Publishing Primary Data on the World Wide Web: Opencontext.org and an Open Future for the Past

Eric C. Kansa

The Internet and Scholarly Communication

More scholars are exploring forms of digital dissemination, including open access (OA) systems where content is made available free of charge. These include peer-reviewed e-journals as well as traditional journals that have an online presence. Besides SHA's *Technical Briefs in Historical Archaeology,* the *American Journal of Archaeology* now offers open access to downloadable articles from their printed issues. Similarly, *Evolutionary Anthropology* offers many full-text articles free for download. More archaeologists are also taking advantage of easy Web publication to post copies of their publications on personal websites. Roughly 15% of all scholars participate in such "self-archiving." To encourage this practice, Science Commons (2006) and the Scholarly Publishing and Academic Resources Coalition (SPARC) recently launched the Scholar Copyright Project, an initiative that will develop standard "Author Addenda"—a suite of short amendments to attach to copyright agreements from publishers <*http://sciencecommons.org/projects/publishing/index.html*>. These addenda make it easier for paper authors to retain and clarify their rights to self-archive their papers electronically. Several studies now clearly document that self-archiving and OA publication enhances uptake and citation rates (Hajjem et al. 2005). Researchers enhance their reputations and stature by opening up their scholarship.

Mounting pressure for greater public access also comes from many research stakeholders. Granting foundations interested in maximizing the return on their investment in basic research are often encouraging and sometimes even requiring some form of OA electronic dissemination. Interest in maximizing public access to publicly financed research is catching on in Congress. A new bipartisan bill, the Federal Research Public Access Act, would require OA for drafts of papers that pass peer review and result from federally funded research (U.S. Congress 2006). The bill would create government-funded digital repositories that would host and maintain these draft papers. University libraries are some of the most vocal advocates for OA research. Current publishing frameworks have seen dramatically escalated costs, sometimes four times higher than the general rate of inflation (Create Change 2003). Increasing costs have forced many libraries to cancel subscriptions and thereby hurt access and scholarship (Association for College and Research Libraries 2003; Suber 2004).

Beyond Papers: Sharing Other Scholarly Media

These policy developments and others show that OA models are gaining in importance. Similar trends toward greater openness are evident for other types of scholarly content, beyond peer-reviewed papers. Besides making distribution highly cost-effective, the Internet provides a powerful means for sharing large collections of rich media and complex data. These types of content are important components of both museum collections and excavation documentation. More efficient and comprehensive sharing of this complex, media rich content is an important goal for many seeking to reform and enhance scholarly communication (Association of Research Libraries 2006:57–59). As demonstrated in ecology and other sciences, reused primary data can be an important resource for advances in understanding (Kansa 2005).

Data sharing does present a new set of technical, conceptual, and incentive problems. Excavation results, specialist analyses, and museum collection databases are highly variable and often complex (Kintigh 2006). Nevertheless, there has been great progress on many of the technical and conceptual problems involved in pooling and integrating the complex and unstandardized data generated by researchers. The Etana Digital Library project (Etana-DL) <*http://feathers.dlib.vt.edu:8080/etana/servlet/Start*>, led by James Flanagan and digital library pioneer Edward Fox, has successfully demonstrated a data-mediation system that uses software to translate local data structures to a more

general data structure. This mediation enables Etana-DL to provide interoperability and integrated search, browse, and analysis tools for several Near Eastern excavation datasets (Flanagan et al. 2004). Dean Snow and colleagues advocate developing advanced text-mining systems to extract comparative data from archaeological reports, including "grey literature" documentation generated from CRM activities (Snow et al. 2006). Following the model of other scientific disciplines, the National Science Foundation recently awarded a group led by Keith Kintigh and colleagues a large "cyber-infrastructure" grant to stimulate data integration and sharing in archaeology. This project is now in its initial stages and aims to begin by developing ontologies for zooarchaeology. Ontologies are formally defined conceptual systems and are often used to support the integration of multiple datasets within a discipline.

An "Open Context" for Excavation Results and Related Collections

Other working systems are now coming online, including two related systems, the University of Chicago OCHRE project <*http://ochre.lib.uchicago.edu/*> and Open Context <*http://www.opencontext.org/*>. Both systems share the same data architecture described by the Archaeological Markup Language (ArchaeoML) and both have similar capabilities for integrating and pooling complex and media-rich archaeological documentation (Schloen 2001; Kansa 2005). While OCHRE provides sophisticated data management tools targeted for active research projects, Open Context (Figure 1) is aimed at streamlined, Web-based access and retrieval of excavation and collections-related content. Funding for the development of Open Context came from the William and Flora Hewlett Foundation, as part of their Education Program's effort to make high-quality instructional resources freely available on the Web.

Open Context enables researchers to publish their primary field data, notes, and media (images, maps, drawings, videos) on the World Wide Web. It provides an easy to use, yet powerful, online database for exploring, searching, and analyzing multiple excavation results, survey datasets, and museum collections. These diverse datasets can be explored by browsing through a map (Figure 1) or through different search options. Open Context is built with standard but powerful Web technologies (MySQL and PHP), making it easy to integrate with a host of other Web services, including weblogs, e-journals, and commercial search engines. Search engine discovery is becoming an increasingly significant factor in determining the impact and uptake of research (Jensen 2005; Vaughan and Shaw 2005).

Types of Material in Open Context

Open Context is best suited for publishing large bodies of complex archaeological documentation. All content is linked together in an integrated and cohesive resource. The types of content in Open Context include:

- Narratives: These include more loosely structured, textual types of content, including excavation notes, observations, and diaries. These narratives are integrated and linked to other types of information, including database records and other media (images, videos, and maps) (Figure 2).
- Analytic (Tabular) Data: Open Context enables publication of database types of content. These include context databases, finds registries, museum registries and catalogs, and specialist analyses. All of these different types of data are automatically integrated together in one cohesive database (Figure 3).
- Media: Open Context can link digital images, maps, drawings, GIS files, videos, and other types of media with other forms of content. For example, a user will immediately know if an item in a finds registry was photographed or drawn because a thumbnail image will appear with the record of that item (Figure 4).

Using Open Context

Open Context makes it easy to browse, search, and analyze data from different projects and collections. It can serve as a reference resource to help researchers find relevant comparative materials. It can also support reanalysis and reinterpretation of excavation results. Finally, undergraduate students can use Open Context as a primary source, so they can develop important analytical skills by exploring and synthesizing primary excavation results. Users have a variety of options to find materials in Open Context, including simple, "Google-like" text searches (Figures 5 and 6), and more sophisticated, advanced searches that use Boolean logic (Figure 7).

Figure 1. A simple map enables users to locate content based on geographic location.

Access, Copyright, and Reuse of Content

Open Context is an OA publication system. All content is freely available on the World Wide Web. Contributors who publish with Open Context retain copyright to their content. This means all contributors are free to publish their material with other venues (including journals, books, and other websites). Open Context is also unique in that it provides a framework for sharing archaeological research, free of burdensome copyright restrictions, while still protecting scholarly attribution. Copyright law prohibits unauthorized reproduction or derivative uses of all expressive works. These legal barriers work against some of the innate advantages of digital content, namely the ease to which digital information can be readapted in new works and then globally shared. In order to encourage scholarly reuse and reapplication of copyrighted digital content, each item in Open Context is licensed with an open, Creative Commons license. These licenses give explicit permissions for users to freely and legally use the material so long as they properly attribute the original creator (Brown 2003) (Figure 8). Creative Commons licenses include machine-readable RDF metadata that is captured by commercial search engines such as Yahoo and Google (Kansa et al. 2005). This metadata facilitates discovery of openly licensed content, including Open Context resources. Such openness ensures that the Open Context content is of maximum value for reuse in both instructional and research applications. Finally, to facilitate scholarly applications, citation information is automatically generated for each item in the database (Figure 8). Stable URLs to each item in Open Context facilitate citation and later retrieval.

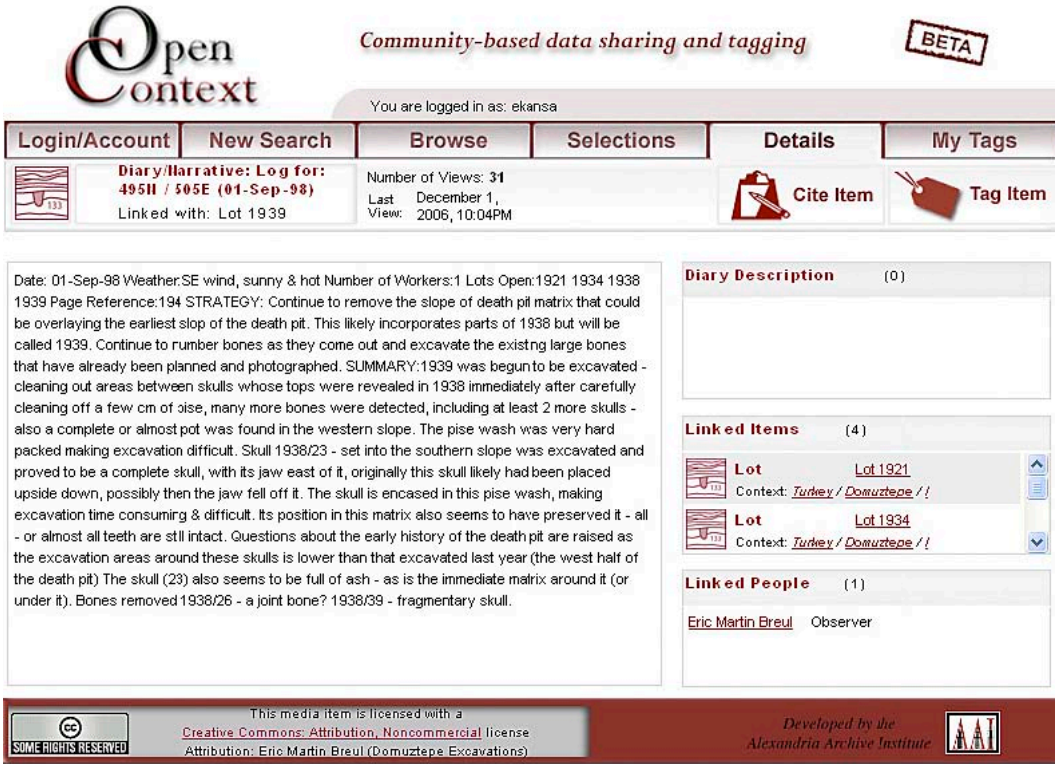

Figure 2. An excavation diary linking to items from the Domuztepe Context Database.

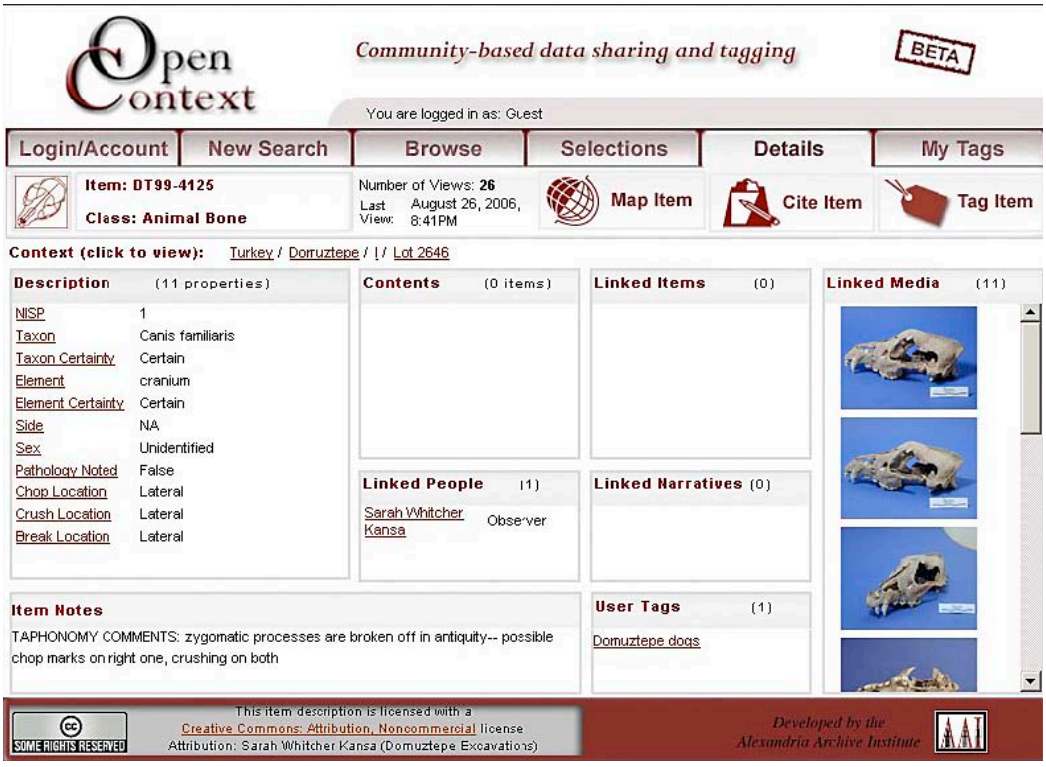

Figure 3. A record from the Domuztepe Zooarchaeological Analysis Database.

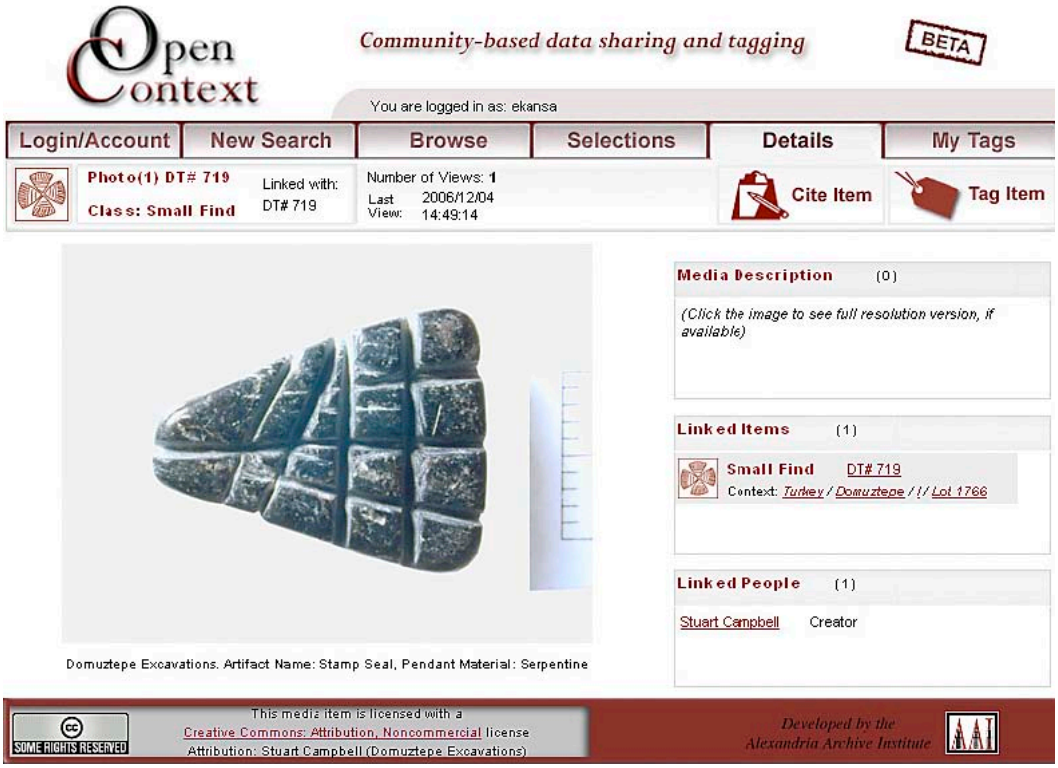

Figure 4. An image linked with its small finds registry record and context.

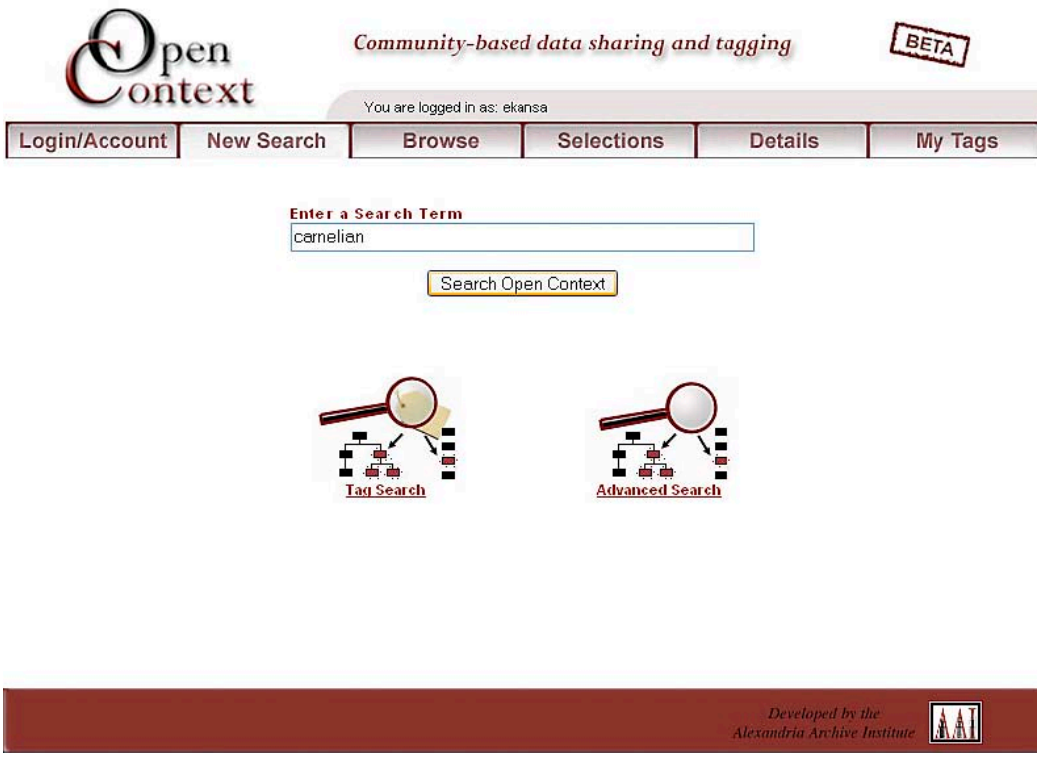

Figure 5. A simple search for "carnelian."

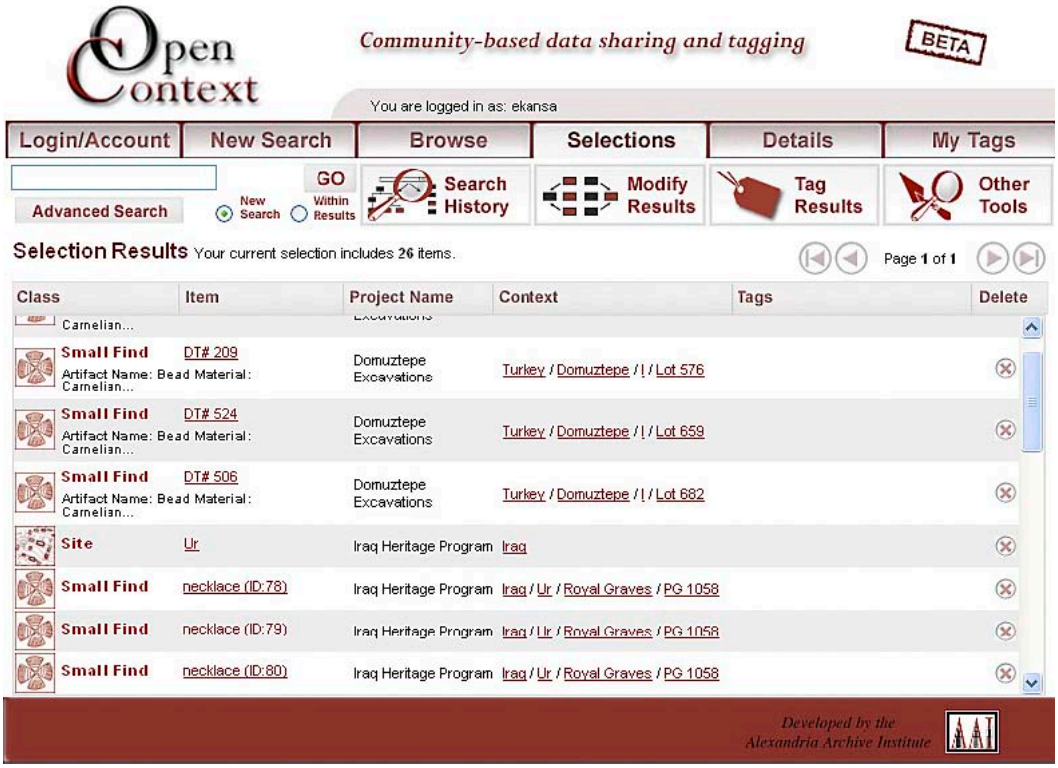

Figure 6. Results of the above "carnelian" search, showing items from multiple projects.

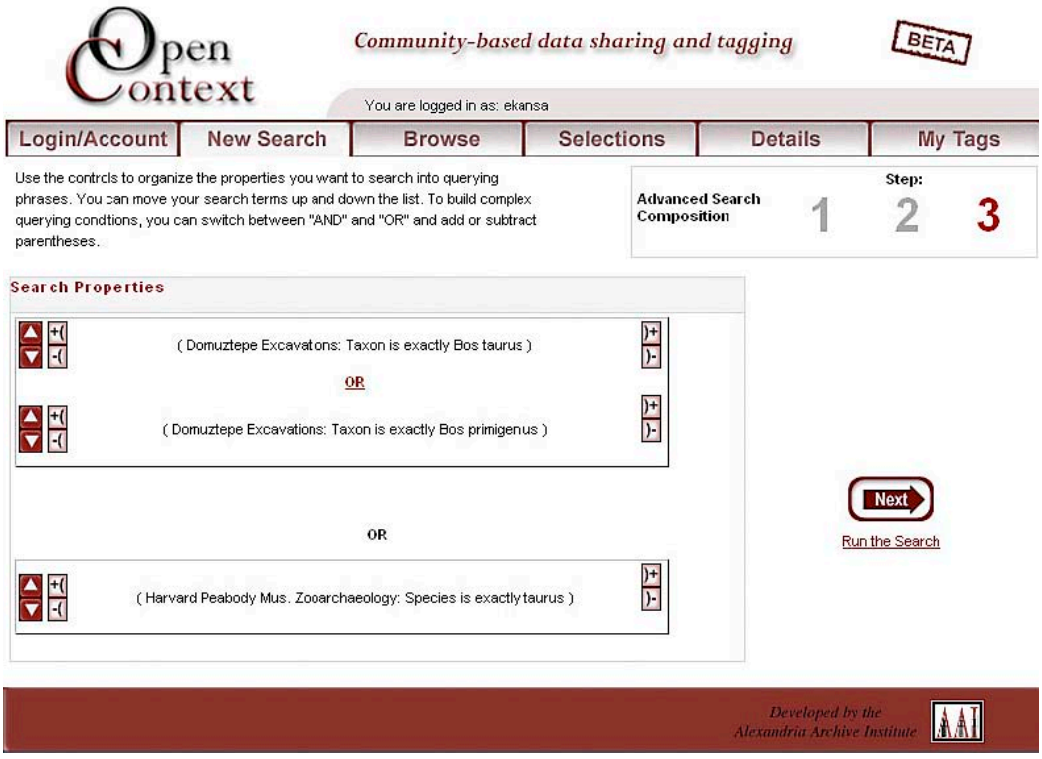

Figure 7. Example of an "Advanced Search" across multiple projects.

Making Sense of Multiple Projects and Collections

The lack of many formal standards in archaeology makes data integration a challenge. Open Context uses special software to import spreadsheets and databases created by individual researchers and collections managers. We are currently completing development of a a web-based version of this import software. The web version will enable individual contributors to upload and document their own data tables and submit them for editorial review. Because Open Context uses ArchaeoML's very flexible architecture, all the original recording systems and terminologies and retained. Open Context enables researchers to publish their data *without forcing them to conform to overly restrictive, predetermined standards*. To help make sense of this widely varying body of material, Open Context has a user "folksonomy" system. Open Context enables users to tag items either individually or collectively (users can assign a tag to items in a query result set). When query result sets are tagged, the history of query composition is automatically linked to the tagging event (Figure 9). Users can also further annotate and explain the rationale behind their tag assignments. Tags can be used to save search selections for future reference and to share sets of items with colleagues. The authorship of each tagging event is documented by Open Context. Users can filter out tags and tag authors they consider unreliable. Currently, Open Context developers, a team led by two archaeologists with doctoral degrees, monitor and moderate tagging activity to insure quality and relevance. As content and tagging activity expands, subject matter experts will be needed to provide editorial review of the folksonomy system. Reviewers will receive email notification of tagging activity and can remove inappropriate tags from public view.

Folksonomies are cost-effective and simple tools that enable a community of users to add value to pooled content by identifying and annotating items of interest. Users can "tag" items with common keywords and phrases and thereby establish and share meaningful links among items from different projects and collections, even if these projects use different recording systems. The folksonomy system can facilitate semantic data integration, and recent experiments suggest such systems offer annotations of sufficient quality to meet some needs of museum professionals (Bearman and Trant 2005; Trant 2006). Open Context is developing several enhancements to this system, including better ways of recognizing professional credentials and scholarly authority, and options for users to apply professionally developed standard vocabularies such as the Getty Art and Architecture Thesaurus or future ontologies such as those advocated by Kintigh's team (Kintigh 2006).

"Pinging" the Past

Folksonomies, once they are properly adapted for professional applications, and other collaborative methods of data analysis can become new areas where researchers can make important scholarly contributions. "Publishing" primary datasets does raise some interesting questions about value and credit. An excavation dataset is not like a peer-reviewed paper, currently the main currency of professional achievement. To begin assessing the scholarly impact of digital datasets, Open Context records information on visits to each record in the system. Recent studies have shown a significant correlation between download counts and more commonly used measures of citation impact in scholarly papers (Brody et al. 2006). Nevertheless, a database in itself is a poor guide to understanding excavation or survey results. In order to be better understood and used, datasets are best linked with papers and narratives that synthesize observations and interpretations in a more meaningful framework (Richards 2003). This will necessitate technical and editorial coordination between archaeological journals and online data publishers.

To facilitate coordination with narratives, Open Context automatically generates reciprocal hyperlinks with other Web services that support the open "ping-back" standard. If a person using a weblog or publishing in a ping-back-enabled e-journal references an item or a set of items in Open Context, the Open Context system will be automatically informed about what items are being referenced. Open Context will then display links back to the weblog post or e-journal article referencing the Open Context database. Once an editorial board is assembled, all links will be subjected to editorial review. This will ensure that Open Context only registers references to trusted sources (such as an e-journal with a peer-review process). As e-journal systems gain popularity, such features will help ensure that Open Context users will easily find scholarly uses and in-

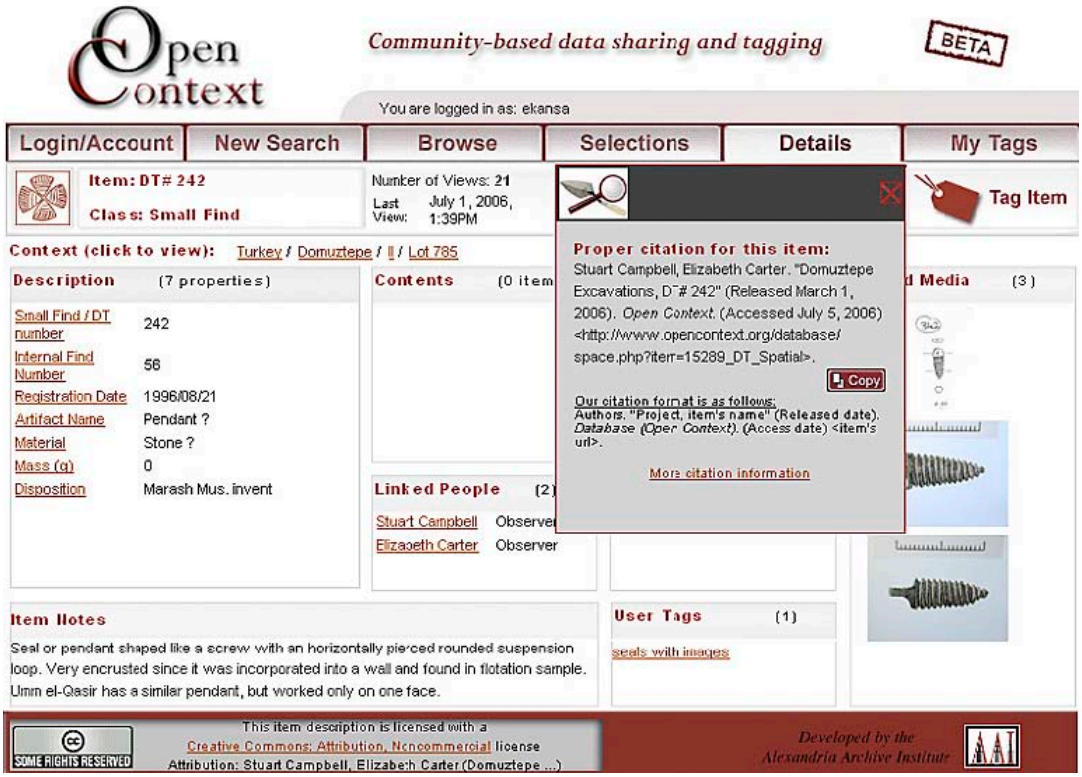

Figure 8. An automatically generated citation for an Open Context item.

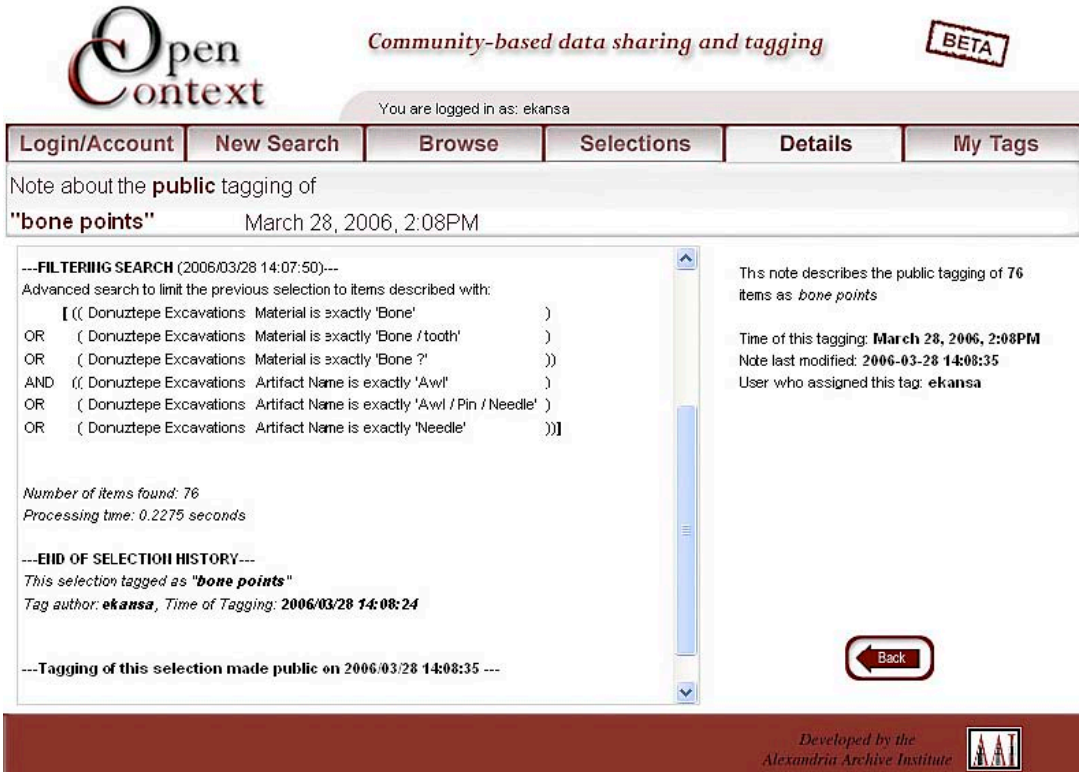

Figure 9. A query history saved with a set of tagged items.

terpretations of Open Context content. Similarly, the ability to reference public databases such as Open Context can enhance e-journal publications by making primary evidence more transparent and open for critical evaluation.

Looking Forward

Open Context has a variety of demonstration datasets now available for exploration and testing. These include field archaeology contextual records and finds registers, geoarchaeological samples, and a variety of zooarchaeological analyses. The system is also adding museum and reference collection datasets. Some projects have rich image collections and narrative material, and others are of primary interest for specialist comparative analyses.

The primary goal now is to build a critical mass of users, contributors, and content needed to sustain Open Context as a valued scholarly resource. Because Open Context supports dissemination of highly structured database content along with textual narratives, such as historical documents, and other media, such as photos or drawings, it can be a useful communication tool for the historical archaeology community. Databases of excavation records, finds registries, and comparative collections can be pooled together with textual syntheses stored in the form of HTML or PDF documents. The flexible data structure can also enable historical archaeologists to relate archaeological datasets with other types of structured information, such as historical census reports, tax records, cemetery records, or even shipping manifests. Relating such varied forms of documentation may open doors for innovative research agendas in historical archaeology. To help demonstrate and refine the applicability of Open Context to support this subdiscipline, its developers invite contributions of excavation and survey data, media, and museum collections from historical archaeologists.

In addition to contributing content, members of the historical archaeological community are also welcome to participate in other capacities. Open Context needs editorial assistance to help insure quality and oversee revision and error correction. Additional support may be provided through open-source software development partnerships. Finally, data sharing and collaboration among multiple repositories is an important digital data longevity strategy (Reich and Rosenthal 2001). Currently, the OCHRE project provides digital longevity support for Open Context pilot projects, and additional archival partnerships will be needed to better secure the often-irreplaceable archaeological content hosted by Open Context. To learn more about contributing to Open Context, please contact the author via email (ekansa@alexandriaarchive.org).

With sufficient community contributions, feedback, and support, Open Context and related OA systems will expedite and streamline reference searches and provide a comparative format for efficiently interpreting and reanalyzing excavation results. Similarly, making primary data efficiently accessible and usable can support research agendas that are not currently achievable. By pooling primary data resources in systems with powerful analytic tools, such systems should enable broad regional syntheses that are more comprehensive and more analytically rigorous than are currently feasible (Kansa 2005; Kintigh 2006).

Contributing to the development and use of systems like Open Context illustrates one way the archaeological community can participate in the broader shifts toward OA. Across the board, OA now has a great deal of momentum and powerful institutional support. As of mid-September 2006, the Alliance for Taxpayer Access (2006) reported that 53 college presidents and 25 university provosts have signed letters in support of OA legislation. This lobbying demonstrates significant institutional backing for the reform of scholarly communication. At the same time, OA journals published by the Public Library of Science (2006) achieved impact factors rivaling *Nature* and *Science*. Other published studies document how OA makes research easier to find and use and allows that research to have greater impact and significance (Harnad and Brody 2004; Hajjem et al. 2005). Data hording, access barriers, and overly restrictive copyright controls increasingly self-marginalize both research and researchers (Willinsky 2006:21). These broader transitions in academic communication, and more discipline specific initiatives such as the development of Open Context, all illustrate how the future of the archaeological past is looking increasingly open.

ACKNOWLEDGMENTS

The author would like to thank everyone who has participated in making Open Context operational, especially David Schloen and the University of Chicago OCHRE

project for their continued support and partnership. Their efforts have provided the essential conceptual groundwork needed for Open Context development. Thanks must also be given for the collaboration of John Wilbanks (Science Commons) and Heather Ford (iCommons) for facilitating Open Context and other open science and open culture initiatives. Sarah Whitcher Kansa deserves special thanks for her comments, ideas, and edits, as do the SHA reviewers for their role in enhancing this text. Any omissions or errors in this work are solely the fault and responsibility of the author. Finally, Open Context and many other open education initiatives directly result from the enabling financial support of the William and Flora Hewlett Foundation. Its support as well as the generous financial contributions of Doris and Donald Fisher and the Joukowsky Family Foundation help ensure that Open Context can serve as a free and open access resource for the community.

REFERENCES

Alliance for Tax Payer Access
2006 Fifty-three College Presidents Lead Newest Show of Support for Public Access Act. Press release, The Alliance for Tax Payer Access. <http://www.taxpayeraccess.org/media/Advisory06-0906.html> 6 Sept 2006.

Association of College and Research Libraries
2003 Journal Pricing. Scholarly Communication Toolkit. American Library Association, Chicago, IL <http://www.ala.org/ala/acrl/acrlissues/scholarlycomm/scholarlycommunicationtoolkit/administrators1/admineconomics.htm> 10 Dec 2006.

Association of Research Libraries
2006 To Stand the Test of Time: Long-Term Stewardship of Digital Data Sets in Science and Engineering. Report to the National Science Foundation from the ARL Workshop on New Collaborative Relationships: The Role of Academic Libraries in the Digital Data Universe, September 26–27, 2006, Arlington, VA <http://www.arl.org/bm~doc/digdatarpt.pdf> 6 Jan 2007.

Bearman, David, and Jennifer Trant
2005 Social Terminology Enhancement through Vernacular Engagement: Exploring Collaborative Annotation to Encourage Interaction with Museum Collections. *D-Lib Magazine* 11(9) <http://www.dlib.org/dlib/september05/bearman/09bearman.html> 10 Dec 2006.

Brody, Tim, Stevan Harnad, and Les 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.

Brown, Glenn Otis
2003 Out of the Way: How the Next Copyright Revolution Can Help the Next Scientific Revolution. *PLoS Biology* 1(1):e9 <http://biology.plosjournals.org/perlserv?request=get-document&doi=10.1371/journal.pbio.0000009> 10 Dec 2006.

Create Change
2003 Scholars under Siege: The Scholarly Communication Crisis. Overview of the Issues, Create Change (Association of Research Libraries, Association of College and Research Libraries, and the Scholarly Publishing and Academic Resources Coalition) <http://www.createchange.org/archive/faculty/issues/silent.html> 10 Dec 2006.

Flanagan, James W., Edward A. Fox, Douglas R. Clark, Weiguo Fan, Rao Shen Ravindranathan, and Marcos André Gonçalves
2004 Jordanian Archaeology and ETANA: Developing a Digital Library for Near Eastern Archaeology. Paper presented at History and Archaeology of Jordan Conference, Petra, Jordan. <http://feathers.dlib.vt.edu/~etana/Publications/ETANAArticleSHAJ.pdf> 10 Dec 2006.

Harnad, Stevan, and Tim Brody
2004 Comparing the Impact of Open Access (OA) vs. Non-OA Articles in the Same Journals. *D-Lib Magazine* 10(6). <http://dlib.org/dlib/june04/harnad/06harnad.html> 10 Dec 2006.

Hajjem, Chawki, Stevan Harnad, and Yves Gingras
2005 Ten-Year Cross-Disciplinary Comparison of the Growth of Open Access and How It Increases Research Citation Impact. *IEEE Data Engineering Bulletin* 28(4):39–47. <http://eprints.ecs.soton.ac.uk/11688/> 10 Dec 2006.

Jensen, Michael
2005 Presses Have Little to Fear from Google. *The Chronicle for Higher Education* 51(44):B16.

Kansa, Eric
2005 A Community Approach to Data Integration: Authorship and Building Meaningful Links across Diverse Archaeological Data Sets. *Geosphere* 1(2):97–109.

Kansa, Eric C., Jason Schultz, and Ahrash Bissell
2005 Protecting Traditional Knowledge and Expanding Access to Scientific Data. *International Journal of Cultural Property* 12(3):97–109.

Kintigh, Keith
2006 The Promise and Challenge of Archaeological Data Integration. *American Antiquity* 71(3):567–578.

Public Library of Science (PLoS)
2006 About *PLoS Biology*. Journal Information. Public Library of Science. <http://journals.plos.org/plosbiology/information.php> 10 Dec 2006.

Richards, Julian
2003 Online Archives. *Internet Archaeology* 15, Council for British Archaeology, Department of Archaeology, University of York, England, UK <http://intarch.ac.uk/journal/issue15/richards_toc.html> 10 Dec 2006.

Reich, Vicky, and David S. Rosenthal
2001 LOCKSS: A Permanent Web Publishing and Access System. *D-Lib Magazine* 7(6). <http://www.dlib.org/dlib/june01/reich/06reich.html> 10 Dec 2006.

Science Commons
2006 Scholar's Copyright Project: Background. Creative Commons and Science Commons <http://sciencecommons.org/projects/publishing/background.html> 10 Dec 2006.

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

Snow, Dean R., Mark Gahegan, C. Lee Giles, Kenneth G. Hirth, George R. Milner, Prasenjit Mitra, and James Z. Wang
2006 Cybertools and Archaeology. *Science* 311(5763): 958–959.

Suber, Peter
2004 University Actions against High Journal Prices. *SPARC Open Access Newsletter* 72. <http://www.earlham.edu/~peters/fos/newsletter/04-02-04.htm> 10 Dec 2006.

Trant, Jennifer
2006 Exploring the Potential for Social Tagging and Folksonomy in Art Museums: Proof of Concept. *New Review of Hypermedia and Multimedia* 12(1):83–105.

U.S. Congress
2006 *Federal Research Public Access Act of 2006*, S 2695, 109th Congress, 2nd. Session, Library of Congress, Washington, DC <http://thomas.loc.gov/cgi-bin/query/z?c109:S.2695:> 10 Dec 2006.

Vaughan, Liwen, and Debora 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.

Willinsky, John
2006 *The Access Principle*. MIT Press, Cambridge, MA. <http://mitpress.mit.edu/catalog/item/default.asp?ttype=2&tid=10611> 10 Dec 2006.

Eric C. Kansa
Open Context
125 El Verano Way
San Francisco, CA 94127
<ekansa@alexandriaarchive.org>