Making Legacy Literature and Data Accessible in Archaeology

Dean R. Snow[1]

[1]Department of Anthropology, The Pennsylvania State University. USA.

Abstract

Archaeological publications, including old ones, are increasingly available in electronic form. JSTOR, Google Books, and other services are digitizing an increasing array of journal back issues and out-of-print books, and providing them as PDF files with optical character recognition enabled. Virtually all publications going forward will be available in searchable electronic formats. The data that underlie publications are also increasingly available, and the Digital Antiquity initiative discussed by other papers in this session is a major step forward in the archiving of databases. However, that still leaves many thousands of unpublished gray literature reports inaccessible or very difficult to access. CRM reports are an important source of legacy information for many researchers, but making them accessible will require us to overcome many obstacles. Government agencies sequester many reports in order to protect resources; older reports are often unreadable by optical character recognition systems; there is and probably never will be a single repository for all such resources. Proposals to overcome these obstacles are discussed. These include the development of additional new tools for searching a distributed network of repositories, further development of ArchSeer and other specialized search tools, and development of a secure software package that will be attractive to the guardians of gray literature repositories.

***Keywords**: ArchSeer, DRM, legacy, gray literature, cyberinfrastructure, search*

1 INTRODUCTION, BACKGROUND

We are in a new era, what Hal Varian calls a period of "combinatorial innovation."[1] As he points out, in the nineteenth century the focus was on interchangeable parts. In the first part of the twentieth century it was on electronics, in the second half integrated circuits. Now it is on open source software, crowdsourcing, and limitless APIs (application programming interfaces).

The urgent need for the development of cyberinfrastructure in archaeology in this new era is all but self-evident.[2] We preserve what we can of the evidence of our past, and we document as best we can both this and the evidence that circumstances do not allow us to preserve in place. The published archaeological literature has grown exponentially over the last century. Preservation of and access to this literature, electronic and otherwise, is a formidable problem for libraries, including digital ones. More serious, because they are far less accessible, is the accumulating legacy of unpublished databases and unpublished "gray literature" reports. Legacy databases exist in both paper and electronic forms, often the cumulative results of long careers in archaeology.

The creation of Digital Antiquity, with the database repository known as tDAR at its core, has been designed to solve the legacy database problem. Files of other types can also be accommodated. This project is the subject of Keith Kintigh's presentation. Fred Limp discusses the technical issues involved in record preservation. My focus is more on the context of archaeological practices, and the consequent problems that face us as we try to secure the future of the past.

Databases and their curation present a host of problems for users, of course. Even if the databases are clean, their terms well defined, associated metadata are intact, and the cyberinfrastructure software is up and running well, using two or more databases together almost always requires a great amount of analytical time and effort. For example, A'ndrea Messer recently used only five legacy databases from the small and circumscribed Mesa Verde region in Colorado for her doctoral dissertation at Penn State. Standardizing their contents, adding missing data, and correcting errors required countless hours of work with original notes and derivative publications.[3] One of the five databases, the one most of us would expect to be the best of the lot, proved to be worse than useless. It was misleading in subtle ways that were only discoverable through hard work and careful comparison with the other databases.

[1] Hal Varian, "Hal Varian on how the Web challenges managers," *The McKinsey Quarterly* (2009), www.mckinseyquarterly.com/Hal_Varian_on_how_the_Web_challenges_managers_2286 (accessed Jan. 15, 2009).

[2] Dean R. Snow et al., "Cybertools and Archaeology," *Science* 311 (2006): 958–959; Dean R. Snow et al., "Envisioning an Archaeological Cyberinfrastructure," *The SAA Archaeological Record* 6 (5) (2006): 15–16; Keith Kintigh, "The Promise and Challenge of Archaeological Data Integration," *American Antiquity* 73 (1) (2006): 567–578.

[3] A'ndrea Messer, *Small Ancestral Pueblo Sites in the Mesa Verde Region: Location, Location, Location* (Ph.D. diss., The Pennsylvania State University, 2009).

The gray literature presents problems different from those presented by databases. Databases are often subsets of largely narrative gray literature reports, but they also exist separately as stand-alone files. In the United States, most gray literature is a vast but largely hidden literature that has accumulated as unpublished reports resulting from cultural resource management (CRM) projects over the last few decades. Thousands of these reports are produced every year and then archived mainly in federal agencies, state agencies, and company libraries. It is likely that in recent years more of the most valuable new information about archaeological resources in countries like the U. S. has been generated by CRM projects than by traditional academic projects. Yet except to the degree that CRM reports are mined by their producers to produce derivative publication in traditional outlets, that information remains sequestered.

It is important to note that it is also in the nature of CRM research that many reports contain valuable negative findings. Negative results are rarely reported at all in most disciplines, to the long-term detriment of those disciplines. For example, a particular medical treatment might be tested but found to be ineffective. The results of such a negative finding never get published, and other researchers having the same idea are doomed to repeat the same initial trials with the same negative results. Archaeological reports that produce useful negative results might not get published, but the findings persist in the gray literature. Thus even archaeological gray literature reports that report negative results are of considerable use and interest. But these too typically remain sequestered and very difficult to access.

2 ACCESSING THE GRAY LITERATURE

My purpose here is to lay out some problems and solutions regarding legacy archaeological resources generally and the gray literature in particular. Penn State University researchers have developed a specialized search engine called "ArchSeer" to extract standard data from such sources. However, many problems remain with respect to accessibility, volume, and the structural peculiarities of the gray literature sources, as well as with respect to the chain of current practices that currently lead from data acquisition through analysis to dissemination and preservation. I will define some problems inherent in any effort to solve these problems and to identify some tools to solve them. My main point is that we must be mindful of some critically important larger problems so that we do not focus too exclusively on the admittedly formidable technical problems we face going forward.

Archaeological publications, including those published long ago as books or in our leading journals, are increasingly available in electronic form. JSTOR, Google Books, and other services are digitizing an increasing array of journal back issues and out-of-print books, and providing them as PDF files with optical character recognition enabled. This is a good thing, of course, as it makes knowledge more widely available to all.

It also seems likely that virtually all publications going forward will be available in searchable electronic formats. There are copyright issues, of course, but it is in the interests of publishers to make works available in as many ways as possible so long as they can profit or at least break even at the same time. The data that underlie publications are also increasingly available, and the Digital Antiquity initiative discussed by other papers in this session is a major step forward in the archiving of databases.

However, the electronic availability of older books and journal articles and the promise of the electronic availability of virtually everything going forward still leave many thousands of unpublished gray literature reports inaccessible or very difficult to access. Moreover, even if we find ways to retroactively make all the legacy reports available, we might succeed only in swamping potential users with more information than they can handle. My purpose here is to explore the range of problems and solutions entailed by these circumstances.

3 LEVELS OF PROBLEMS TO BE SOLVED

We can usefully consider the issues before us in strategic, tactical, and technical terms in turn. Strategy defines the tactical problems we must solve and they in turn frame the technical problems. I will set the last aside for present purposes for a variety of reasons, not the least of which is that others are more skilled at addressing them than I am. But perhaps most importantly, I think that we tend to misperceive the larger problems as sets of technical ones. If we do not solve the strategic and tactical problems as we develop the technology, we could fail in the effort.

4 STRATEGIC PROBLEMS

The overall strategy should be to make archaeological knowledge widely available to both professional archaeologists and society at large. When I was young, knowledge was expensive, in the sense that it was difficult to find and access. Researchers acquired it, sometimes through heroic effort, often sequestered it, and parceled it out in publications designed to maximize credit for their work. Today knowledge is comparatively cheap, or at least it appears to be, and there is a concomitant poverty of attention. So much information is available on line that most people assume that all knowledge is or should be instantly available. Many people are unaware of how much information remains sequestered, sometimes deliberately, sometimes inadvertently. Of course it is also the case that many consumers of information lack the skills to distinguish reliable information from nonsense in an electronic world where both abound.

Nevertheless, we are mindful that society at large has already paid for the knowledge archaeologists have generated. The gray literature generated by cultural resource management projects in particular is a vast knowledge base, and we owe it to the citizens who funded the projects that generated it to make it all available in one form or another. Our strategy is to find ways to do so, for this would also redound to our benefit in the long run. Quite apart from an ethical obligation to make gray literature information available, it is an important source of reliable information to counter widely available misinformation. It is also a fertile source of reliable data that students and researchers can use to advance disciplinary research.

But recent experience has shown that it is not possible to simply throw open the gates. Even if it were possible to do that, it is not reasonable to expect happy results. Facts do not speak for themselves, and they never have. Most people are overwhelmed by the amount of information that is already available, and offering to double or triple the amount that is available every few months is not a viable solution. The information explosion will not be tamed by merely adding to it. There are, in addition, tactical problems that must be overcome, and they are considerable.

5 Some Tactical Problems

Having identified a major strategic problem, I turn now to some specific tactical problems that are entailed by it.

Gray literature reports are distributed and resist centralization.

Collections of gray literature reports can be found in government agencies, museums, universities, and at private firms, depending upon who produced them and for whom they were produced. Comprehensive repositories tend to be those of government agencies. These are national agencies in many instances, but in the United States and Canada primary government repositories are at the state or provincial level.

Some initiatives have been undertaken to encourage professional archaeologists to make their unpublished reports and data more accessible. The Alexandria Archive is a good example. The Digital Antiquity database repository system currently being incubated at Arizona State University is another. The National Archeological Database, which was initiated by the National Park Service and is currently hosted by the University of Arkansas, is still another. These can be grown and linked, but they remain voluntary and could die young of neglect. When it comes to gray literature reports, none of the options developed so far is likely to capture more than a small minority of the thousands of reports that have been generated in recent decades.

Information on site locations is often suppressed.

It is common practice and often a legal requirement for government agency repositories to limit repository access to authorized individuals. Repositories also often suppress locational information in order to protect archaeological resources. While this appears to deter the casual looter, it is often said that more persistent and sophisticated looters probably already know the locations of productive sites better than do most professional archaeologists.

In some countries, such as the United Kingdom, archaeological site integrity is often maintained through the recruitment of nearby residents as stewards of local archaeological resources, a successful proactive approach. Archaeologists in the United States are increasingly of the opinion that a similar approach would be more productive than the secretive one currently in place. Experimental efforts in states such as Louisiana and Utah have shown that identifying sites with prominent signs that implore people to protect and preserve archaeological resources is a more effective approach than secrecy. However, this idea remains counterintuitive to many archaeologists and in any case contrary to federal policy in the U. S. I do not expect the policy to change soon.

There is too much unprocessed information.

Archaeological databases, documents, images, GIS files, and the like must be curated and made available to professionals, but we are not the only legitimate audience. Even professionals must now concede that there is too much unprocessed information for everyday purposes. If we professionals are overwhelmed, consider what the problem must look like to the nonprofessional.

NSF established the National Science, Mathematics, Engineering, and Technology Education Digital Library (NSDL) in 2000 as a vehicle for delivering content to anyone that wanted it, but with secondary school teachers as a primary target audience. Various pathways were established to package information in digestible amounts. One of these is BEN (biosciednet.org). So far BEN has 11,000 peer-reviewed resources, and the goal is to push that total to 25,000 by 2010. But users are already overwhelmed, and they have not made much use of the resource. There is too much unprocessed information for the nonprofessional, and too little that is new for the professional. For those reasons the resource is not much used.

The problem is that more is not necessarily better. One discouraged user said "I'd rather have a nice-sized catalog of peer-reviewed material that promotes active learning than a vast amount of stuff that hasn't been vetted."[1] As an analogy, it is useful to remember that some stores are too big; they discourage shoppers by offering more than can be comprehended and processed in the shopping time available. This may seem counterintuitive in the information age, but it is a tactical problem that should loom large in our thinking.

Google is very good at finding information, but even the most constrained search can sometimes return many thousands of hits. Most of them are junk, of course, but how does a middle school teacher sift through it all and pull out reliable information for a lesson that has to be delivered in a matter of hours? For that matter, how does a professional accomplish it?

Adequate review and validation requires professional mediation.

It is the obligation of professional archaeologists to process, synthesize, and make the results available via peer reviewed outlets. We have traditionally done this through popular articles and books, encyclopedia entries, film, and video presentations. That is no longer good enough, either for ourselves or for the public that supports us. Nonprofessionals expect to find the equivalent of encyclopedia entries on line, and most of them are smart enough to realize that much of what they are finding there cannot be trusted. I am still writing the occasional encyclopedia entry, but I have no expectation that more than a small fraction of the people who need access to such short contributions will ever find them. This is an expensive, cumbersome, and quickly obsolete solution to the problem, and it is no longer viable. Where is the archaeological equivalent of WebMD, a professionally mediated resource that is very dependable? How do we go about creating such a resource for archaeology?

Spatial references are often fractional, not Cartesian.

Tools like ArchSeer have the capacity to automatically extract even well hidden standard data from gray literature reports. But there are limits to this capability. Modern GPS coordinates are recorded in a handful of alternative formats, each of which can be made recognizable to the program so long as reports are in proper electronic format. This will be effective going forward, as reports are increasingly submitted in electronic form.

However, older reports often record site locations in terms of nested and fractional systems of land tenure rather than in terms of Cartesian coordinates. On the East Coast of the United States this form of reckoning is usually laid out in terms of a hierarchical listing of nested categories: state, country, minor civil division, and parcel. In the interior of the country the nineteenth-century Township and Range system is more often used. Most of the landscape was surveyed into a checkerboard of 36-square-mile townships, each subdivided into 36 mile-square "sections." If considered an archaeological site, the house I grew up in would be located by means of the following statement: "The northwest corner parcel of the Northeast Quarter Section 32 of Township 110 North, Range 32 West, Brown County, Minnesota. I know of no way to turn such a statement into GPS Cartesian coordinates without expert human intervention.

Furthermore, much locational information from within archaeological sites, particularly those excavated when all data were recorded on paper, conforms to the same nested, fractional approach. The location of a object from one of my own early excavations originally took the form "subquad 2, quad A, square 56 (N27W3)." These were later converted to simple Cartesian coordinates, but only with considerable human mediation. We can expect that most excavation reports from before the 1980s will resist easy extraction of locational data.

Older reports are not usually available in PDF format and when they are it is typically without optical character recognition (OCR).

We cannot expect that the staffs of repositories holding gray literature reports will scan older reports and produce PDF versions of them with OCR invoked. This is a very expensive proposition, and the physical conditions of many older reports will frustrate attempts at optical character recognition. Some reports might even be handwritten. My own experiments with older documents in typescript or from dot matrix printers have been discouraging. The best we can reasonably hope for is that only the abstracts of many legacy reports will become widely available, and without OCR. We will have to find ways to work within those constraints.

6 Some Solutions

It is our professional duty to make information available to each other and to the public at large. But data do not speak for themselves and some data are lies. Some are merely obsolete. False hypotheses may be winnowed out of the recent scholarly consensus, but they live on in older literature, sources that are as easily accessed as the newer propositions that we think have replaced them. Thus it is not good enough to simply make information available, although we certainly must do that rather than sequester it. We can let commercial outlets like The Archaeology Channel do it for us, but in that case the profession has too little quality control and thus insufficient influence on content. A better solution is a resource like WebMD hosted by one or more

[1] Jeffrey Mervis, "NSF Rethinks its Digital Library," *Science* 323 (2009): 54–58.

professional organizations, with mechanisms to maintain peer review and quality control.

Hierarchical Access

If we spend all our time and effort speaking and writing only for each other we do not deserve the support we expect from society at large. Viewed this way it appears that we should be thinking about how we propagate archaeology in hierarchical terms. Most users need to be able to access archaeology at the most abstract level. That is where most of them will stop, because they will lack either the time or the inclination to dig deeper. That's alright, for if we want middle school teachers to use archaeology, we have to make it efficient for them to do so.

Digital Antiquity

The Digital Antiquity project is a great start, but we have a long way to go. Sensitive data can be sequestered, but we are obliged as a consequence to provide benign derivatives. Thus, if we have to sequester sensitive data, then part of our professional obligation must be to provide second-order descriptions that provide the basics in a form abstract enough to be harmless yet descriptive enough in character to be meaningful. But while this is necessary, it is not sufficient, given the vastness of the information available on most topics. Thus we must consider problems of critical evaluation, synthesis, security, and dissemination as we are working mainly on technical problems in the early going.

Government Policies

We will have to work with government agencies everywhere to find ways to make databases, documents, images, and so forth easily available to users. In many cases it will be possible to work with agencies to require electronic reporting and their mandatory submission to trusted repositories. But that is only the first part of the larger problem. Making the electronic resources not just available but attractive must also be a major concern as we work through the inevitable technical problems.

Viral Tools

We are professionally obliged to generate third order synthetic overviews about archaeological information. This is a step beyond first or second order description and it entails the drawing of at least some inferences. At this level we begin to enter the realm of potential controversy. The way to balance the need for generalization while accommodating the possibility that reasonable people will differ is to use a wiki. This will work so long as everyone has access to the results on line but only professionals have editorial rights. I believe that there are reasonably easy ways to manage this. The way to counter the attention deficit that this much information produces is to build viral tools that will attract users and get them to do much of the work that is needed. Google Maps is a good example. If you are in the map business you are not just competing against a couple hundred Google employees. Hundreds of thousands of users are using open APIs to create new map products on the fly. We cannot succeed if we do not make our resources widely available in similar ways and provide tools that will prompt thousands of users to participate in refining our intellectual products.

We might need to find ways to reward professionals for working at the production of these kinds of dynamic resources, but my guess is that most will participate rather than let someone else author entries on topics for which they perceive themselves to be the experts. Professional activity should include the production of popular publications, including on-line ones, and whoever is counting the beans should be encouraged to take note of this kind of academic activity. Properly structured, a wiki has viral characteristics that draw in users. The primary tool at the heart of Digital Antiquity is tDar. It is not yet a viral tool, but it could be.

Collaboration with Archivists

We should follow the lead of forward-thinking archivists, who are currently working through the emerging issues surrounding document preservation.[1] Archivists are urging us to distinguish between digitization for the purposes of preservation and digitization for the purposes of access. At the same time they argue that we have to distinguish between a growing number of alternative media (paper documents, photographs, audio files, born digital documents, and so forth). There are issues of authenticity and integrity to worry about, particularly in the case of paper records. Also entailed are serious questions regarding evaluation and selection, something that any archaeologist who has ever had to decide what to keep during the course of excavation should be familiar with. It is possible to create a digital collection based on inappropriate criteria, that is, criteria that do not match the purpose of a digital repository. We need to solve some very specific problems, including:

1. the most appropriate ways to preserve through digital reformatting;
2. incentives to get archaeologists to be more proactive in archiving their materials;
3. propagation of metadata standards;
4. promotion of appropriate criteria for evaluation and selection; and
5. making training in preservation and archiving as important as acquisition and analysis in graduate training.

[1] Doris J. Malkmus, "Documentation Strategy: Mastodon or Retro-Success?" *American Archivist* 71 (2) (2008): 384–409.

The last point is particularly important. The “compleat archaeologist” is one that responsibly carries out the tasks of research design, data acquisition, analysis, and publication of results. But the cycle of project tasks is not completed by those four steps. We must also be committed to digitizing all of our records and depositing the resulting files in suitable repositories where others can make full use of them.

We have only just begun to solve this complex array of interlocked problems, but I can think of little that is more urgent for the long-term health of our profession. I am very pleased to be involved in this symposium, and hopeful that our collective efforts will propel us all forward as our discipline evolves and grows through the course of this new century.

ACKNOWLEDGEMENTS

I thank Bernard Frischer for inviting me to represent the Society for American Archaeology in this symposium, and I thank Jeffrey Altschul, Keith Kintigh, Fred Limp, and Julian Richards for their participation.

BIBLIOGRAPHY

Kintigh, Keith. “The Promise and Challenge of Archaeological Data Integration.” *American Antiquity* 73 (1) (2006): 567–578.

Malkmus, Doris J. “Documentation Strategy: Mastodon or Retro-Success?” *American Archivist* 71 (2) (2008): 384–409.

Mervis, Jeffrey. “NSF Rethinks its Digital Library.” *Science* 323 (2009): 54–58.

Messer, A’ndrea. *Small Ancestral Pueblo Sites in the Mesa Verde Region: Location, Location, Location*. Ph.D. diss., The Pennsylvania State University, 2009.

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

Snow, Dean R., Kenneth Hirth, and George Milner. “Envisioning an Archaeological Cyberinfrastructure.” *The SAA Archaeological Record* 6 (5) (2006): 15–16.

Varian, Hal. “Hal Varian on how the Web challenges managers.” *The McKinsey Quarterly* (2009). www.mckinseyquarterly.com/Hal_Varian_on_how_the_Web_challenges_managers_2286 (accessed Jan. 15, 2009).