data, so until some critical data mass is reached, the value of adding some data now is less than it would be if it were part of a larger effort. That is not to say that it is without any value—far from it. In the same way that any act of individual scholarship is of use to the larger community, any, even small, individual-scale programs utilizing Web 2.0 are of merit. Done properly, they are at least additive and may have an even

<sup>5</sup> <http://www.iadb.org.uk/>

<sup>6</sup> <http://www.english-heritage.org.uk/>

<sup>7</sup> <http://www.daacs.org>

<sup>8</sup> <http://www.perseus.tufts.edu>

<sup>9</sup> <http://stubbs.arms.state.nm.us/arms/>

<sup>10</sup> <http://www.statemuseum.arizona.edu/crservices/azsite/index.shtml>

<sup>11</sup> <http://www.uark.edu/campus-resources/archinfo/cspdatabases.html>

<sup>12</sup> <http://www.nps.gov/nr>

<sup>13</sup> Formerly available at <http://www.nps.gov/history/museum/publications/>



greater multiplicative effect. We can think of it, perhaps, as analogous to a transportation network. Having a railroad around your hometown and even many others in isolation is valuable, but each becomes much more valuable when linked by an intercontinental network. The great genius of the Web 2.0 environment is that by simply participating in the community and utilizing the existing technology standards, at least the possibility of this interconnect-edness is in place. To return to our railroad analogy, adopting and using stan-dards is important—it is only when all the railroads have the same track size that we can effectively link them.

### ON THE ARCHAEOLOGY WEB— WHO WINS AND WHO LOSES AND WHY?

Many Web 2.0 developers fundamentally believe that data wants to be free, that crowd sourcing is good, and that open source is better than commercial software. These precepts may all be true, but it is essential to understand that hardware costs money, which programmers need in order to eat (typically Jolt Cola and Ding Dongs—but still) and pay mortgages, and that “there is no such thing as a free lunch.” Something may be free but that does not mean it has no cost or, conversely, that there is no measureable benefit to the “free” contributors. Mozilla may be free, but its existence reflects basic business de-cisions of people who wanted a balance to Microsoft’s Internet Explorer. None of this is to diminish the truly great societal value of FOSS (free and open software), but it is important to understand who wins and (perhaps) who loses. But first let us look at the real and enduring values of FOSS. A ma-jor positive value is transparency. If you have access to the underlying code, you can verify yourself what is happening. It can also provide increased flex-ibility, allowing you to make the changes you need. There are many other benefits, which are beyond the scope (or the author’s capabilities) to consider for the Internet writ large (cf. Brabham 2008 for an interesting assessment). However, we can consider the values of FOSS for archaeology.

One very important issue is that archaeology can clearly piggyback on many of the massive efforts made by others in Web 2.0. We need not create a Google Maps-like system or design a replacement RSS for archaeology. Archaeologists simply need to adopt and adapt these existing Web 2.0 struc-tures and tools to our needs. When we do this, our own development costs are greatly lessened. That said, the standard commercial balance sheet does

not really apply to archaeology specifically or to research generally. The archaeological enterprise (and research generally) is seen as a societal good and thus is usually subsidized. Within this context, the creation and adoption of Web 2.0 becomes a viable and sustainable effort if the overall value (measured in intangibles such as improved research, public involvement, and understanding of the past) outweighs the initial (usually grant-subsidized) costs plus the continuing costs for operation, data entry, and the like. Costs here need not be just money but include the time involved and diversion from other tasks. Every article written about Web 2.0 in archaeology may be viewed, in part, as one *not* written about the archaeological record itself. Furthermore, there are costs involved in an archaeologist's learning Web 2.0 methods—this is called an “opportunity cost”—as they are not simultaneously doing “real” archaeology. From a pure cost-benefit perspective, only if the archaeologist who “spent” time learning Web 2.0 and “spent” time creating new tools also created powerful or faster ways to do new analyses (or to do old ones faster) could we say that Web 2.0 was valuable.

The papers by Kansa and Kansa, Wendrich, Eiteljorg (Chapters 2, 8, and 10, respectively), and others have emphasized that the canons of scholarly life (generally, and archaeological research life, specifically) do not appear to provide strong motivation for Web 2.0-type efforts, at least not currently. There are powerful forces in archaeology (and many disciplines) that are very contrary to data wanting to be free (in the many meanings of that word). If we are to achieve the value that interoperable, scalable access to data can yield, it is critical that these canons of the scholar change. To do this, we must alter the current cost-benefit equation. That is, we need to lower the cost of entry through free, easy-to-use Web 2.0 tools, and we need to increase the “benefit” of data publication and new tool development to those seeking promotion, tenure, and related personnel advancements in museums, universities, consultancies, and government service. There are some technical steps that can take us a long way in this direction; the tools discussed by Kansa and Kansa are particularly relevant, and their example should be widely duplicated. The adoption of persistent URLs that can be cited and tools that make citations accessible are essential, such as those developed for **Open Context**.<sup>14</sup> Citation counts for data as well as for publications need to be de-

---

<sup>14</sup> <http://opencontext.org>

veloped and used in evaluating scholarship. If we imagine a setting where the number of citations to a data set published on the Web would be as important a factor in promotion and tenure as the number of citations to an article, we can quickly see how the community good is aligned with the individual scholar. There are many examples of strong movements in this direction in many fields (e.g., Schwartz et al. 2010). If we consider the consulting community, which is probably the main source of essential data, we also need to consider how we increase the value to these companies and agencies while lowering the cost. At a minimum, we can see that if data are accessible online and easily accessed, this will lead to reducing the cost and time to access the data, which clearly has real implications in reducing budgets or increasing the speed at which decisions can be made.

We need to assess academic and research institutions not only on the aggregate of their scholarship and museums on the breadth of their collections, but on the accessibility and comprehensiveness of their data archives. Just as we judge a great university in part on the number of volumes in its library, we should now also judge it on the breadth and number of archaeological records that are exposed to search and analysis.

## SUSTAINABILITY

While looking at who wins and who loses, we must also consider sustainability. In 1981, Sylvia Gaines edited the landmark volume *Data Bank Applications in Archaeology*, which included a number of articles covering a range of digital databases and systems used in research and site management. Now, 30 years later, few of them still exist. To the best of my knowledge, of the research-oriented systems described in the book in 1981, only the Arkansas Archaeological Survey's **AMASDA system**<sup>15</sup> continues as a robust and operational system that has migrated its data repeatedly from earlier software architectures and platforms to its current and greatly expanded one—all the while maintaining data integrity.

More critically, of all the systems listed, including those in this current book, other than AMASDA and the state and national site management systems, none of the data systems described have, as yet, outlived their creators;

---

<sup>15</sup> <http://www.uark.edu/campus-resources/archinfo/cspdatabases.html>

this applies to even the most robust to date: ADS. Archaeological data systems, at least ones that go beyond site management functions, have a limited record of sustainability. This problem is not restricted to archaeology. The National Science Foundation (NSF) has recognized that the great majority of the cyberinfrastructure projects it has supported have not been sustainable. In fact, the recent DataNet initiative has as one of its key objectives to “provide reliable digital preservation, access, integration, and analysis capabilities for science and/or engineering data over a decades-long timeline” (NSF 2009). Another key supporter of a sustainable digital infrastructure (and one applicable to archaeology) is the Mellon Foundation’s Scholarly Communication’s Program. All archaeologists are already indebted to them for the **JSTOR system**,<sup>16</sup> and many of the examples highlighted earlier in this book have been Mellon-funded (for example, **DAACS**<sup>17</sup> and **UEE**<sup>18</sup>). As they note:

Studies by the Association of Research Libraries and others suggest that a significant percentage of the primary source materials that cultural institutions have painstakingly collected to fuel humanistic scholarship remains uncataloged and effectively “hidden” from scholars (Mellon 2008:23).

As we look at the projects discussed in this book, the question of sustainability is central. This is not to criticize the developers, far from it, but simply to point to the challenges. With what appears to me to be no exceptions (though perhaps ADS and UEE are examples, to some degree), all the projects described here are not institutionally driven but are scholar/researcher driven. The question, then, is what happens when the individuals behind the process leave, die, or become disinterested? The Web 2.0 tools are only relevant if they can continue to work from a digital database. Of course, the same challenge can be put to any archaeologist who conducts any fieldwork. How are the physical materials and records being “sustained” (that is, preserved)? One conclusion may be that we must align institutional objectives with the sustainability of these digital systems/services. One idea that is growing in interest is the role to be played by university library systems in hosting and continuing digital systems created by their faculty.

---

<sup>16</sup> <http://www.jstor.org>

<sup>17</sup> <http://www.daacs.org/>

<sup>18</sup> <http://uee.ucla.edu/>

As Mellon and NSF support initiatives to make digital information more accessible, they are requiring that these initiatives develop a sustainability plan before initial funding is granted. In the case of JSTOR, funds to continue and expand its operation come from universities and other libraries that pay annual subscription fees. In such cases, Mellon serves as a venture capitalist, providing initial support to create and populate a system, after which the system then must sustain itself. A recent Mellon-supported archaeological initiative to create an ADS-like system for North American digital data, *Digital Antiquity*,<sup>19</sup> has sustainability as a central objective. In the recent NSF DataNet calls for proposals, the successful proposals will be evaluated, at least in part, by the viability of their plans for sustainability. Again, it is important to recognize that the entire enterprise of archaeology is one that has as a core premise that the preservation and recovery of the record of the past has intrinsic societal value. In a similar manner, a Web 2.0 archaeological effort should also be structured and focused on this intrinsic criterion.

We can imagine that there may be alternatives to the fee-for-service or subscription approach. There can be “community good”-based models. Museums, for example, create and maintain physical archives without the need to rent them out, but we are all aware of the tremendous fiscal challenges even the strongest museums are facing.

## INTEROPERABILITY

For a Web 2.0 strategy to be successful, not only must there be a critical mass of base digital data, but the data needs to be interoperable. Interoperability among and between digital data is a key “enabler” of Web 1.0 and Web 2.0 technologies. In fact, interoperability may be *the* essential element underlying the success of the Web. There is a power law at work here. The value of any single data element grows along with our ability to relate it to other elements.

Because interoperability in “simple” things like web addresses, character sets, and image file formats has already been solved and is ubiquitous (except in moving videos from a Windows operating system to a Mac), we often fail to appreciate the complex processes and considerable efforts that have gone into their development and acceptance. It is my opinion that recognition and

---

<sup>19</sup> <http://www.digitalantiquity.org>

adequate control of ontologies and semantics have been fundamental to all these successful efforts. This does not mean that *all* terms and relationships must be defined and structured before work can begin, but it is critical that fundamental concepts and terms be set early on and that a dynamic process (and tools) be in place to encourage and extend these rapidly through time.

I have had the (dubious) pleasure to be associated with the development of interoperability for one such area, geospatial data and processes, as part of the **Open Geospatial Consortium (OGC)**<sup>20</sup> and, more recently, the **Open Geospatial Interoperable Institute**.<sup>21</sup> The common thread in my experience was the extreme challenge in coming to consensus on the very basic terms of practice. In the OGC setting, it took literally years to develop a common semantic basis around such fundamental terms as point, line, and polygon. That said, once the very fundamental terms had been defined and structured, further ontological/semantic efforts proceeded relatively easily and quickly. In the OGC example, once semantic agreement was accomplished for the very basic terms, a rapid test-bed structure was created where, in a six- to eighteen-month period, a community of interested parties developed both the semantic and technology infrastructure to address a well-defined problem. I believe that the initial hard ontological/semantic work has not yet been done for archaeology. In fact, it could be argued that the reward structure in archaeological scholarship provides a powerful *disincentive* for participation in the development of semantic interoperability and, instead, privileges the individual to develop and defend individual terms/structures and categories. This is particularly troublesome with respect to ontology and semantic interoperability because of the strong linkage between increased prestige and fighting for a specific ontological position. Look at the historical trajectory of these issues in the field. We have only to review the epic battles over the Midwestern Taxonomic System or other naming schemes to see that a model of conflict, rather than consensus, is historically embedded. It is also clear that there is perceived value in the naming of archaeological entities and in creating new, distinct ones.

How do we address this issue? It seems to me that one important step is to separate more forcefully data (observations) from the higher-level abstractions derived from them. In archaeology, much of what passes for data is in-

---

<sup>20</sup> <http://www.opengeospatial.org>

<sup>21</sup> <http://www.ogcii.org>

stead an *n*th-order abstraction, approaching information but not there yet either. We can quickly illustrate, using the simple example of almost any common artifact term—say, “celt” (but you can pick almost any one). The term itself simultaneously embeds ideas of (a) shape, (b) use, (c) material, and frequently (d) time and (e) place, as well as imputed/inferred parameters of social role, trade, and on and on. We must unroll this complex web of meaning into its constituent parts if true interoperability is to be realized.

If we can separate the formal characterization of the dimensions of an entity (such as size, shape, material, place, and time) from its imputed archaeological meaning (trade, social status, cultural affiliation), then logically we can provide different incentives for the dissemination of each. The multiple objectives of archaeological systematics, ontologies, semantics, classification, and taxonomy have been commingled from the earliest days of the field, and the literature is replete with discussion over the proper procedures and objectives, but it has always been clear that one of the first roles of an archaeologist is to “properly” classify materials, which means, effectively, processing a multidimensional matrix of variables into the proper single classificatory category. Thus a sherd with specific temper, thickness, surface treatment, and surface decoration is a sherd of type “X.” But once classified as X, other archaeologists impute meanings. This traditional approach also requires that individual properties that may be more or less continuous in their distribution are necessarily made discrete to allow categorization, reducing our ability to see variability. If formal characterization is separated from imputed meaning, as two separate but related properties each with its own semantics and ontologies, then there is clear value in publishing the basic “data” as well as the second- (and *n*th-) order abstractions. In a paper-based linear (pre-Web 2.0) discourse, we were (and are) obligated to conflate multiple dimensions of objects, sites, or other things simply to make any discussion possible. In doing so, we do speed up communication but we also lay the foundation for confusion, conflict, and endless arguments. With Web 2.0 tools, we can isolate these areas both conceptually and operationally. Separation of these independent properties as data from their (presumed) meaning would serve to provide a critical role for the Web 2.0 tools described in this book.

## FINAL COMMENTS

A final general thought: Particularly in information technology, “meaning well” and “doing well” are two quite different things. It is clear that the authors in this text all mean well. In order for their vision to be accomplished, however, we will need fundamental institutional and sociological changes to create a setting where the promise of the technological tools can actually be realized. It is because of the promise that is clearly evidenced in these contributions that, hopefully, our field can be motivated to do the hard work necessary to fulfill this promise—and to do well.

## REFERENCES CITED

- Brabham, D. C.  
 2008 Crowdsourcing as a Model for Problem Solving: An Introduction and Cases. *Convergence* 14/1 (February 1): 75–90.
- Katz, M., and C. Shapiro  
 1986 Technology Adoption in the Presence of Network Externalities. *The Journal of Political Economy* 94/4 (August): 822–841.
- Gaines, S. (ed.)  
 1981 *Data Bank Applications in Archaeology*. Tucson: University of Arizona Press.
- Mellon Foundation  
 2008 2008 Annual Report. Retrieved from [http://www.mellon.org/news\\_publications/annual-reports-essays/annual-reports/content2008.pdf/view](http://www.mellon.org/news_publications/annual-reports-essays/annual-reports/content2008.pdf/view) (accessed August 28, 2009).
- NSF  
 2009 DataNet Solicitation Guidelines. Retrieved from [http://www.nsf.gov/funding/pgm\\_summ.jsp?pims\\_id=503141](http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=503141) (accessed August 28, 2009).
- O'Reilly, Tim  
 2006 Web 2.0 Compact Definition: Trying Again. Retrieved from <http://radar.oreilly.com/2006/12/web-20-compact-definition-tryi.html> (accessed May 10, 2010).
- Schloen, D.  
 2009 System Design. Retrieved from [http://ochre.lib.uchicago.edu/index\\_files/Page632.htm](http://ochre.lib.uchicago.edu/index_files/Page632.htm) (accessed December 7, 2009).
- Schwartz, A., C. Pappas, and L. Sandlow  
 2010 Data Repositories for Medical Education Research: Issues and Recommendations. *Academic Medicine* 85/5: 837–843.