2008

Where Keywords Fail: Using Metadata to Facilitate Digital Humanities Scholarship

Sheila A. Bair
Western Michigan University, sheila.bair@wmich.edu

Sharon Carlson
Western Michigan University, sharon.carlson@wmich.edu

Where Keywords Fail: Using Metadata to Facilitate Digital Humanities Scholarship

Importance of Primary Sources

The recent final Report of the Library of Congress Working Group on the Future of Bibliographic Control, *On the Record*, encourages libraries to "enhance access to rare, unique, and other special hidden materials," noting that "as educators seek to engage more students, including undergraduates, in research that utilizes primary sources, these materials are increasingly important for teaching and learning" (Library of Congress, 2008, p. 21). Humanities scholars have long agreed on the importance of studying primary sources. Presnell calls primary sources "the foundation of historical research" (Presnell, 2007, p. 92) and Williams notes that "[p]rimary and secondary sources are the crucial tools of the historian" (Williams, 2007, p. 56). A primary source can be defined as "a work that was written at a time that is contemporary or nearly contemporary with the period or subject being studied" (Cantor & Schneider, 1967, p. 22). They have also been described as "items that are directly associated with their producer or user and the time period in which they were created" (Presnell, 2007, p. 93). They are created by eyewitnesses to and participants in events and time periods and include sources such as diaries, journals, memoirs, newspaper accounts, and letters. A primary source may be understood in contrast to a secondary source which is an account of the past created by people writing about events after they occurred.

Introducing primary sources as part of curriculum-specific bibliographic instruction may be of benefit to students in several disciplines (Allen, 1999). The analysis and interpretation of these documents may be a valuable part of introducing critical

thinking skills to students at all levels, as research of primary sources requires considering questions of sources, quality of information, interpretation, implication, as well as questions of point of view (Robyns, 2001). Scholars at the 2008 OCLC symposium, Digitization and the Humanities, described how the digitization of primary sources has dramatically changed pedagogy and research and "unanimously called for expanded collection and digitization of primary sources" (Proffitt & Schaffner, 2008, p. 5).

Limitations of Keyword Searching

Mere digitization, even with full-text searchability, may not be enough, however, for digital humanities scholarship. Shawn Martin of the Text Creation Partnership at University of Michigan has noted that "sophisticated functionality is essential to new digital humanities scholars" and that "humanities scholars in the vanguard want to do research that cannot be done even with standard digitization." Martin references the Virtual Modernization project at Northwestern University where researchers are mapping variant spellings of words in early printed works to modern spellings in order to track their use through history. This type of word-level research, he notes, "requires materials much more accurate than Google would ever be likely to create," referring to Google's mass digitization efforts (Martin, 2007).

The many idiosyncrasies of primary source materials, including alternate spellings, abbreviations, obsolete and regional word usage, idioms, misspellings and omissions make full-text keyword searching difficult at best. This paper will show how the addition of controlled vocabularies for personal, corporate, and geographic names, and pre-

coordinated topic searches to transcribed and marked up primary texts increases their research value, provides searchability far beyond mere full-text keyword, and can facilitate scholar and student access to these materials.

Background of Project

In 2005 the Western Michigan University Libraries received a Digitization for Preservation and Access grant for Michigan libraries through a program sponsored by Michigan's Department of History, Arts and Libraries to create the *United States Civil War Collection* (Western Michigan University, 2006). Eight Civil War diaries, owned by the Western Michigan University Libraries and housed at the Archives and Regional History Collections, were selected, representing the experiences of seven Union soldiers and one contract employee. They presented a broad spectrum of experiences, with seven diarists having affiliations with Midwestern regiments. The remaining diarist, from Michigan, brought a particularly unique perspective as a contractual employee of the Quartermaster Department and included accounts of making and repairing carts and wagons.

The diaries had been used by scholars in the past but were difficult to access because of the handwriting. Nineteenth century handwriting includes specialized script for some letter combinations, such as words ending in two s's, that are not used today. Each diarist had different handwriting. Seven of the diaries included pre-printed spaces for entries and the minimal allotment sometimes resulted in text that was written very small or, when the diarist ran out of space, continued into margins and sideways up pages. The diaries written in pencil had become so faded as to make them nearly illegible. These

factors were further complicated with irregularities in spelling, abbreviations, and punctuation. A couple of the diaries also had serious preservation issues with fragile bindings. Generally, these diaries were not used by undergraduate students for these reasons.

Selection of the diaries for the project was predicated on the high level of interest of this period in history. According to the American Library Association's *Booklist*, in the 140 years since the United States Civil War ended, it continues to attract more public interest than perhaps any other event in American History. Well over 50,000 books have been written about the war since its conclusion in 1865 (American Library Association, 2001, p. 1006). James Geary's study of Civil War scholarship in the years after 1965 reveals that battles and campaigns, history of specific units, and works relating to "raising and maintaining and army" continue to make up large numbers of monographs (Geary, 1990, p. 19). Increasingly, scholars and enthusiasts have called for research into the non-military aspects of the war (Freehling, 2002).

Librarians on the project sought to standardize the names of people, places, battles, military units, and ships to benefit scholars and make the diaries available to students in new ways. The additional access points would also facilitate ease of use. As the diaries were gathered, transcribed, and marked up, personal, corporate, and geographic names were tagged and the authorized headings included in the markup. Additional markup, including 15 topic areas and dictionary definitions of obsolete words, was added to highlight other aspects of Civil War history.

Importance of Names

Historically, humanities scholars have found searches on personal names as subjects and geographical names very important to their research (Bates, Wilde & Siegfried, 1993). Flanders, Bauman, and Caton noted that researchers are interested in names mentioned in primary sources, and that "[t]he name can be a starting point for research on biographical, historical, or literary issues" (Flanders, Bauman & Caton, 1998, p. 285). For instance, scholars and students of the Civil War may be particularly interested in names of people, places, battles, military units, and maritime vessels mentioned in manuscripts. However, time period, regional language and spelling differences, way of life and education of the writer, and point of view can result in differences in terminology and naming. Presnell, for instance, notes that "while Americans refer to 'Queen Anne's War,' the British call this same military action the 'War of the Spanish Succession" (Presnell, 2007, p. 97). The inclusion of authorized name headings within the transcribed text, using TEI XML markup (Text Encoding Initiative, 2007), can provide the researcher with pre-coordinated browse lists of all names mentioned in the manuscripts, and can also improve the recall of keyword searches. Studies have shown that over one-third of keyword search hits would not be found if controlled subject vocabularies were not included in databases (Gross & Taylor, 2005; Norwick & Mering, 2003). The addition of controlled name headings provides a similar benefit. Normalizing names, by including an authorized name heading in the XML markup, enables specific searches, limits search results, and makes recall possible when names are misspelled, abbreviated, merely referred to, or unknown by the writer.
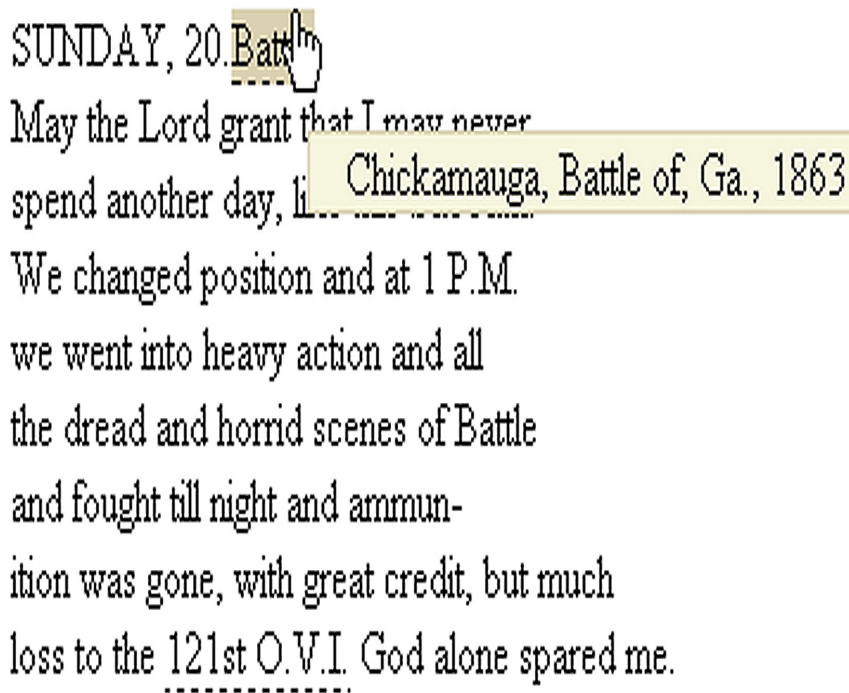
Benefits of Controlled Vocabularies

Inclusion of controlled vocabularies in the XML markup helps to disambiguate between names and commonly used words. For instance, the words cotton, hill, gray, wood, and cousin are also names of people and places in the diaries. A keyword search on the word "boots" will recall 25 hits, but only two of them are for James Boots of the 100th Illinois. The 107 hits recalled for a keyword search on the words "hill" or "hills" include ten mentions of Private George W. Hill of the 100th Illinois and two for General Daniel H. Hill. There are also mentions of Chapel Hill, North Carolina and the Muldraugh Hill escarpment. These specific references could only be found after laborious search or serendipitously without the inclusion of the authorized headings in the markup, as the men are referred to merely as "Hill" or "Hills" in nearly all of the passages, and Chapel Hill and Muldraugh Hill are spelled as Chapple Hill and Muldrah Hill.

Controlled vocabularies also enable discovery of new connections between the authors and recorded events – connections that may have taken years for researchers to discover, if ever. With one mouse click, for instance, the references across all eight diaries of a particular battle can be recalled. This is immensely helpful to students and teachers who may have the time limitations of a semester to complete their research. The inclusion of authorized headings for battles is important for recall as soldiers in the midst of, as yet, unnamed battles may relate valuable eye-witness accounts, but without the essential naming terminology needed for successful keyword searches. Alonzo Ide, of the 2nd Michigan, wrote "name of this Battle field as yet not Known" during the battle of Spotsylvania Courthouse.

Authorized headings in the XML markup can also reveal connections or shared experiences among the diarists. For instance, researchers can find that two soldiers fought

in the same battle or traveled on the same ship. Augustus Yenner, 121st Ohio, and

Eugene Sly of the 100th Illinois both fought at the battle of Chickamauga, September 20,

1863. Yenner recorded an entry the day of the battle (Fig. 1), while Sly remembered his

capture at the battle a year later from Andersonville Prison (Fig. 2). Sly notes, "Not over

1/3 of the Chicamauga prisoners are now alive." Researchers may never have discovered

this connection from the diaries in their original print format. The underlying markup of

the digitized version, which includes the Library of Congress authorized heading for the

battle, allows the searcher to find them both, though Yenner does not know the name of

the battle and Sly refers to the name of the battle as "Chicamauga." A keyword search on

the word "Chickamauga" would not have found either of these references.

Fig. 1. Mouse-over, pop-up function shows the authorized heading in the underlying

markup.

TUESDAY 20

One year today since I came
into this (So called) Confederacy
and as wonder would have
it am Still alive. I would have
thought it an impossibility to live one
year under such treatment as I have
received in the Confederacy. I can
See no better prospect of geting
exchanged this fall than there
was a year ago. Not over 1/3 of the
Chicamauga prisoners are now alive

Chickamauga, Battle of, Ga., 1863

Fig. 2.

People may be referred to or their names spelled differently across manuscripts

and even within the same source. Henry Wirz was the commander of Andersonville

Prison during the Civil War and was executed for war crimes when the war was over. In

his diary, Eugene Sly refers to Wirz as the "old Dutchman," "Old D," the "old Dutch

Capt.," and "Capt. Wirz." All references to Wirz can be easily found with one click on

his name in the browse menu or with a keyword search on Wirz because the underlying

XML markup includes the authorized heading (Fig. 3).

<p>Five tunnels were found in the pen today by the Officers where the prisoners were trying to get out. The **<name reg="“Wirz, Henry, 1823?-1865" type="person">**Old Dutchman**</name>** has sworn vengeance on us</p>

Fig. 3

Searches on military units and geographic places also benefit from the inclusion of controlled vocabularies. Throughout two diaries, the Michigan Engineers and Mechanics are referred to variously as "Engineers & mechanics," "M & Engeeners Regt.," "M & Engeeners," and "MichEng & Mach." The authorized, Library of Congress heading, "United States. Army. Michigan Engineers and Mechanics Regiment, 1st (1861-1865)," in the XML markup provides normalized spelling and a way to collocate all references with one search. Geographic names in the diaries frequently have alternate spellings, are abbreviated, or merely referred to with initials after the first reference. Authorized headings in the markup not only provide browse function and normalized keywords, but the mouse-over function in the database allows readers to view the place name (Fig. 4).

[page 164]
October, TUESDAY 24, 1865
cloudy & windy we
go a board the stern
wheel Steamer Clinton
10 am Start for Newburn
3Pm via the canal 3 Pm
run on a snag & was ship wrecked
the Steamer Alice
came along & took
us aboard we anchor
in C Sound at night
sleep on th Floor on the cabin
Currituck Sound (N.C.)

Fig. 4

Sources for Normalized Names

Many, if not most, of the names mentioned in primary sources will not have an
authorized heading created in a standard controlled vocabulary such as the Library of
Congress Authorities. Metadata creators, in collaboration with archivists and subject
specialists, must then choose a normalized form of the name. While several of the WMU
diaries referenced well known people, i.e., President Lincoln, Jefferson Davis, most of
the individuals named in the diaries were common soldiers or civilians. The diaries did
not include sufficient information in most instances to assign correct names to referenced
individuals. A number of online and print resources provided additional information,

including the United States Census, Civil War pension records, county histories, and regimental histories. Information, such as middle initials and birth and death dates, can be discovered and included to help disambiguate between similar personal names. Rules for formatting name headings can be found in the *Library of Congress Rule Interpretations (LCRIs)* and the *Anglo-American cataloguing rules, 2nd ed., 2002 revision.* Some of these guidelines can be found online at the NACO (Name Authority Cooperative Program of the PCC) website.

The diarists selected for this project included men who found themselves in surroundings that were unfamiliar and different. Often they had sketchy information about the place names and spelled them phonetically as they heard them. Sometimes it's hard to determine if a soldier who relates camping "near Beach grove" is referring to camping near a town called Beach (i.e. Beech) Grove, or camping near a grove of Beech trees. During the transcription and markup of the diaries, it was helpful to follow the progress of the soldier on a map. Maps created as near to the time period as possible were most useful as place names sometimes change over time, or towns become extinct or are subsumed by larger principalities.

There are many excellent Internet resources for researching Civil War era names, places, battles, and maritime history. The National Archives is a good source for personal name information and maintains a website offering guides to researchers. The National Parks Service also offers online information on people and places, including the *Civil War Soldiers & Sailors System*, which contains information on servicemen, regimental histories, and prisoner of war records. For some battlefields that are also national parks, such as Antietam and Gettysburg, the National Parks Service offers online information

about timelines, order of battle and people present. More information on Civil War battles can be found at The Civil War Sites Advisory Commission *CWSAC Battle Summaries* website. Information useful for creating authorized headings for Union and some Confederate ships can be found in the online Navy Department Library *ZC (Ship) Files in the Navy Department Library*. More on Confederate and Union ships can be found at Kenneth W. Jones' online *Index of Civil War Naval Forces Confederate and Union Ships*. Historical maps online can be found in the Library of Congress *American Memory Map Collections* and the *David Rumsey Map Collection*. Authorized headings for geographic place names may be found in the *Getty Thesaurus of Geographic Names Online* and the *Library of Congress Authorities*. These resources are listed with their links under "Metadata Creation Aids" at the end of this article.

Pre-Coordinated Topic Search

In addition to the traditional study of the military and political aspects of the Civil War through the use of primary sources, researchers have increasingly seen them as valuable windows into the social aspects of that period. Coinciding with the growth and interest in social history in the 1980s, historians called for the inclusion of social and cultural studies in Civil War history research (Vinovskis, 1989, p. 59). By 2002, more monographs began to explore the experiences of the common soldier, African-Americans, civilian life, and other topics of social history (Faust, 2004, p. 377). Librarians involved in this project have noted the increasing number of reference questions in the last decade about non-military aspects of Civil War history such as clothing, health, leisure, and religion.

Because of the interest in these topics, a decision was made to incorporate subject analysis at the word level in the XML markup. In collaboration with the University archivist, 15 topics of interest to scholars were chosen that could add intellectual content, assisting in keyword searching, and be the basis of a browseable pre-coordinated topic search function. In addition to the markup of names of people, places, battles, military units and maritime vessels, the following topics were also identified: African-Americans, clothing, death and casualties, desertion, food, health and medicine, leisure, money, music, religion, and transportation. An example of the markup for the topic "food" is shown in Fig. 5. Many of the selected topics occurred in the majority of the diaries but keyword searching was difficult for a variety of reasons.

Significant changes in the language and terms for African-Americans made finding references difficult.

<p>On the march A splendid fire of Cedar rails afforded us to make a
<seg type="food">cup of coffee and fry fresh mutton**</seg></p>
Fig. 5

Seven of the eight diaries reference African-Americans. Various terms across the diaries include "colored," "negroes," and numerous slang terms. While a seasoned historian may be aware of both the slang and socially accepted terms associated with African-Americans in the nineteenth century, students may not be aware of the changes in language in the last 140 years. Successful keyword searching may require a deeper understanding of the history of the Civil War and the meaning of terms that one may not

immediately associate with African-Americans. In the George Harrington diary, two references to "contraband" appear when searching for references to African-Americans (Fig. 6). An added note in the XML markup indicates that slaves that escaped to or were brought into the Union lines were referred to as contraband. Clicking on "African-Americans" in the topic browse and recalling references to "contraband" provides a learning tool for budding historians.

[page 108]

September, TUESDAY, 22, 1863.
left the C.H and went
across the Rapidan had
a fight too 30 Prisoners
took some Contraband[42]
Sent them back to
Maddison Camp across
the River

Fig. 6

Similarly five of the diarists reference leisure in sixty-seven entries, spanning a wide range of activities. On August 1, 1863, Augustus Yenner wrote about a Saturday involving both work and leisure (Fig. 7). In one passage alone, he writes about men writing, reading, doing handicrafts, and playing cards. Other diarists wrote about dancing, swimming, chess games, and baseball. Without these added access points, researchers

looking for references to leisure would be left to guess at the activities, how the diarist described them, and the keywords required for recall.

August, SATURDAY, 1.

No drill at all, nor Dress Parade
The skies were clear, we kept the shade
And spent time as suits the mind
writing, reading things refined
Washing, mending, playing cards
Making rings & plates of clams ^{shells}~~most~~ hard
Thus tediously though fast the time fly ~~does~~
While we for Home & peace do sigh
But no news came to make a change
Of which we think indeed most strange

Fig. 7

Conclusion

Primary sources, such as diaries and letters, are foundational to digital humanities research. Their study may be a valuable part of developing critical thinking skills in students. Enhancing access to these "hidden" resources through digitization is a valuable service to scholars, students, and educators. However, merely scanning and providing full-text keyword searchability may not fully meet the needs of digital humanities scholars. Abbreviations, obsolete and regional word usage, idioms, misspellings and alternate spellings, and omissions in primary sources make keyword searching, especially across many items in online collections, unproductive. The addition of metadata in the form of normalized name headings and topics can greatly enhance the research

experience and save the time of users, especially students inexperienced with possible terminology and spellings and who may have the time limitations of a semester to complete their studies.

Civil War era letters are now being added to the *United States Civil War Collection*. As the collection grows, the added functionality will become more important. Added controlled vocabularies and topic searches will continue to create the opportunity for students and researchers to make connections and discoveries they may otherwise have never been able to do.

References

1. Allen, Susan M. (1999). Rare books and the college library: Current practices in marrying undergraduates to special collections. *Rare Books and Manuscripts Librarianship, 13*(2), 110-119.

2. American Library Association. (2001). *Booklist, 97*(9/10), 1006.

3. Barrett, Andy. (2005). The information-seeking habits of graduate student researchers in the humanities. *The Journal of Academic Librarianship, 31,* 324-331.

4. Bates, Marcia J., Wilde, Deborah N., & Siegfried, Susan. (1993). An analysis of search terminology used by humanities scholars: The Getty online searching project report number 1. *The Library Quarterly, 63*, 1-39.

5. Cantor, Norman F. & Schneider, Richard I. (1967). The materials of history. In *How to study history.* (pp. 22-38). New York: Thomas Y. Crowell.

6. Faust, Drew. (2004). "We should grow too fond of it": Why we love the Civil War. *Civil War History, 50*(4), 368-383.

7. Flanders, Julia, Bauman, Syd & Caton, Paul. (1998). Names proper and improper: Applying the TEI to the classification of proper nouns. *Computers and the Humanities, 31*, 285-300.

8. Freehling, William. (2002). Why Civil War military history must be less than 85 percent military. *North & South, 5*, 14-24

9. Geary, James W. (1990). Post-centennial trends in American Civil War monographic literature, 1966-1990. *Collection Management, 14*(3/4), 3-14.

10. Gross, Tina & Taylor, Arlene. (2005). What have we got to lose? The effect of controlled vocabulary on keyword searching results. *College & Research Libraries, 66*, 212-230.

11. Library of Congress. (2008, January 9). *On the Record: Report of the Library of Congress Working Group on the Future of Bibliographic Control*. Retrieved June 21, 2008, from http://www.loc.gov/bibliographic-future/news/lcwg-ontherecord-jan08-final.pdf

12. Martin, Shawn. (2007). Digital scholarship and cyberinfrastructure in the humanities: Lessons from the Text Creation Partnership. Retrieved August 7, 2008 from, http://repository.upenn.edu/cgi/viewcontent.cgi?article=1058&context=library_papers Also available in, *The Journal of Electronic Publishing*, *10*. Retrieved June 21, 2008, from http://hdl.handle.net/2027/spo.3336451.0010.105

13. Norwick, Elaine A. & Mering, Margaret. (2003). Comparisons between Internet users' free-text queries and controlled vocabularies: A case study in water quality. *Technical Services Quarterly*, *21*(2), 15-32.

14. Presnell, Jenny L. (2007). The thrill of discovery: Primary sources. In *The information-literate historian: A guide to research for history students.* (pp. 92-135). New York: Oxford University Press.

15. Proffitt, Merrilee & Jennifer Schaffner. (2008). *The impact of digitizing special collections on teaching and scholarship: Reflections on a symposium about digitization and the humanities.* Dublin, Ohio: OCLC Programs & Research. Retrieved August 6, 2008, from http://www.oclc.org/programs/publications/reports/2008-04.pdf

16. Robyns, Marcus C. (2001). The archivist as educator: Integrating critical thinking skills into historical research methods instruction. *American Archivist, 64*(2), 363-384.

17. Text Encoding Initiative. (2007). *TEI guidelines.* Retrieved August 6, 2008, from http://www.tei-c.org/Guidelines/index.xml

18. Vinovskis, Maris A. (1989, June). Have social historians lost the Civil War? Some preliminary demographic speculations. *The Journal of American History, 76*(1), 34-58.

19. Western Michigan University. (2006). *United States Civil War Collection.* Retrieved June 24, 2008, from http://quod.lib.umich.edu/c/civilwar1/

20. Williams, Robert C. (2007). Sources and evidence. In *The historian's toolbox: A student's guide to the theory and craft of history* (2nd ed.). (pp. 56-78). Armonk, New York: M.E. Sharpe.

Metadata Creation Aids

1. Cartography Associates. *David Rumsey Map Collection.* Retrieved June 21, 2008, from http://www.davidrumsey.com/

2. The Getty. *Getty Thesaurus of Geographic Names Online.* Retrieved June 21, 2008, from http://www.getty.edu/research/conducting_research/vocabularies/tgn/

3. Joint Steering Committee for Revision of AACR. (2002). Headings for Personal Names. (chapt. 22). In *Anglo-American cataloguing rules, 2nd ed., 2002 revision*. Ottawa: Canadian Library Association; Chicago: American Library Association.

4. Jones, Kenneth W. *Index of Civil War Naval Forces Confederate and Union Ships.* Retrieved June 21, 2008, from http://www.tarleton.edu/%7Ekjones/navy.html

5. Library of Congress. *American Memory: Map Collections.* Retrieved June 21, 2008, from http://lcweb2.loc.gov/ammem/gmdhtml/gmdhome.html

6. Library of Congress. *Library of Congress Authorities.* Retrieved June 21, 2008, from http://authorities.loc.gov/

7. Library of Congress. (2007). Chapter 22: Personal names. In *Library of Congress Rule Interpretations (LCRIs).* Washington, D.C.: Cataloging Distribution Service, Library of Congress. Retrieved June 24, 2008, from http://www.loc.gov/catdir/pcc/naco/trainers/lcri22.pdf

8. Library of Congress. (2006). Chapter 23: Geographic names. In *Library of Congress Rule Interpretations (LCRIs).* Washington, D.C.: Cataloging Distribution Service, Library of Congress. Retrieved June 24, 2008, from http://www.loc.gov/catdir/pcc/naco/trainers/lcri23.pdf

9. Library of Congress. (2006). Chapter 24: Corporate names. In *Library of Congress Rule Interpretations (LCRIs).* Washington, D.C.: Cataloging Distribution Service, Library of Congress. Retrieved June 24, 2008, from http://www.loc.gov/catdir/pcc/naco/trainers/lcri24.pdf

10. Library of Congress. *NACO: Name Authority Cooperative Program of the PCC*.

Retrieved June 21, 2008, from http://www.loc.gov/catdir/pcc/naco/nacopara.html

11. National Archives. *Archives.gov*. Retrieved June 21, 2008, from

http://www.archives.gov/

12. National Park Service. *Antietam: History & Culture.* Retrieved June 21, 2008, from

http://www.nps.gov/anti/historyculture/index.htm

13. National Park Service. *Civil War Soldiers & Sailors System.* Retrieved June 21, 2008,

from http://www.itd.nps.gov/cwss/

14. National Park Service. *CWSAC Battle Summaries*. Retrieved June 21, 2008, from

http://www.nps.gov/history/hps/abpp/Battles/tvii.htm

15. National Park Service. *Gettysburg: History & Culture.* Retrieved June 21, 2008, from

http://www.nps.gov/gett/historyculture/index.htm

16. Navy Department Library. *ZC (Ship) Files in the Navy Department Library*.

Retrieved June 21, 2008, from http://www.history.navy.mil/library/guides/zcfilelist-

ab.htm