Peter J. Cobb. Collaborative Online Bibliography for Archaeology. A Master's Paper for the M.S. in I.S degree. July, 2008. 39 pages. Advisor: Bradley M. Hemminger

Annotated bibliographies are a key tool for students in any academic discipline because they improve the efficiency of the discovery process for relevant resources. The field of archaeology offers a unique challenge to the bibliographer because it encompasses so many dimensions, including time and space. This paper documents my evaluation of the top social networking and bibliographic software on the internet for their suitability for use by archaeology students. Web-based software improves the bibliographic process by allowing users to share their work. For archaeologists, online software also provides innovative techniques for indexing citations, such as geographical browsing interfaces. As a result of my research for this paper, I have chosen two systems that I will continue to experiment with in my own archaeological research process.

Headings:

Archaeology literature/Bibliography
Social networks/Evaluation
Geographic names
Information systems/Evaluation

COLLABORATIVE ONLINE BIBLIOGRAPHY FOR ARCHAEOLOGY

by
Peter J. Cobb

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Information Science.

Chapel Hill, North Carolina

July, 2008

Approved by

____

Bradley M. Hemminger

Table of Contents

Introduction to the Problem

Two years ago, I joined an archaeological excavation at a small site in the middle of Turkey. I had a few months to prepare for this excavation before we actually went out into the field. I was determined to locate and read all of the published material I could find about this obscure site. The task was made difficult by the fact that the site had not been excavated in 15 years and that all the material about it was published in Turkish. I was very lucky to locate via a Google search the PDF version of a recent article in English that analyzed some of the cultural remains from this site. This article's citation list was critical to my ability to quickly locate the main excavation reports. I also relied on senior members of the current Turkish and American team to provide me with citations, although none of us were able to locate one particular article that was rumored to have existed.

Specialists in any academic discipline are expected to attain a broad familiarity with published scholarly sources on their topic. One of the challenges of being a student is to locate and learn about all of the relevant sources. Traditionally, a number of tools have been leveraged to facilitate this process. In addition to social networks and the citation lists at the ends of articles, there are also annotated bibliographies which are constructed specifically to increase the pace of evaluation and access to sources on a specific topic. Naturally, digital network technologies have offered improved methods of utilizing all of these old tools. The goal of this SILS master’s paper is to research, evaluate, and begin to implement a solution for collaborative, online bibliography designed specifically for the academic field of archaeology. This bibliography would leverage the efforts of the community to guide researchers to useful sources, regardless of the type and format of the sources. Annotations and reviews would aid the researcher in making informed decisions about which sources to utilize.

Archaeology offers some unique challenges to the bibliographer because it encompasses so many dimensions. Subject classification is only one way to organize sources for discovery. It would also be very useful to discover archaeological sources using geographical and temporal information. For example, if a student wished to locate all published excavation reports from a particular region, they should be presented with a map interface that might allow them to outline a boundary as a search criterion. There are also different types of sources. Original excavation reports are as close as one can usually come to a "primary" source. There are also many secondary sources that reference data from the original excavations. These secondary sources often enhance our understanding of material remains, usually by comparing data from different excavations. Sources can come in multiple formats, such as satellite photos. Finally, the ancient objects and the sites themselves can be considered sources, and these should be referenced from a bibliographic system. An archaeologist may wish to revisit architectural ruins if they still exist, or to view an object in a museum.

One of the best online bibliographies for an archaeological site is that of the Gordion site in Turkey.[1] It consists of a series of manually-maintained html pages with

[1] http://home.att.net/~gordion/bibliography/bibliography.html

a subject index. It is thorough and relatively up to date, but it takes a lot of effort to maintain and it is static. The “state of the art” for research bibliographies is much different in other academic disciplines. Perhaps the best existing example of an online bibliographic resource is MEDLINE, a database of articles in the biomedical field. This database is very comprehensive and can be searched on standard fields such as author, title and topic. It is able to provide a high level of service since it is maintained by the U.S. government. There are also many new internet technologies that have been developed over the last five years that help people share resources with each other. By looking at other academic disciplines and other technologies, it should be possible to begin to build a better tool for the creation and maintenance of archaeological bibliographies.

In this master’s paper, I will first evaluate the existing technologies that can be leveraged to create an online collaborative bibliography. As an archaeologist, I do not have the time to devote to building and maintaining my own system. However, none of the existing systems yet fit the specific needs of the archaeological community completely. There may be opportunities to combine existing tools in new ways in order to meet these needs. With this master’s paper, I will begin to articulate these additional needs so that I can communicate these to the developing communities. After choosing the two best tools to start with, I will use these in my own archaeological research process over the next few years. This will enable me to share my archaeological research process with other researchers and help them find useful resources. The ultimate goal is that online collaborative bibliographic tools will enhance archaeological students’ ability to conduct research efficiently.

Related Digital Archaeological Research

Collaborative bibliography is only one way that computers are used to improve the way archaeologists work. In fact, archaeologists have experimented with computer systems for a longtime,[2] and within the past decade, computers have become an indispensable archaeological tool. One extremely important type of information source in archaeology is raw excavation data, the records about everything a field archaeologist discovered while digging. Unlike published journal articles referenced from bibliographies that can be found in most academic libraries, it has never been easy to track down original data. Today, many digs are producing large amounts of digital data, which opens the possibility of universal online access directly to the data. Incorporating pointers to original excavation data from an online bibliography will help improve scholars' ability to quickly locate comprehensive information about an excavation. This section reviews important research and development work that aims to enable efficient data access and to help speed the adoption of digital tools by the archaeological community. Many of these existing digital archaeology projects provide excellent integration points for this bibliographic project.

Access to original excavation data in digital format from multiple archaeological sites can improve the efficiency of cross-site analysis. For instance, automated queries could discover the distribution of related objects across a landscape. Governments have often taken a role in providing access to multi-site data. One example of this is the

[2] Chenhall, 1968.

Archaeological Data Services (ADS) in the UK[3], and its European cousin, Archaeological Records of Europe - Networked Access (ARENA).[4] In addition to cataloging most archaeological sites in the UK, the ADS archives and provides access to the raw data from some of them. They are also developing web-based tools to improve a user's ability to browse the data. In the US, the National Park Service maintains geographical information on many American sites in its National Archaeological Database MAPS project.[5]

A number of organizations also disseminate information about computer technology in archaeology. The promotion of standard practices between digs will help to improve interoperability in the future. The Society for American Archaeology (SAA) has an interest group called Digging Digitally[6] that communicates about technological developments. Another group, the Center for the Study of Architecture/archaeology[7] (CSA) publishes a newsletter with technical recommendations for using computers in the field of archaeology.

A geographical framework and community-based content creation are two important factors in our ability to create useful multi-site information systems as well as collaborative bibliographies. The Ancient World Mapping Center (AWMC)[8] at UNC is creating a website, called Pleiades, which will provide scholars a forum for associating ancient place names to actual geographical locations. By providing a forum for

[3] http://ads.ahds.ac.uk/
[4] http://www.archaeoinformatics.org/
[5] http://www.cast.uark.edu/other/nps/maplib/
[6] http://www.alexandriaarchive.org/blog/
[7] http://csanet.org/
[8] http://www.unc.edu/awmc/

discussion open to anyone who wishes to contribute, Pleiades will leverage the combined resources of the community to produce the best results. In addition, an editorial process will ensure the quality of the system's final data. While the maps will be based on those published in the AWMC's *Barrington Atlas*[9], the system will also integrate with popular GIS systems such as Google Earth. Thus, Pleiades will provide a geographical framework for other archaeological systems to tie multi-site data together spatially. There is also a bibliographic component in Pleiades, to allow contributors to support place naming decisions with both ancient and modern sources. When this component is built, it will be one of the best tools for integrating geographic and citation data.

A few groups have already built systems that actually provide the ability to search excavation data from multiple sites. The Electronic Tools and Ancient Near East Archives (ETANA) is one project working to connect data from multiple dig sites.[10] The main focus of the ETANA project is the creation of a catalog of resources about Near Eastern archaeology. This catalog contains http links to articles, web sites, and digitized public domain books, and thus will be an important resource for an online archaeological bibliography. In addition, for one of their projects they created an information system[11] for sharing excavation data. This system combines data from eight excavations[12] and enables searching for data across all excavations at once. It also supports the Open Archive Initiative Protocol for Metadata Harvesting (OAI-PMH),

[9] http://www.unc.edu/awmc/batlas.html
[10] http://www.etana.org/
[11] http://digbase.etana.org:8080/etana/servlet/Start
[12] http://digbase.etana.org:8080/etana/htmlPages/etanadl_collections.htm

which enables sharing of metadata between web-based systems. ETANA has also been developing an ontology for archaeology, a tool for organizing and mapping the semantic relationships of archaeological concepts and vocabulary.

The Online Cultural Heritage Research Environment (OCHRE)[13] project has developed an eXtensible Markup Language (XML) schema for flexibly encoding archaeological data called ArchaeoML.[14] Through the Open Context[15] project, another organization, called the Alexandria Archive Institute,[16] provides an interface for searching and browsing data encoded in this format. Eleven projects or collections[17] currently have contributed data to Open Context. The Alexandria Archive also encourages data publication under the Creative Commons[18] copyright license, which makes the data available for academic use but still protects the rights of the data owners. Microsoft Research has provided support to develop another online database tool for archaeologists called Nabonidus.[19] However, this project does not use a widely-accepted license like Creative Commons to make data ownership and usage issues explicit. The project also raises concerns because it is dependent upon non-open-source Microsoft coding technologies and thus it is tied to the survival of a single private corporation.

[13] http://ochre.lib.uchicago.edu/
[14] http://ochre.lib.uchicago.edu/index_files/ArchaeoML_Schema.htm
[15] http://www.opencontext.org/
[16] http://www.alexandriaarchive.org/
[17] http://www.opencontext.org/database/browse_summary.php
[18] http://creativecommons.org/
[19] http://nabonidus.org/

A group called Archaeoinformatics.org seeks to apply informatics techniques developed for the sciences to archaeology.[20] The work of the Bioinformatics community, for example, should serve as a model for archaeology. All of the diverse projects mentioned in this section exhibit that there is general agreement among archaeology technologists about the need for computer systems to share archaeological research. A collaborative online bibliographic system would be a key part of the solution by providing an organized way for people to find all of the other resources.

Related Information Science Research

The process of organizing textual documents for improved access is at least as old as the Hellenistic period of the civilizations around the Mediterranean Sea. During the final few centuries BCE, libraries like those at Alexandria and Pergamon collected such a large number of books that organization became essential. Callimachus, who worked at the library of Alexandria, is credited with writing one of the first bibliographies. His work, *Pinakes* cataloged Greek literature and organized it by subject.[21] Yet, bibliographic organization surely began much earlier in the Bronze Age archives of Egypt and Mesopotamia. Research into the organization of information continues today and now we have an extensive set of tools to both create more information and to improve our access to this information. This section reviews some

[20] http://www.archaeoinformatics.org/

[21] Oxford Classical Dictionary, 1996.

research and development in the field of information science which is related to bibliographic work.

One useful way to organize resources is the technique of citation indexing. In 1955, Eugene Garfield published a small article in the journal *Science* where he argued in favor of developing a citation index for scientific literature. He argued that scientists were too often referencing data and ideas only from the early articles on a topic, and they were missing later articles that had refuted some portion of the original article. A citation index, by providing a comprehensive list of article cross-citing, was an efficient way to prevent anything from being overlooked. He was especially interested in the way scholars could follow the thread of an idea by following citations.[22] Writing at the beginning of the computer age, he did not yet have the best tools to accomplish the task. Since then, companies have developed such indices, though only for certain disciplines. Citation indexing is resource intensive, which means that it could only be accomplished by a collaborative online system. The tools for creating such an index in an open collaboration have not been developed yet, but it should be considered for an area of possible future enhancements. The organization CrossRef.org, which is funded by publishing companies, has created a database linking many recent articles using Document Object Identifiers (DOI). Hopefully, it would be possible to leverage this data even for archaeology articles.

Traditional citation indexing depends only upon following the references from the end of an article. The Internet enables a number of other ways to track the impact of

[22] Garfield, 1955, pp108.

an idea that is published online. As part of their relevance judgments, search engines like Google consider how often other webpages hyperlink to a particular page.[23] Links from different types of webpages can be judged in different ways. For instance, if an idea in a scholarly article is referred to by a mass media news agency, this may indicate that the idea is of interest to a public audience. Additionally, it is possible to track how many users visit a resource on the web. Web statistics can also provide information about a user's location and how they arrived at a website. For scholars, all of this information can give a better idea of an article's impact on the community.[24] The end goal is to map a knowledge domain, to make it easier for people who want to learn about that domain to gain knowledge.[25] An online collaborative bibliographic tool will greatly enable the ability to do this successfully because it acts as a filter of the best resources and it ties these together. The information contained in such a system is created by people, such as students, who are themselves seeking knowledge.

A tangential development in the distribution of scholarly works is the Open Access (OA) movement.[26] The technological developments of the last decade have lowered the barrier for publication on a global scale. Most scholarly journals now have online editions and that has led to the question in the minds some academics, why shouldn't anyone in the world have access to the latest scholarly developments? Private publishers still tend to charge a fee for access to online articles. Through subscription plans, most large research university members have access to these, but often through

[23] Brin and Page, 1998.

[24] Kleinberg, 2004.

[25] Shiffrin and Börner, 2004.

[26] Harnad et al, 2008.

an overly complex proxy server system when away from campus. To solve these problems, open access journals and repositories shift the cost of operation to authors, advertisers or libraries and provide free access to anyone with an internet connection. A 2001 editorial in the journal *Science* exalted the value of making all peer-reviewed research findings publically available.[27] A number of the signatories to this editorial went on to help found the Public Library of Science (PLoS), which publishes a number of OA biomedical journals online. Due to their wide availability, these journals have already achieved high citation index impact factor ratings.[28] In addition, Harvard University's faculty of Arts and Sciences and the Law School recently implemented a policy where copies of all published literature must be given to the library for placement in an OA repository.[29] OA is crucial for online bibliographic creation because when a link to a resource is posted online, it increases demand for that resource. Yet, some portion of that demand will probably come from people who do not have access to university library subscriptions.

Finally, online collaborative communities offer their own potential for information science research. It will be useful to study how scholars interact in a collaborative virtual community. There are a number of issues that might arise. The academic environment for a discipline like archaeology is focused on individuals and small groups. Promotions such as tenure decisions for professors depend a lot on publication. What happens if many authors from across the world, who may not know each other, develop an online resource together? How is credit distributed for

[27] Roberts et al, 2001.
[28] Patterson, 2008.
[29] Harvard Law School Press Release, 2008.

intellectual contributions? On a different topic, an online bibliographic resource built by multiple experts would be an excellent dataset for further analysis. By connecting resources together in an intelligent way, the bibliography presents an opportunity for machine learning. Alternatively, queries run against such a data set could discover connections not previously noticed. Perhaps a certain metal is used for the same tool at archaeological sites on different continents.

Evaluation of Online Collaborative Systems

One prevalent type of site on the internet now is the bookmaking site. There are two types of these, including personal and social. In the first category is a useful tool called http://www.spurl.net/, which allows users to capture links to websites as bookmarks, but from any computer since they are stored online. The second category offers similar functionality but combines the work of multiple users, thus it is called social bookmaking. One popular example is the site del.icio.us, which allows users to save website URLs in their profile. These links are organized using single word tags as descriptors, in what is referred to as a folksanomy.[30] Multiple users can tag and then add comments to the same link. This type of system is an excellent way to distribute the work it takes to organize and describe many potential resources. It can also help to indicate popularity if many people tag the same link and it keeps links up-to-date as valuable links continuously receive tags. As people assign similar tags to a link, the description becomes more accurate because the most important tags are emphasized,

[30] Terdiman, 2005.

especially when displayed in a “Tag Cloud.” Most of the social functions of social bookmaking sites would be useful to a Bibliographic system, and have already been included into existing systems such as Connotea, discussed below. The two key concepts of these types of sites are the distribution of labor required for organization and the ability for all users to work on a single outside source, a URL in this case.

Social bookmaking sites do not offer an inherently flexible data model, however. A user can do little more with a link than add tags and comments. Yet there is so much additional useful information that users want to connect to their links that workarounds have been developed by the community. The ability to associate an archaeological resource with the location of an excavation site is crucial for bibliography. This is sometimes referred to as geocoding. Social bookmaking sites use tags to store geographic information, in a process called geotagging. Thus, in del.icio.us a user can add three tags to each link they want to geotag: "geotagged”, “geo:lat=y”, “geo:long=x”, where x and y are the actual spherical coordinates on earth.[31] Every night, a script runs that exports all the geotagged bookmarks to Google Earth. This system limits the type of geographical data that can be attached to a bookmark to one or two coordinates. It also adds extra tags to each link that are not designed for a human user and thus impede comprehension. Furthermore, any desired change to the location data would need to be applied to every single link that had that tag.

[31] Torrone, 2005.

One of the main problems of social bookmaking sites is that they lack flexibility for users to provide information about the resources. On the other hand, systems such as Wikis allow all users full control over the information they can add to a webpage. In Wikipedia, information in the form of text, images, and audio, etc is used to describe a topic. Since all users collaboratively edit the same information, the ideal is that the best information appears in each article over time. In a bibliographic system, it would be useful for users to be able to add whatever information they want to a resource. For example, notes and quotes from the resource may improve others' ability to comprehend the resource before reading it. For archaeological sources, it would be useful to add other types of information, such as images, geographical locations and time periods.

An additional important aspect of wikis is their ability to record the history of modifications made to information in the system. Change history is a useful collaborative concept that developed within the software programming field over many decades. By associating each change with a user identifier, it is possible to know who was responsible for each modification. It also makes it possible to "roll back" selected changes in order to return the information to a correct state if a mistake was made. In a bibliographic system, the crucial information is the metadata of the citation. This data, after being entered the initial time is unlikely to undergo many subsequent modifications, so change management is less important. However, any additional information such as description, tags or subject headings, and reviews would be updated more frequently and thus could benefit from such a system.

Whereas Wikipedia uses an article structure, other websites allow the user to define structure for their information. By structuring data, queries can be run against the data and it can be reused in new ways. The Semantic Mediawiki project expands the wiki functionality to add additional information about the relationship between text articles. This system extends the technology of Wikipedia to implement the design ideas of the W3C's Semantic Web initiative. Practically speaking, this means that relationships between articles can be indicated at each link. For example, in an article about a city, a link to the article about the mayor can encode this semantic relationship with a phrase such as "is mayor of". Later, it is possible to automatically create a list of city mayors based on this encoded information. Data structure can also be defined using Ontologies. An advantage of this system is that it should not be much harder to use than a basic wiki, since there is only one additional step.

Other websites allow users to structure data using traditional database concepts. Thus, fields or columns can be defined to hold specific types of values. The fields are combined into useful rows or records, and records exist together in a table. Finally, records from different tables can be related to each other depending on fields. Freebase.com, created by Metaweb Technologies, Inc. is such a technology. It allows custom structures to be built and filled with data by anyone who registers for the site. There is also an API for pulling data from the database from external websites. According to one article, it is even possible to modify data through the APIs without visiting the home site.[32] There are however a number of potential downsides to Freebase.com. The openness of the project is easy to question. The data itself is

[32] Mattison, 2008.

released under a Creative Commons license that is similar to Wikipedia's, however unlike Wikipedia, the organization that runs the site is a for-profit business. Unlike Google, Metaweb does not have a long public record of action that can be evaluated for future intent. The initial implementation of their website did not allow public access even for reading data. The software itself is not open source, and neither are the APls for use from external sites. Even in his glowing review of Freebase, columnist David Mattison ominously states: "I could find virtually no reference in the Freebase help to technical aspects of the backend infrastructure."[33] Another disadvantage is the complexity of relational databases. In the same way that many users feel uncomfortable moving beyond Microsoft Excel to Access, it may be too much.

Evaluation of Current Bibliographic Software

Beyond these general web tools that allow users to create and share data online, there are also websites and software designed specifically for bibliography. Before reviewing these tools, it is useful to examine the most basic type of online bibliography – static webpages that simply list sources. These types of bibliographies inherit the weaknesses of the old paper bibliographies on which they are based. Take for example the Gordion Bibliography mentioned above, or another online bibliography of Old Testament sources called "Prophecy and Apocalyptic: An Annotated Bibliography." [34] This online bibliography is a supplement to a printed publication of the same name. It

[33] Mattison, 2008.
[34] Sandy and O'Hare, 2007.

includes annotations and is organized topically. Although the quality of the data is likely high given that the two authors are probably authorities on their topics, there are a number of downsides. Sources that could be linked to two or more topics are only listed under their most important topic. There is no facility for users to add sources, and thus dynamically grow and keep the list up to date. Sources are not related to each other beyond the topical category. The topical organization denies the user the ability to search and browse on any other relevant criteria.

Better bibliographies will be created by leveraging software to move beyond static html pages. Bibliographic software falls into a number of categories. The first point to consider is whether the software is used online through a webpage or must be downloaded and run locally. For collaborative bibliography, the first option is preferable so that anyone can use and view the citations quickly, so mostly this type will be reviewed here. The next point to consider is whether or not the software requires a fee for use. For many years, digital bibliography has been done using purchased software packages loaded onto a personal computer that interacted directly with word processing software. A major software product in this category was Endnote. These commercial products have migrated online, and new competitors have been created to provide this service, such as Refworks. Even the latest version of Microsoft Word has a built-in bibliographic management component. However, since proprietary software effectively divides the world into users who have access and those who do not, they are ill-suited for enabling collaborative bibliographies. A comprehensive list of reference management software options is available in a Wikipedia article and has been reproduced here in appendix A. This section provides a review of the major options of non-proprietary, web-based bibliographic software. These options are evaluated for their applicability to a collaborative bibliographic tool for archaeologists.

The first examples of reference management software worth reviewing are the open-source, self-hosted packages. Two prominent examples in this category are RefDB and refbase. RefDB was originally designed with a command-line interface and only recently has a web-based PHP interface been added. It was designed to integrate with structured documents such as XML. The RefDB project homepage does not maintain a list of links to existing installations, therefore it is difficult to evaluate this software. However, it does seem overly complex for academic work and it does not add useful functionality not found in other software packages. In contrast, refbase has been adopted by multiple projects. It offers a simple interface for adding, searching, browsing and exporting citation records. There are a large amount of metadata fields available to describe each resource. However, there is not a way to directly import data from an external website, at least some data needs to be copied and pasted into the refbase screen, which slows a researcher down. Furthermore, since it does not depend on a universally unique identifier for each resource, such as a URL or DOI, it is possible for users to accidently enter redundant citation records.

One of the main issues with tools like RefDB and refbase is that they are designed to be installed and maintained by the user community. Thus, it requires that a user have access to her own web and database servers. The user must also have the expertise to be able to install, configure, update and maintain such software. This may lead to other important maintenance issues such as dealing with security breaches and operating system patches. Thus, self-hosting these types of software can turn out to be quite time and resource intensive. On the other hand, the person maintaining the software has complete control over both their data and the functioning of their environment. It would be possible, for instance, to modify the html pages in order to customize the look and feel of the reference software. Refbase itself does maintain a generic install that can be used by the public, however it has not been as popular as some of the other centrally hosted websites discussed below. Both of these software packages are open source and appear to be developed by volunteers. This second point may put it at a disadvantage as compared with software with commercial backers since updates may not be as frequent. However, refbase was recently updated and it is probably the best reference management software package among these open source packages. On its development wiki, a list of future enhancements attempts to address some of the functional shortfalls of this software.[35]

Two of the more popular bibliographic websites are Connotea and CiteULike, which are both based on the social bookmaking concept discussed above. They expand this model by including metadata fields specific to published resources associated with each link. Since these sites understand the data model of a citation, a key advantage is their ability to import this data automatically from a resource. For example, if a user visits a journal article on a webpage which they want to catalog in Connotea, they click a JavaScript enabled browser bookmarklet. Using techniques described below, Connotea is able to extract information from the article such as title and author. With this data in its database, the bibliographic web tool is then able to export a list of

[35] http://wiki.refbase.net/index.php/Planned_feature_additions

citations for many potential uses. These could be formatted for copy and paste into a word processor or structured for export to other bibliographic software. Like other social bookmaking sites, all resources can be visible to all users and anyone can create tags for each resource.

These social bibliographic bookmaking sites do still have some serious drawbacks. Connotea does not yet have a way to display a citation formatted using any of the main academic citation styles, functionality which CiteULike has. On the other hand, CiteULike does not have a built-in wiki like Connotea. The wiki provides space for community members to communicate whatever information they choose. A tool of this flexibility could help archaeologists to develop a group of related resource links by setting standard tagging vocabulary. Another weakness of both sites is that, like other social bookmaking sites, the data model is inflexible. There are no extra fields that could be used to denote the time periods or geographical locations of archaeological sources. These values must be entered as tags, and thus their semantic importance cannot be differentiated from other tags. Connotea does recommend assigning geographical locations to sources using the same workaround as del.icio.us, and thus has the same limitations. One user has already requested a feature upgrade that would make it easier for resources to be integrated with Google Map. Another disadvantage of these sites is that they were designed for use with web links and journal articles, because their initial target audience was scientists. For archaeologists, printed books are still key resources that reference management systems must recognize. Both of these sites can capture book metadata in the article metadata fields. However this type of data is only automatically imported from Amazon.com. There is also no easy way to move from the resource link to library holdings information or a scanned copy of the book.

Another consideration with both Connotea and CiteULike are ownership issues. The first consideration is the software itself. Connotea has released all of its software through the open source process, using the sourceforge development website. CiteULike has not yet open sourced its software code. There are many advantages for a user of open source coded software applications.[36] Most important for a user of a bibliographic system is the possibility of continuity. Even if the website itself were to close down, the data could be exported to the software running on another server. This leads to the important question of who owns the actual data that a user has entered into a website. The user is expending effort to enter the data, so they should have an assurance that no one else will directly profit from their work. Sites like Wikipedia and Freebase.com make information ownership explicit by using a Creative Commons license or similar device. Finally, continuity of a reference management website is influenced by the stability of the organization that supports the site. For instance it is reassuring that the Nature Publishing Group is behind Connotea.

Finally, a very interesting piece of software in this category is Zotero. This is a plugin that adds functionality to the Firefox open source web browser that was developed by historians. It allows a user to capture a link to a webpage in a personal library. Like Connotea, it also can automatically extract metadata from each webpage and then create formatted citation lists with this metadata. However, since it was

[36] Wheeler, 2007.

developed by historians, it is designed to work better with printed materials such as books. It also has a folder system for organization of resources, in addition to tag functionality. Finally, it has good notetaking functionality and even allows a user to make notes and annotations directly on to locally cached copies of web pages. The biggest disadvantage of Zotero is that it was designed for individual use. Therefore the data it records is only stored on the local computer, which makes it difficult for a user to switch computers. More importantly, it prevents researchers from collaborating by sharing, discussing and organizing their resource links together. Naturally this problem has not gone unnoticed by the developers and the next two future releases will add server-side capabilities to Zotero. It will be interesting to see how they merge their current functionality with features similar to Connotea's. Other projects, such as SILS's NeoNote project, are also focusing effort on improving the server-side capabilities of Zotero.[37]

Data Entry and Import

In order to improve the efficiency of a bibliographic system, whenever possible metadata about a resource should not be entered by the user. Much of the important information about scholarly resources has been digitized by someone else. The information that must be gathered for a bibliography includes basic metadata about resources, such as title, author and publication date. The various types of resources each have different places online where this data can be found and reused. Monographs

[37] Hemminger et al, 2008.

and other paper books are important resources for archaeologists. Most books have already been cataloged by organizations like OCLC or the Library of Congress. OCLC's Worldcat website provides the metadata and library holdings about these books. The Internet Archive's Open Library project has amassed similar information, but its data is more freely available, such as through an API, and can be edited by anyone.

Metadata for journal articles is available from different sources. For example, citation information about articles in the biomedical field is found in the Medline database.[38] There are also a number of private companies that maintain article databases, but a fee may be required to access these databases. One system that has developed over the last ten years is the Document Object Identifier (DOI).[39] The identifiers provide a persistent link to the online copy of the article. They also provide metadata about the article that can be used in a citation. This metadata can be retrieved as XML using OpenURL queries, however this does require registration with CrossRef.org and the use of their schema. Unfortunately, publishers must purchase DOIs and thus the coverage may not be universal or continuous. For example, in one citation list I found a DOI link to an article in the *American Journal of Archaeology*. This link must have been valid at some point, but now it appears to have expired as it no longer resolves properly when entered into the official DOI website.

Web resources are easy to link to given the ubiquity of Uniform Resource Locators (URL). However, extracting metadata from webpages can be difficult. There

[38] U.S. National Library of Medicine, 2008.
[39] Rosenblatt, 1997.

are many potential standards webpage authors can use to supply metadata about their pages. However, the extra effort required to implement a standard usually dissuades a page creator from providing this type of data.[40] The Dublin Core is a standard that was designed to be simple to use and very flexible, to help website authors overcome these hurdles.[41] Yet this flexibility causes inconsistency in the data that is placed on webpages, and thus it often is not possible to use this data. The OpenURL ContextObject in SPAN (COinS) specification provides a means for bibliographic information to be encoded in a normal Html page. The wide adoption of the OpenURL standard will help improve the quality of COinS implementations. However, these standards are intended more for use with published journal articles and it remains to be seen if varied resources like blog posts will start to contain such data. Finally, it is possible for software to make an educated guess about the best way to extract bibliographic metadata from webpages. Search engines like Google have developed sophisticated algorithms to locate important meta-information about pages. Systems such as Zotero also scrape obviously useful information from pages - such as the page title.

Archaeologists deal with data and resources beyond text. Satellite photographs and raster images from non-visual instrumentation can provide valuable data about climate, vegetation and land-use. Photographs of architectural remains, the excavation process, ancient documents and artifacts all contain useful information. Audio such as podcasts and video are fast becoming important tools for the sharing of knowledge

[40] Thomas and Griffin, 1999.
[41] Apps, 2005.

among researchers. Perhaps most important of all are excavation datasets. These can be referenced at multiple levels of detail. For example, the data collected about an entire trench might be leveraged to bolster an argument, or the information about pottery shards collected from a single context. How are all these different types of resources referenced in a scholarly work so that other researchers are able to locate them and follow an author's thought process? How can these discreet resources be cataloged for discovery and reuse in a bibliographic management system? Information scientists have only just begun to consider these questions and modify existing standards or create new ones to deal with them.

A new project of the Open Archives Initiative aims to make it easier to maintain connections among scholarly works. This project is called Object Reuse and Exchange (ORE). By leveraging Semantic Web technologies such as RDF, the ORE provides a way to aggregate related resources.[42]

Archaeological Bibliography Features

This review of existing online collaboration and bibliographic software has provided a view into the current state of scholars’ ability to share information digitally. Future applications will be built upon this foundation of the current technology. It will be useful to articulate which existing features are useful to the specific field of archaeological bibliography, and what new features would be most useful to

[42] Lagoze et al, 2008.

archaeologists. There are a number of options for adding new functionality to existing systems, especially if they are open source. For example, one could combine existing software or functionality, such as in a “mash-up.”[43] Alternatively, it might be possible to request functionality from a developer community, or directly contribute source code if the project is open.

I see one of the main advantages of collaboration as the centralization of the space where people work. On the one hand, this has the distinct advantage of decreasing redundancy. Each citation is its own object, as are authors, publishing houses, archaeological sites and archaeological concepts. Duplicate data is less efficient to find because of effort wasted in identifying and eliminating duplicates. Another advantage of an online collaboration space is the uniformity of the user interface. This decreases the learning curve for users, but may have the unintended consequence that the best possible interface is not available. The system should also allow for shared editing of bibliographic records. As with Wikipedia, the amount of data available will grow quicker if resources are distributed as wide as possible. However, it might be useful to add an editorial layer on top of the system to maintain quality. Collaboration allows human agents to combine their expertise and their efforts into a whole. Thus, a centralized collaboration space for all scholars to share is essential for an archaeological bibliographical tool.

Another key feature is the availability of intuitive interfaces for searching and browsing each unique dimension of the resource metadata. The system should allow

[43] Pietroniro and Fichter, 2007.

organization, browsing and searching based on topics, resource type, geography, and time period. The solution should understand archaeological concepts that affect the way sources are used. It should be able to deal with the multiple types and formats of content used by archaeologists. Faceted classification search systems provide one good way to interact with data in different dimensions. For example, the Triangle Research Library Network is experimenting with the Endeca search interface for finding library books.

Control and ownership of the data is also an important consideration. Since systems are changing rapidly, the option of moving the data to a better system must be kept open. New features will improve the ability to enter and utilize data – but they may only be added to a competing system. Consider the case where a system requires external users to attribute the origin system for use of the data, even this might be a stumbling block to moving to another site. Yet, the citation metadata is not itself the original intellectual creation. Rather, the important contribution is the structure of the data, the organization of the citations and the relationships between them. In the same way, the system itself is not important. Ideally, the data creator would maintain ownership. Finally, the solution should be built with popular, open technologies that can be enhanced and maintained in the future by a large community of developers.

Initial implementation

The next step is to begin to implement an online bibliographic tool for archaeology. Experimentation will lead to ideas for new features or new ways to combine existing tools. More importantly, it provides the opportunity to begin to collect and organize the actual citation data. Even if the underlying systems change over time, the data itself will remain relevant. I will be working on my PhD in archaeology over the next few years. During my classwork and while I write my dissertation, I will obviously be constantly interacting with scholarly resources and writing papers that will require citation lists. By organizing these resources online, I will have the opportunity to share my work with other people and hopefully benefit from the work of others. The main focus of my archaeological research is ancient Turkey, so this initial implementation will only be scoped to include sources and topics from this region. The system should handle books, monographs, articles and web resources, but also have the ability to expand to other sources later.

Based on the evaluations of software done for this paper, I have decided to try two software packages as the core of the bibliographic system, Zotero and Connotea. Experimentation with Zotero will begin after they finish developing their online collaboration component. The initial implementation of this system is therefore based on Connotea. The main advantages of Connotea are that it is social, designed specifically for bibliography, open-source and it has an API that enables data access from other websites. The main disadvantages are that it does not have a robust geographical component, it does not import book metadata automatically, and there is not a quick way to do citation style formatting.

For the initial system, a few important books, articles and web sites were entered into Connotea. A tag prefix was then used as a way to identify the archaeological sites. For example, a book about the Gordion site in central Turkey was assigned this tag: "site:Gordion." This identification allows the resource to be associated with the geographical location of the site as well as other resources about the site. Resources can be queried by tag names by using Connotea's web API. Unique identification of ancient sites is a challenge because ancient names are rarely certain. Modern scholars may disagree about the identification of a place, or the old name may not be known at all. In the latter case, modern names are often used to identify a site. There is also the possibility of multiple sites having the same name. The Pleiades project is currently developing a system that will provide standard, unique identifies for ancient Classical places, and I hope to eventually leverage this in my bibliography.[44]

By mapping the resources on a Google Maps mash-up, the initial implementation works around Connotea's limited geocoding functionality. However, integrating geography into the initial system turned out to be the greatest challenge due to the lack of quality location data available about these ancient sites. Initially, I planned to pull coordinate data out of Wikipedia for each site. There are two major ways to retrieve geographical coordinates from Wikipedia. The first is DBpedia, a tool that uses semantic web technologies to structure the information found in Wikipedia.

[44] Elliott, 2008.

This structured data is then exposed for SQL-style database queries over the internet. A second service is GeoNames, which associates location names with coordinates. This website aggregates data from many sources including Wikipedia, and provides standard web APIs to query this data. I tried to use GeoNames to convert sites names to coordinates because it is simpler and it is designed for this purpose. However, many sites were not contained in its extensive dataset and names that did return matches often returned multiple, inconsistent coordinates. I was also unable to find a way to perform batch queries on multiple locations. I realized that I would need to supply the coordinate data to these systems myself. But instead of trying to populate unstructured Wikipedia pages, I entered coordinates for my test sites into Freebase.com. I pull these data based on the site names using a web API. These coordinates are placed on a Google Map and then associated with Connotea references based again on site name.

The resulting mash-up thus still requires manual data manipulation, but it is a good step forward. By leveraging the existing social tools Connotea and Freebase.com it will be possible for anyone to contribute to this bibliographic system.

Sources

Apps, Ann. "Guidelines for Encoding Bibliographic Citation Information in Dublin Core Metadata," Dublin Core Metadata Initiative Citation Working Group, 13 June 2005. <http://www.dublincore.org/documents/dc-citation-guidelines/>

Apps, A. and R. MacIntyre. "Why OpenURL?" *D-Lib Magazine*, v12 no5, May 2006.

Brin, S. and L. Page. "The Anatomy of a Large-Scale Hypertextual Web Search Engine," *Computer Networks and ISDN Systems*, v30 no1-7, 1998, pp107-117.

"Callimachus (3)," *The Oxford Classical Dictionary, Third edition*, edited by Simon Hornblower and Antony Spawforth, Oxford University Press: Oxford, New York, 1996.

Chenhall, Robert G. "The impact of computers on archaeological theory: An appraisal and projection," *Language Resources and Evaluation*, v3 i1, 1968, pp15-24.

Elliott, Tom. "Barrington Atlas IDs," *Horothesia Blog*, 10 July 2008. <http://horothesia.blogspot.com/2008/07/barrington-atlas-ids.html>

Garfield, Eugene. "Citation Indexes for Science: A New Dimension in Documentation through Association of Ideas," *Science*, v122 i3159, July 1955, pp108-11.

Harnad, Brody, Vallieres, Carr, Hitchcock, Gingras, Oppenheim, Hajjem, and Hilf. "The Access/Impact Problem and the Green and Gold Roads to Open Access: An Update," *Serials Review*, v34 i1, March 2008, pp36-40.

"Harvard Law faculty votes for 'open access' to scholarly articles," Harvard Law School Press Release, 7 May 2008. <http://www.law.harvard.edu/news/2008/05/07_openaccess.php>

Hemminger, Brad, Jie Jin, and Peiwen Zhu, "NeoNote: A User Interface for 'Memex'," YouTube video, 9 April 2008. <http://youtube.com/watch?v=PUn09--HRaw>

Hendry, David G., J. R. Jenkins, and Joseph F. Mccarthy. "Collaborative bibliography," *Information Processing & Management*, v42, May 2006, pp805-825.

Kleinberg, Jon. "Analysing the scientific literature in its online context," *Nature Web: Focus on Access to the Literature,* April 2004. <http://www.nature.com/nature/focus/accessdebate/18.html>

Lagoze, Van de Sompel, Johnston, Nelson, Sanderson, and Warner. "ORE User Guide – Primer," Open Archives Initiative, 11 July 2008. <http://www.openarchives.org/ore/primer>

Mattison, David. "The Freebase Experience," *Searcher*, v16 i2, February 2008.

Patterson, Mark. "2007 Impact factors for PLoS Journals," *PLoS Blog*, 18 June 2008. <http://www.plos.org/cms/node/366>

Pietroniro, Elise and Darlene Fichter. “Map Mashups and Rise of Amateur Cartographers and Map Makers,” *ACMLA Bulletin*, no127, University of Saskatchewan, 2007, pp26-30.

Roberts, Varmus, Ashburner, Brown, Eisen, Khosla, Kirschner, Nusse Scott, and Wold. “Building A ‘GenBank’ of the Published Literature,” *Science*, v291 no5512, 23 March 2001, pp2318-2319.

Rosenblatt, Bill. “The Digital Object Identifier: Solving the Dilemma of Copyright Protection Online,” *Journal of Electronic Publishing*, v3 no2, Ann Arbor, Michigan: Scholarly Publishing Office, University of Michigan University Library, December 1997. <http://hdl.handle.net/2027/spo.3336451.0003.204>

Sandy, D. Brent and Daniel O’Hare. *Prophecy and Apocalyptic: An Annotated Bibliography [Additional Bibliography]*, Institute for Biblical Research, 2007. <http://www.ibr-bbr.org/IBRStudies/Sandy-ApocalypticBiblio/ApocalypticBibliographSupplement_SandyOHare.aspx>

Shiffrin, R. and K. Börner. “Mapping knowledge domains,” *Proceedings of the National Academy of Sciences*, v101, 6 April 2004. <http://www.pnas.org/content/101/suppl.1/5183.full>

Terdiman, Daniel. “Folksonomies Tap People Power,” *Wired Magazine*, 1 February 2005. <http://www.wired.com/science/discoveries/news/2005/02/66456>

Thomas, C. and L. Griffin. "Who will create the Metadata for the Internet?" *First Monday*, v3 no12, December 1999. <http://www.firstmonday.dk/issues/issue3_12/thomas/index.html>

Torrone, Phillip. “HOW TO geotagg del.icio.us bookmarks,” *MAKE Magazine blog*, O'Reilly Media, 27 July 2005. <http://blog.makezine.com/archive/2005/07/how_to_geotagg_1.html>

U.S. National Library of Medicine, “Fact Sheet MEDLINE,” 22 April 2008. <http://www.nlm.nih.gov/pubs/factsheets/medline.html>

Wheeler, David. “Why Open Source Software / Free Software (OSS/FS, FLOSS, or FOSS)? Look at the Numbers!” Revised 16 April 2007. <http://www.dwheeler.com/oss_fs_why.html>

Appendix A: Comparison chart of reference management software from Wikipedia <http://en.wikipedia.org/wiki/Comparison_of_reference_management_software>

Software	**Developer**	**First public release**	**Latest stable version**	**Cost (USD)**	**Open source**	**License**	**Notes**
2collab	Elsevier	2007-11	?	Free	No	proprietary	centrally-hosted website, web-based
Aigaion	Aigaion developers	2005-01	2.0.2 (2008-03-11)	Free	Yes	GPL	web-based
BibDesk	BibDesk developers	2002-04	1.3.14 (2008-02)	Free	Yes	BSD	BibTeX front-end + repository
Biblioscape	CG Information	1998	7.19 (2007-11-15)	US$79-299[1]	No	Proprietary	ODBC; web access in Pro ed; optional client/server
BibSonomy	U. Kassel	2006-01	?	Free	No	proprietary	centrally-hosted website
Bibus	Bibus developers	2004-06-03	1.4.2 (2008-03)	Free	Yes	GPL	integrates with Word and OO.o Writer
CiteULike	Richard Cameron	2004-11	?	Free	No	proprietary	centrally-hosted website
Connotea	Nature Publishing Group	2004-12	1.7.1 (2006-02-01)	Free	Yes	GPL	centrally-hosted website, web-based
EndNote	Thomson Corporation	1988	X2	US$299.95[1]	No	proprietary	often used in academia
JabRef	JabRef developers	2003-11-29	2.3.1 (2007-11-29)	Free	Yes	GPL	Java BibTeX manager
Papers	Mekentosj	2007	1.6 (2007-	US$42	No	proprietary	search repositories

			09-06)				from interface; supports plug-ins
ProCite	Thomson Corporation	1984 ?	5.0.3	US$299.95[1]	No	proprietary	supports network access
Pybliographer	pybliographer developers	?	1.2.11 (2007-09-25)	Free	Yes	GPL	Python/GTK2
refbase	refbase developers	2003-06-03	0.9.0 (2006-10-23)	Free	Yes	GPL	web-based for institutional repositories/self-archiving[2]
RefDB	refdb developers	2001-04-25	0.9.9 (2007-11-05)	Free	Yes	GPL	network-transparent; XML/SGML bibliographies
Reference Manager	Thomson Corporation	1984	11.0.1	US$239.95[1]	No	proprietary	network version available; built-in web publishing tool
RefWorks	RefWorks	2001	2007-08	US$100 per year	No	proprietary	centrally-hosted website
Scholar's Aid	Scholar's Aid, Inc.	1998	4.1 (2008-4-1)	US$149[1] / Free Lite version	No	proprietary	integrates with Word and OpenOffice
Sente	Third Street Software, Inc.	2004	5.5 (2008-5)	US$129.95[1]	No	proprietary	integrates with Word, Mellel, Pages, and Nisus
Zotero	Center for History and New Media	2006-10-05	Beta 1.0.6 (16 June 2008)	Free	Yes	ECL	Firefox extension