Metadata Strategies for Cultural Heritage Information

Lewis R. Lancaster, University of the West

In 1997, a group of scholars meeting in Berkeley formed the Electronic Cultural Atlas Initiative (ECAI). Since that time, the group has grown to nearly 1000 affiliates and has held 18 conferences for the purpose of developing the strategies for archiving, display, and retrieval of cultural heritage information. At first, we discussed the use of static maps with TEI metadata categories embedded in the text data. However, neither of these procedures proved to be effective. First, having hundreds of static maps created the problem of sorting and retrieving. Second, the embedding of detailed metadata into text material required a great deal of time and expertise.

Over time, our discussions of these problems has led to a series of decisions and implementations. At the moment, the metadata needs for cultural heritage have been broken into four main categories: Where, When, Who, and What. Each one of these has been a major challenge and we have by no means reached a point of sufficient solutions for any of them.

The first category to occupy ECAI members was "Where." It was recognized that static maps, even though digital, would not be adequate for the large amount of data that comprises cultural heritage history. What was needed was a way to have any number of maps created to represent changing parameters. In order to do this, it was necessary to make use of the Geographic Information Systems (GIS) softwares. Fortunately, industry provides robust programs that operate on the basis of latitude and longitude designations. This has proved to be one of the easiest methods to categorize large amounts of information. Using the markup of latitude and longitude for any piece of text or image allows us to archive and retrieve that data by using a map rather than searching for words that have been used as titles

The second category of concern was "When." It is not enough for humanities to have the location on earth of a site. History involves temporal dimensions that define the concerns of the scholar in equal importance to space. With time included in the metadata, it would allow scholars to limit the search to a site such as Alexandria but for the years of the Roman era rather than all of the information for all time. Working with the University of Sydney, ECAI helped with the development of TimeMap software that permits metadata for latitude and longitude to be matched with the metadata for a specific time. Once the mapping area is set and the time bar determines the metadata category of the period under consideration, maps can be created with dots the represent available information. TimeMap has now been made open source and is freely available for users.

Having settled the basic concerns of metadata needs for time and place, ECAI members began to look for ways to automatically use these two elements with large sets of data. We do not yet have metadata TEI categories embedded in hundreds of thousands of pages of digital texts. Our search engines can retrieve words and phrases but cannot

tell us whether these words or phrase refer to a place, an object, or a personal name. With reference to the category of "Where" a problem arose when we sought to find the latitude and longitude for place names used in historical documents. Large on-line gazetteers exist that have thousands of place names tied to latitude and longitude but ancient names for sites that are no longer habitated are not included. This meant that we had to work with the issues of creating historical gazetteers for place names found in ancient texts. With a list of ancient place names, tied to latitude and longitude, it would be possible for scholars to search large digital documents and find places where those names in the list appear in the text. Such a search would not require embedding the markup in the data itself. The search would be referencing the gazetteer listing and finding similar strings within the documents.

At the same time as the automatic search for place names identified by comparison with a historical gazetteer listing, scholars in humanities will seek to narrow the retrieval by temporal settings. This is not as easy as first hoped because ancient texts do not use modern Common Era numerical designations for the time periods. The pattern of most textual data is the use of names for time periods. For example, " Parthian Empire", "Han Dynasty," "reign of Darius" may be the only temporal references. The creation of a list of named time periods is a lengthy process, especially if it requires scholars to scan, process, and identify thousands of such designations. This challenge was taken up by Professor Michael Buckland and Professor Larson at Berkeley. With a grant from the IMLS, they explored a method of using the Library of Congress Catalogue as the data source for named time periods. The results were quite promising, as hundreds of named periods emerged from the fields in the LC digital catalogue. Not only did they names appear but Common Era years are often found beside them. This has allowed the construction of a list of named time periods.

As the work proceeded, it was apparent that the category of "Who" could play a major role in the automatic search and identification of places and time frames. Part of this research was taken up by Harvard University and Fudan University in Shanghai. Under the project of the Chinese Historical Atlas, the researchers identified the boundaries of political divisions in China over the centuries. At the same time, they began the work of making a list of the government officials who were associated with the polygons of counties and prefectures. The places where officials were born, served, and died have been recorded. This will allow searches in the future that identify place and persons. When these two are put together, the accurate identification of both is increased. For example, finding a name such as Arlington can be enhanced if we find the personal name George Washington in the same paragraph. We can be more certain that the reference is to the Arlington in Virginia not an "Arlington" in another location. The Harvard project added another dimension to the ECAI efforts. We might call it "Who Else?" That is, they began the task of tracing the network of contacts between an official and the other people in his biographical information. When these contacts are also marked up for time and place, it is possible to have dynamic maps created that show the distribution of those who had a relationship to an official. Much more needs to be done to determine how "Who Else?" can be done in a more automatic fashion. It is under

study by Professor Buckland and will be a valuable extension to his current work on time periods.

Finally, we recognize that "What" is an important element in our study of culture. "What" implies that we need to carefully study events and define how we can track them in digital information. Already in the work of determining place, time, and persons, a whole range of events are implied in those categories. There are births, positions, deaths, exiles, reigns, and a whole host of happenings that are implied in the textual data referring to an individual or a place. If we can develop a search comparable to the one that Professor Buckland has done for named time periods, our efficient and effective use of information will be much improved.

Digital libraries with enormous amounts of data will be much more useful to all scholars when they can have search abilities to find items for Where, When, Who and What. These searches require semantic backups such as gazetteers, named time periods, biographical references, and methods for identifying events. Because of the size of data banks and the rapid expansion of them into the foreseeable future, we need to have machine search and retrieval as well as dynamic mapping and display of metadata and information. ECAI has worked on such problems for over eight years and yet there is still much to be done by people in the fields of the humanities. Some delegates at the last meeting of ECAI at the University of Hawaii, suggested that the next "W" to be added to our list should be "Why." The search for "Why" will require a great deal of careful thought and algorithms of analysis. The promise of the digital library is still exciting and we believe the potential can be achieved as scholars, librarians, and technicians work in collaboration.