

Quantity has a Quality all its Own: Archaeological Practice and the Role of Aggregation in Data Sharing

Eric Kansa, University of California, Berkeley (ekansa 'at' ischool.berkeley.edu)

Joshua J. Wells, Indiana University South Bend (jowells 'at' iusb.edu)

"Quantity has a quality all of its own."

(-Joseph Stalin, referring to the relative merits of investing in quantity over quality for the Red Army)

Abstract

Archaeological information on the Web is changing in ways that impact archaeological practice. Technical standards, copyright licensing, and Web services all blur boundaries between disciplines and organizations. They also make data aggregation easier. Consequently, the scope of "archaeological data" may grow beyond traditional field, survey, and collections data. Aggregators can also document how data are combined, navigated, and used. In other fields, aggregation services evolve into primary channels for information retrieval. Because aggregators enjoy increasingly privileged positions in new information environments, this paper explores documented benefits and drawbacks of imminent issues affecting archaeological research opportunities, professional expectations, and ethical challenges.

Introduction

In considering Web-based data sharing and its potential research impacts, quantity becomes a central issue. Any level of data sharing may spark some interest, and may improve efficiencies in some research practices, especially searching for references or comparative materials. But for really transformative changes that can promote wholly new ways of exploring the past, archaeological data must be available in both huge quantities and via dynamic services, to a community that is prepared to creatively utilize them. As described in this paper, these issues of quantity and dynamism have a profound impact in shaping technical design issues. Beyond technology, the challenges inherent in working with increasing quantities of dynamic data will also play increasingly important roles in shaping the instructional and training directions in our discipline. These changes will raise a host of methodological and even ethical challenges, as they expand options available for researchers to make new contributions to the discipline.

Quantity's Quality

The current landscape, which sees archaeological data fragmented among various content silos, hampers demonstration of clear data sharing benefits. Most online data sharing initiatives have primarily focused on developing "destination web sites." Such sites implicitly put people into the position of being a passive audience, clicking and accessing content through a predefined user interface offered by the site. A site report or a museum catalog is essentially reproduced on the Web.

Such approaches typically leave content in splendid, and sometimes beautifully designed, isolation. Usually, little thought is given to making collections open for interaction through alternative interfaces, or for making content open for use in alternative contexts. Potentially, other data published on the Web can be used by researchers to contextualize content published in a given collection. However, few archaeological data sources offer alternative interfaces, especially Web-services and APIs ("application program interfaces"),

that would allow content to escape a given silo so that it could be used and compared to data published by other sources.

While the "destination web site" approach does offer access to archaeological content, it does not fundamentally transform how the content can be used. One of the great potentials for data sharing is enabling wholly new research programs unprecedented in scope and analytic rigor (Snow et al 2006; Kintigh 2006). But this goal cannot be realized simply through access to data. Data needs to be portable, and open for aggregation and comparative analysis with other datasets, some of which may come from outside the disciplinary boundaries of archaeology.

Archaeologists widely vary in their research interests. Often, an individual's interests are quite specific and even esoteric. Typically, no one single collection grows large enough to obtain a "critical mass" required to sustain wide interest. It is more usually the case that bodies of relevant content will be published by several different sources. If one cannot effectively work across these different sources, the network effects expected from data sharing will never be realized.

So a key issue for the community to finally enjoy the research benefits of data sharing is to enable the data to reach the researchers, rather than expect the researchers to find the data. In other words, you may not be able to provide the best and most innovative way of presenting and using your archaeological data. But if your data is open and portable, someone else may be able to provide additional content and context that can unlock the hidden value of your data. Data portability is thus a key requirement for building a critical mass of useful content and reaching levels of quantity needed to enable transformative research.

Dynamic Services: A Foundation for Collaboration

We've already alluded to dynamic services in the discussion above. Dynamic services are another way of describing APIs and Web-services. These enable data portability so that content can flow out of one system for reuse in other contexts, especially other contexts that enable aggregation and comparison with other data sources. For example, the dynamic Web services in use on Open Context (<http://opencontext.org>), a free and open access data publication system, enable Open Context content to appear in a variety of other web based applications and collections. In effect, users can interact with Open Context content without having to navigate to Open Context. These capabilities now primarily serve search and discovery functions. Aggregation sites such as the OERCommons.org, opened.creativecommons.org, and Folksemantic.org all use Open Context Web services for search functions.

Besides search and aggregation functions, Dynamic Web services have other more specialized advantages. Google Earth represents the most widely used example. It would be impractical for archaeologists to develop their own Google-Earth-like interface for visualizing data. It is far more practical to publish data in the KML format, which in effect, means offering a KML Web service, so that one can take advantage of Google Earth's powerful visualization capabilities. Similarly, there are graphing and other data visualization applications freely available on the Web, all of which can be used if archaeological data publishers offer Web services for their data.

Use of Web services can allow archaeologists to use an emerging infrastructure of services specifically built to support scientific data sharing and curation. Open Context is now drawing upon data preservation and curation services offered by the University of

California's California Digital Library (CDL) 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"): These are special identifiers managed by an institutional repository. The repository has taken on a responsibility for insuring that the objects associated with these identifiers can be retrieved into the future, even if access protocols such as "HTTP" change.
- Data archiving: The California Digital Library also has data curation and stewardship processes to help maintain the integrity of digital data and also to migrate these data into new computing environments as required.

With these CDL services, citation and archiving of Open Context content is put on a strong institutional foundation. In this case, services provide more than convenience. It would be absolutely infeasible for Open Context to try to meet these essential preservation needs by itself. Data preservation requires hefty technical and institutional commitments beyond the means of Open Context's developers. Web services, such as those provided by the California Digital Library, can help enable lower barriers to participation in scholarly computation. The data curation needs can be handled by organizations with the required capacity. This leaves Open Context developers free to focus on meeting more discipline specific needs while still offering its users state-of-the-art levels of data permanence. In effect, dynamic Web services mean that a single effort does not have to solve all the problems inherent to data sharing. Rather, the job of solving these problems, including various visualization, preservation, and content development challenges, can be distributed among various collaborating projects and institutions. Such services are working to make a distributed data sharing and preservation infrastructure a reality.

We use the term "dynamic" services because that helps capture a crucial aspect of services valuable to data-sharing. Data delivered via services means more than access to content. Rather, dynamic services deliver data together with "back-end" processes that add value to those data. These may include data curation processes, as is the case with the CDL. Other back-end processes may include software-mediated responses to queries that filter or summarize a given dataset or collection. For example, Open Context provides a service summarizing the overall characteristics of its content via its faceted search application (see Hearst 2006 for general definition and discussion). By providing this summarized information, users gain a greater understanding of a collection as a whole (Kansa and Kansa in press; Jeffrey et al. 2009). This service requires Open Context to process data "on the fly." If Open Context only offered static datasets (downloadable from a fileserver), we could not offer this kind of summarization service, or enable users to select and filter for the specific data that matches their interests. Thus, there are important advantages to be gained in moving beyond Web access to data, but also making the data available via Web-based services that help with data processing and manipulation.

Secondly, publishing data in dynamic services helps make sure that downstream users of those services gain access to updated data. While an individual dataset may be relatively static and "frozen" at publication, the data publishing service itself will likely continue to publish more and more data. Over time, the service offers a more comprehensive and richer picture that may help inform some interpretation. Dynamic services help make sure downstream interpretations are based on up-to-date information.

This dynamism opens a new door for scholarship, allowing some researchers to make scholarly contributions "as a service." Traditional artifacts of scholarly production are static objects, such as papers, monographs, or more rarely, an archaeological dataset. However, as dynamic services increase in importance, activities relating to designing, selecting,

manipulating, and choreographing across these services will become an important area of research. In that sense, some important scholarly contributions will not take the form of a static article or book, but rather a service that dynamically responds to different requests. In some disciplines this trend is already impacting researcher practice. For instance, biomedical fields are extremely fast-paced and see volumes of publication far in excess than what can be read by practitioners. To better manage this information deluge, software agents are becoming increasingly important "audiences" for biomedical publications (see review by Markel 2009). These data and text-mining software agents are the outcomes of significant research and scholarly investment. The agents power dynamic services that respond to user (and machine) queries and return useful but ever changing results. There is an inherently dynamic nature of these research products, since these products dynamically generate responses from changing queries and expanding collections.

Data Portability First, Semantics Second

Not only does data portability make it easier to work with quantities of content capable of transforming research, but data portability also can enable researchers to take advantage of serendipitous opportunities. Researchers may discover interesting ways to combine a given archaeological dataset with a dataset from a different region, time period, or even a different discipline. The audience for archaeological data may be wider than the archaeological community, and at the same time, archaeologists may choose to work with datasets and services offered from other scientific disciplines, government and commercial sources (Borgman et al. 2007; Onsrud and Campbell 2007). The fact that not all research agendas and questions can be anticipated ahead of time must inform standards discussions, particularly in the difficult area of semantics.

Of course, common semantics and standards for archaeological data can greatly reduce the costs of finding and working with archaeological data. However useful, such semantics represent only a facet of the larger problem of interoperability. It would be too much to expect every publisher of possibly relevant content to adopt an elaborate standard promoted by archaeologists. Similarly, many potential consumers of archaeological data come from outside the discipline, and it would be a shame for data standards specific to archaeology to act as a barrier and complication for use by nonspecialists. The 2006 American Council of Learned Societies report highlighted how openness to collaboration and multi-disciplinary reuse of data represents a key requirement for humanities and social science cyberinfrastructure (2006: Chapter 3).

Fortunately, relatively simple and widely used standards can help bridge archaeological information with the wider context of data available on the Web. As witnessed by the vibrant "mashup" community on Web, even simple steps to make machine-readable data can be sufficient to enable many creative and some useful new applications of data. One of the most exciting areas of serious mash-up and programmable web development is in the world of "Open Government" (Wilde et al. 2009), where software developers, journalists, and advocacy groups have collaborated to make better sense of many different information streams made available from federal government sources. The semantics of these various sources often vary widely, and figuring out how to effectively combine disparate data sources is often not straightforward. Yet, because these sources offer sufficient quantities of data relevant to many public interest concerns, access to machine-processable data is enough to encourage their aggregation and analysis.

Thus, while archaeological standards and semantics are important concerns, they do not need to be "solved" to have useful levels of data interoperability. In the near term, archaeological data publishers should not neglect making their services mashup ready, using widely used open standards such as Atom, GeoRSS, and KML. Atom is useful because

it provides a widely understood "standard-container" for content, even content expressed according to a very specialized standard (such as a discipline specific markup language or RDF representation). Thus, even if downstream users lack the tools needed to process the specialized data, they can get some useful information from reading metadata expressed in Atom. As such, Atom underlies the most widely used types of Web services currently deployed on the Web (Wilde 2008). Besides Atom, Creative Commons licenses also help cross-disciplinary interoperability (Kansa et al 2005). These licenses offer both a common legal-licensing framework and a common technical metadata standard for expressing copyright related permissions and obligations. Atom, GeORSS, Creative Commons and the like are already widely supported by a large variety of free commercial services as well as open source software libraries. These simple tools and standards can build bridges across various archaeological data silos, and across different disciplinary boundaries.

As a demonstration of this "programmable web" or "mashup" approach, Open Context currently serves some 200,000 different contextual and finds records from 15 different projects and collections. For a sense of scale, the Metropolitan Museum of Art has about 150,000 different objects in its online collection. Since Open Context represents a small project, resulting from the mainly part-time effort of a few individuals, it is easy to see how even relatively modest efforts at data publication can quickly generate large bodies of content! Open Context makes this content available via several dynamic web services, using Atom, KML, and JSON formats. The [Bade Museum](#) and the [San Diego Archaeological Center](#) are currently using and testing Open Context's APIs for presenting their materials on their own websites. While these API applications are mainly for presentation purposes, we also have active collaborations with the NEH funded Pleiades Project (at New York University) and the German Archaeological Institute's Arachne project to use such web services for more substantive data sharing. These efforts will facilitate cross-collection searches and even semantic annotation (Kansa et al. in press).

Who Will Work in the New Environment

We wish to be explicit that this paper should not be interpreted as a single-track advocacy for open data as a self-realizing answer to current and future issues. Within this new environment of data quantity will be a need not just for niche archaeoinformaticians to oversee compliance, government, and research facilities; there must also be a parallel gearing-up of archaeological competencies with digital data in general. Repeated evidence within informatically-transformed sciences, humanities, and industries has shown that the achievements of communities in overcoming semantic issues and other issues of interoperability and organization are correlated with stakeholder training about data parameters and access to actionable models of data usage (Kling et al. 2005). To have a data-sharing community, and to achieve successful "mashups" with archaeological data means satisfying very different data criteria than are involved with mixing media, creating web-applications, or trading code. Openly producing and sharing useful data within a scientific community will require important standards of training beginning at the undergraduate level.

Archaeologists can learn from other scientific and applied disciplines that have crossed some of this new ground ahead of us. For instance, the American Medical Informatics Association has developed their "AMIA 10x10" program to drive medical workforce education and professional development in informatics through outreach (Hersh and Williamson 2007). 10x10 program instructors and students have a variety of new media and social networking outlets through which program participants can engage in a longer process of community learning through communication of shared experiences and new strategies.

A sample list of competency goals for a 15-week 10x10 course at the University of Illinois at Chicago ([AMIA 2010](#)) indicates that students entering the program with no prerequisite information technology training should leave the program with the ability to:

- Describe the concept and certification of an electronic health record (EHR)
- Explain bases for EHR standards, security, and privacy
- Formulate EHR implementation plans that respect “organizational culture and standards as well as personal work preference”
- Evaluate EHR outcomes
- Apply informatics theories to professional practice

If we consider the similarities between archaeological site records and EHRs the potential strength of the 10x10 model becomes more apparent; both types of records contain data on unique-yet-standardized investigations into individual incidences of meaningful traits that are defined by shared concepts, yet are recorded for local purposes with local schema. Furthermore, both forms of records contain sensitive information that must be considered in terms of professional ethics, legal restrictions, and privacy concerns.

Obviously ad hoc instances of these training processes are already present in archaeological science education. Various programs in the vein of NSF Research Experience for Undergraduate courses can offer specialized technical training. Many traditional undergraduate laboratory, fieldwork, and CRM courses include specific technological instruction that can be transferred to professional practice. However, there is no standard set to define what level of informatics competency an archaeological practitioner should be expected to have that would enable them to (in 10x10 terms) describe the concept of an electronic site record or excavation record and to recognize that (certification of quality aside) it has strengths and weaknesses for interoperability. A potential list could include:

- the general functions of a relational database
- basic differences in recording and querying qualitative and quantitative data in a database
- an awareness of point, vector, and raster data capabilities in GIS products
- basic competence in using the "programmable-Web" (web services, APIs), and the "Semantic Web" (linked data systems)
- ethics of archaeological information, including cultural and intellectual property issues, legal restrictions, privacy, and the political context of media and representation

These data are part of many day-to-day realities of working in archaeological science, and a functional understanding of them should not be learned on the job for finite purposes, but instead be considered as pedagogically important as measuring provenience, analyzing a diagnostic artifact, or researching a literature review. The need to develop core informatics competencies will only grow as the quality and quantity of Web-based archaeological data grows. As data resources grow in comprehensiveness, investing in informatics training will yield higher returns.

Future Directions

Expanding the quantity of professionally developed research data available via dynamic services creates network effects that expand research opportunities. As the amount of data expands, the range of possible ways to combine and recombine datasets also expand. While the majority of such combinatorial possibilities will yield meaningless or uninteresting results, some will yield significant new insights. At the same time, expanding Web-based publication of archaeological data will spur the development of new whole new types

of archaeological data.

For example, Open Context is collecting a growing corpus of how different datasets map to the ontology described by the University of Chicago developed "Archaeological Markup Language." These mappings, provided by a user through a special interface, are recorded so that eventually it will be feasible to (semi)automate semantic mappings. More automation in such mappings will enable Open Context to crawl different data providers and index their materials according to a common ontology. Potentially, the records of these mappings could also be used to facilitate text mining efforts on publications, especially the data tables presented in monographs and papers. This kind of semantic metadata collected by Open Context (and also related efforts of Digital Antiquity, the Archaeology Data Service, and others) is an emerging new form of archaeological documentation.

In addition, server logs and the like represent fascinating new types of "archaeological data." These records document how people find and use archaeological data. Search engines and many websites collect detailed interaction data on individuals as identified by IP addresses and user accounts. Detailed interaction data on individual users can better target advertisements and improve quality of search and recommendation services. Similar such data can be collected by archaeological data providers. These data are used to evaluate and rank impact and interest in different content. Not only are such records useful to improve user-interfaces and services, but they can help better understand what makes a certain dataset more valuable (or, more precisely, "used") than other datasets. More specifically, they may be invaluable for improving field documentation strategies, and improving semantic mark-up. In other words, metrics on how datasets are used may help drive convergence toward those semantic standards, services, and documentation practices that actually prove valued to the community. Potentially, rich metrics on how users interact with archaeological data will be of research and interpretive interest since such data will inform investigations of the archaeological research process.

Thus, there will be strong pressures to collect data about individual users of archaeological data. However, these pressures will tend to conflict with privacy concerns and norms well-established within traditional academic repositories, namely libraries. Libraries traditionally destroy all record of loans to patrons once the content has been returned. The rationale for destroying these data is to protect the anonymity of patrons as they access collections. This ethical stance has been shaped by decades of legal and political battles over issues of national security, limits on government powers, and freedom of speech. However, the migration of knowledge onto the Web changes this expectation of privacy. Amazon books, and especially, Google with its Web, book, and scholarly literature search facilities have no such ethical reservations in capturing data on individual user interactions. Because more and more scholarship is mediated through Google services, even if archaeological data publishers themselves forgo collecting user detailed metrics, Google would have likely already captured such metrics (note: we retain no individualized usage data with Open Context).

Beyond privacy concerns, the shift of research toward a Web dominated by commercial giants such as Google points to other potential worries. Google is, for all intents and purposes, a "black box" in how it finds, indexes and ranks Web content (Kansa and Kansa in press). While the PageRank algorithm that initially catapulted Google to success is published and well-known (Brin and Page 1998), Google continually revises and increasingly personalizes the ways it computes search results. As we come to rely upon Google more and more, this will impact our literature searches, research into comparative materials, and searches for primary datasets. Search engines like Google may bias the selection of research materials in unknown ways, and in doing so, search engines may bias some meta-

analyses that synthesize masses of data from disparate sources (Kansa and Kansa in press). Thus, the transformative research that archaeological informatics seeks to enable will likely be greatly shaped by a media landscape dominated by powerful commercial search-engines. Understanding these issues points back to the need for training. New scholars need to better understand this new information environment, including its biases as well as its capabilities.

The point of this discussion is not to condemn the use of Google for scholars. Google offers invaluable services, and exposing one's data to Google can do much to encourage wider use. Indeed, opening a collection to Google's Web-crawlers is one important step in making a collection less siloed. But the archaeological community should be wary of over-dependence on one, less-than-transparent service provider. Instead of exclusively relying upon Google, archaeologists should seek ways to make their research collections easier to find and use via alternative services. The kinds of dynamic services discussed above can help reduce dependence on Google. Because such dynamic services reduce the costs and complexity of looking across distributed Web-based collections, the researcher community can build their own aggregated, cross-collection search, retrieval, and query services, and do so in a way that is more transparent and open to scholarly scrutiny than Google.

Conclusions

We are in the midst of a rapid shift toward Web-based research. There are many, many collections of archaeologically relevant data on the Web, some of them quite large. However, the impact of this data explosion remains difficult to discern. There are still few examples of research outcomes based on these increasingly rich data resources. Training researchers in how best to use and contribute toward this expanding body of archaeological data is a critical issue. Also, the current fragmentation of archaeological data across a number of closed silos tends to inhibit innovative uses of these data. It is still too difficult to search across and use content locked in different collections, even if these collections are Web accessible. Unlocking these collections and enabling researchers to work with truly Web-scale quantities of data will require much greater attention to dynamic Web-services. Instead of expecting users to come to a collection, even a very significant collection, we should find ways of having collection content go to users. In doing this, we need to acknowledge that software-agents that provide and consume dynamic services will be an increasingly important "user community" to serve.

The Web is more than a medium for making data accessible, increasingly it is a platform for dynamically processing and transforming data. Thus, archaeological data publishers need to recognize the growing importance of automated processes and software that aggregate, republish, index and archive data in a host of different contexts. To best use the Web as a research platform, current and future archaeological data publishers must understand and leverage such software agents and services.

References Cited

ACLS (American Council for Learned Societies)

2006 *Our Cultural Commonwealth: The report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences.* <http://www.acls.org/uploadedFiles/Publications/Programs/Our_Cultural_Commonwealth.pdf>

AMIA (American Medical Informatics Association)

2010 AMIA 10x10 Partner Programs Description for University of Illinois at Chicago. <<https://www.amia.org/10x10/partners/uic/description.asp>> (Accessed Mar. 2010).

- Ankolekar, Anupriya, Markus Krötzsch, Thanh Tran, and Denny Vrandečić.
2008 "The two cultures: Mashing up Web 2.0 and the Semantic Web." *Web Semantics: Science, Services and Agents on the World Wide Web* 6: 70-75.
- Borgman, Christine, Jillian Wallis, and Noel Enyedy.
2007 "Little science confronts the data deluge: habitat ecology, embedded sensor networks, and digital libraries." *International Journal on Digital Libraries* 7: 17-30.
- Brin, Sergey, and Lawrence Page
1998 "The anatomy of a large-scale hypertextual Web search engine." *Computer Networks and ISDN Systems* 30: 107-117.
- Hearst, Marti
2006 "Clustering versus Faceted Categories for Information Exploration." *Communications of the ACM* 49: 59-61.
- Hersh, William and Jeffrey Williamson
2007 Educating 10,000 Informaticians by 2010: The AMIA 10 x 10 Program. *International Journal of Medical Informatics* 76: 377-382.
- Jeffrey, S. et al.
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.
- Kansa, Eric C, Tom Elliot, Sebastian Heath, and Sean Gillies
In Press. "Atom Feeds and Incremental Semantic Annotation of Archaeological Collections" In *Computer Applications and Quantitative Methods in Archeology – CAA 2010*. Eds. Javier Melero & Pedro Cano (Editors)
- Kansa, Eric C and Sarah Witcher Kansa.
In Press. "'Mashable' Heritage: Formats, Licenses and the Allure of Openness", Pp. 105-112 in *Heritage in the Digital Era*, Multi-Science Publishers, London.
- Kansa, Eric C., Jason Schultz, and Ahrash N. Bissel.
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.
- Kintigh, Keith, W.
2006 "The Promise and Challenge of Archaeological Data Integration." *American Antiquity* 71: 567-578.
- Kling, Rob, Howard Rosenbaum, and Steve Sawyer
2005 *Understanding Social Informatics: A Framework for Studying and Teaching the Human Contexts of Information and Communication Technologies*. Information Today, Inc. Medford, New Jersey.
- Markel S,
2009 BioLINK Special Interest Group Session on the Future of Scientific Publishing. *PLoS Comput Biol* 5(5): e1000398. doi: 10.1371/journal.pcbi.1000398

Onsrud, Harlan, and James Campbell

2007 "Big Opportunities in Access to "Small Science" Data." *Data Science Journal* 6: OD58-OD66.

Snow, Dean R. et al.

2006 "Cybertools and Archaeology." *Science* 311:958-959.

Wilde, Erik, Eric C. Kansa, and Raymond Yee.

2009 "Web Services for Recovery.gov." *UC Berkeley, School of Information Technical Reports* <<http://escholarship.org/uc/item/0fv601z8>>.

Wilde, Erik.

2008 "The Plain Web." Beijing, China: WWW2008 <<http://dret.net/netdret/docs/wilde-wsw2008-plain-web.pdf>> (Accessed May 28, 2008).