Marking Up Cultural Materials for Time and Geography

Fredric Gey, Ryan Shaw, Ray Larson,
Michael Buckland, Barry Pateman and Dan Melia

[1] University of California, School of Information,
South Hall, Berkeley, CA 94720-4600 USA
{gey,dmelia}@berkeley.edu {ryanshaw,ray,buckland}@ischool.berkeley.edu

Abstract. This paper describes two projects which fundamentally change access to cultural materials by automatically marking up textual materials for time, events, geography, and people. Because of the imperfect nature of named entity recognition within natural language processing, it is also important to allow for flexible, interactive markup of these materials by editors, scholars and compilers of cultural collections. Relating people in time and place may help to explore hitherto unrecognized relationships in history and social movements.

Keywords: biography, document markup, named entity recognition, Irish and Celtic studies, Emma Goldman

1 Introduction

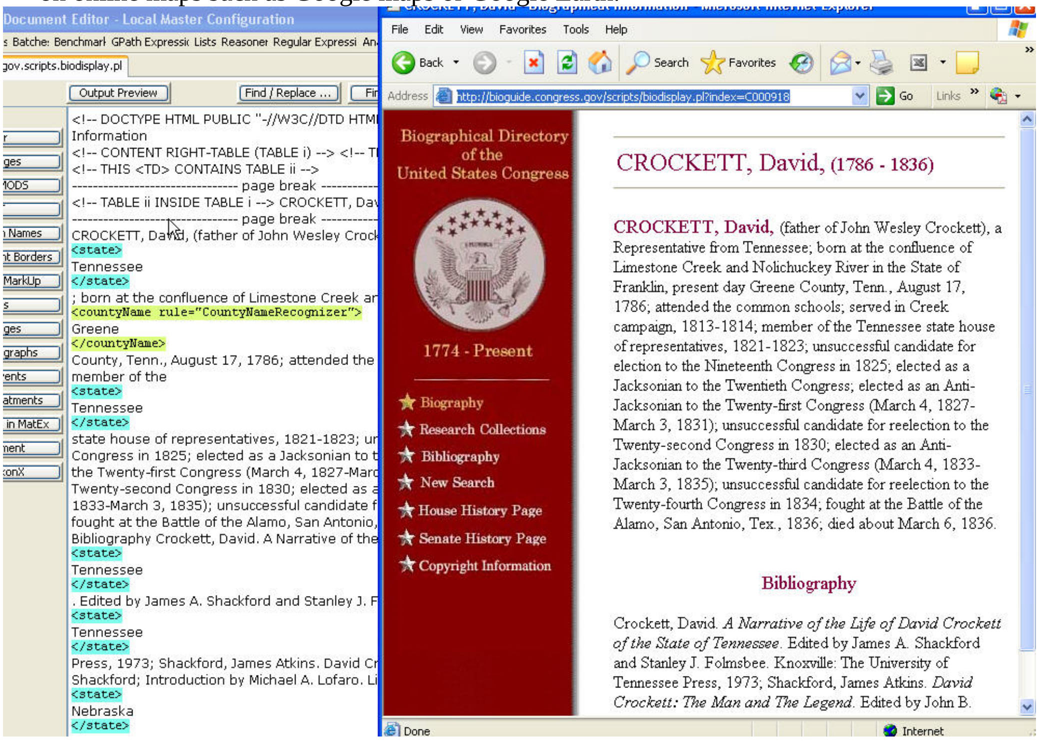
Cultural heritage, history, and social sciences are fundamentally about human activity. Everyone is interested in what other people do and have done. Biographies are regularly among the best-selling books. Not only History, but also Geography and most other subjects can come alive in the travelogues, journeys of discovery, and the life-stories of those involved. But mere narrative is not enough. *Understanding the context* differentiates education from memorizing. Building and supporting a community of learners needs more than facts. Understanding the *circumstances* of people's actions illuminates their lives. However, there is a significant gap in the infrastructure developed by libraries, museums, and publishers in this area. We have standards for the computerized handling people's *names*, but not for their *lives*. Take an example from American history, the US frontier hero Davy Crockett – how many people know that after serving as a Colonel under General Andrew Jackson, he was twice elected a congressman from Tennessee? These facts leads to other contextual questions: What was the population of Tennessee when Crockett was a congressman? Who were his contemporaries in Congress (e.g. Daniel Webster)? We will show mapping of biographies and historical events and their demographic contexts being developed in two closely-related projects: *Bringing Lives to Light: Biography in Context* and *Contexts and Relationships: Ireland and Irish Studies*. [1]

[1] http://ecai.org/imls2006, http://ecai.org/neh2007

2 Biography markup

Our presentation will include prototype interfaces and an “under-the-hood” look at pioneering XML event markup standard proposals which facilitate the demonstration. We have developed named entity markup grammars using the Gate technology [1] and, in particular, the Golden Gate XML editor from Karlsruhe [2] to automatically markup biographies for time and place. The work is similar in spirit to that of Smith and Crane [3]. Figure 1 below shows parts of two screens – to the right is the official USA congressional biography for Davey Crockett from the Congressional Research Service web site.

To the left in the figure is an XML version of the biographical text which has automatically marked up state and county names (counties are political sub-jurisdictions of states in the USA) using named entity recognition. Combining such markups with lookup from a gazetteer such as the Geonames server,[2] will add latitude and longitude to the marked up data, enabling locating parts of a personal biography on online maps such as Google maps or Google Earth.



Document Editor - Local Master Configuration
s Batche: Benchmarl GPath Expressic Lists Reasoner Regular Expressi An
gov.scripts.biodisplay.pl
Output Preview
Find / Replace ...
Fir
r
ges
MODS
n Names
nt Borders
MarkUp
s
ges
graphs
ents
atments
in MatEx
nent
onX
File Edit View Favorites Tools Help
Back
Search
Favorites
Address
Go
Links
Done
Internet

```
<!-- DOCTYPE HTML PUBLIC "-//W3C//DTD HTM
Information
<!-- CONTENT RIGHT-TABLE (TABLE i) --> <!-- T
<!-- THIS <TD> CONTAINS TABLE ii -->
------------------------------- page break ----------
<!-- TABLE ii INSIDE TABLE i --> CROCKETT, Dav
------------------------------- page break ----------
CROCKETT, David, (father of John Wesley Crock
<state>
Tennessee
</state>
; born at the confluence of Limestone Creek ar
<countyName rule="CountyNameRecognizer">
Greene
</countyName>
County, Tenn., August 17, 1786; attended the
member of the
<state>
Tennessee
</state>
state house of representatives, 1821-1823; ur
Congress in 1825; elected as a Jacksonian to t
the Twenty-first Congress (March 4, 1827-Marc
Twenty-second Congress in 1830; elected as a
1833-March 3, 1835); unsuccessful candidate f
fought at the Battle of the Alamo, San Antonio,
Bibliography Crockett, David. A Narrative of the
<state>
Tennessee
</state>
. Edited by James A. Shackford and Stanley J. F
<state>
Tennessee
</state>
Press, 1973; Shackford, James Atkins. David Cr
Shackford; Introduction by Michael A. Lofaro. Li
<state>
Nebraska
</state>
```

http://bioguide.congress.gov/scripts/biodisplay.pl?index=C000918

Biographical Directory of the United States Congress

1774 - Present

★ Biography
★ Research Collections
★ Bibliography
★ New Search
★ House History Page
★ Senate History Page
★ Copyright Information

CROCKETT, David, (1786 - 1836)

CROCKETT, David, (father of John Wesley Crockett), a Representative from Tennessee; born at the confluence of Limestone Creek and Nolichuckey River in the State of Franklin, present day Greene County, Tenn., August 17, 1786; attended the common schools; served in Creek campaign, 1813-1814; member of the Tennessee state house of representatives, 1821-1823; unsuccessful candidate for election to the Nineteenth Congress in 1825; elected as a Jacksonian to the Twentieth Congress; elected as an Anti-Jacksonian to the Twenty-first Congress (March 4, 1827-March 3, 1831); unsuccessful candidate for reelection to the Twenty-second Congress in 1830; elected as an Anti-Jacksonian to the Twenty-third Congress (March 4, 1833-March 3, 1835); unsuccessful candidate for reelection to the Twenty-fourth Congress in 1834; fought at the Battle of the Alamo, San Antonio, Tex., 1836; died about March 6, 1836.

Bibliography

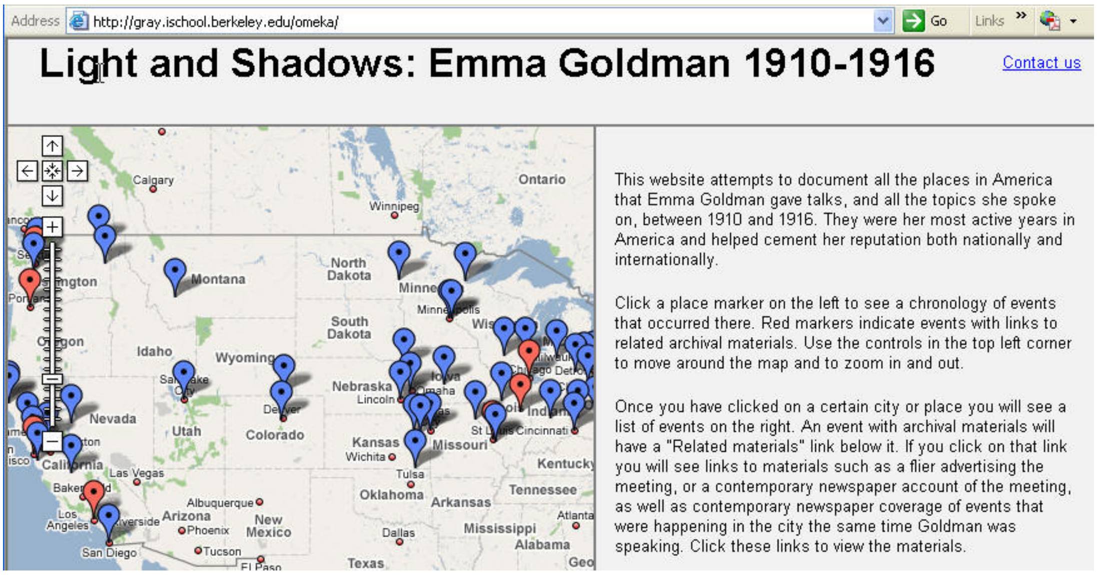
Crockett, David. *A Narrative of the Life of David Crockett of the State of Tennessee*. Edited by James A. Shackford and Stanley J. Folmsbee. Knoxville: The University of Tennessee Press, 1973; Shackford, James Atkins. *David Crockett: The Man and The Legend*. Edited by John B.

Figure 1: Split screenshots of a biography page and its XML markup of geographic names

[2] http://www.geonames.org/

3 Markup for Time and Geography

In addition to working with USA congressional biographies, we have followed the timeline of the great anarchist thinker and lecturer Emma Goldman as she toured the USA from 1910 to 1916, lecturing about drama, status of women, and anarchists such as Francisco Ferrer y Guardia, the Spanish-Catalan anarchist. Developing a timeline marked up with geography using the XML Atom format[3] allows us to input the data file directly into Google Maps and provide a geographic user interface to the Goldman travels and lectures. Working from a 5-page textual narrative of the Goldman travels provided by our partner, the Emma Goldman papers project,[4] the application of a Gate grammar marked up over 900 personal names of individuals who co-lectured or were otherwise associated with Emma Goldman. Figure 2 provides a screenshot of the resulting interface.[5]



Light and Shadows: Emma Goldman 1910-1916

This website attempts to document all the places in America that Emma Goldman gave talks, and all the topics she spoke on, between 1910 and 1916. They were her most active years in America and helped cement her reputation both nationally and internationally.

Click a place marker on the left to see a chronology of events that occurred there. Red markers indicate events with links to related archival materials. Use the controls in the top left corner to move around the map and to zoom in and out.

Once you have clicked on a certain city or place you will see a list of events on the right. An event with archival materials will have a "Related materials" link below it. If you click on that link you will see links to materials such as a flier advertising the meeting, or a contemporary newspaper account of the meeting, as well as contemporary newspaper coverage of events that were happening in the city the same time Goldman was speaking. Click these links to view the materials.

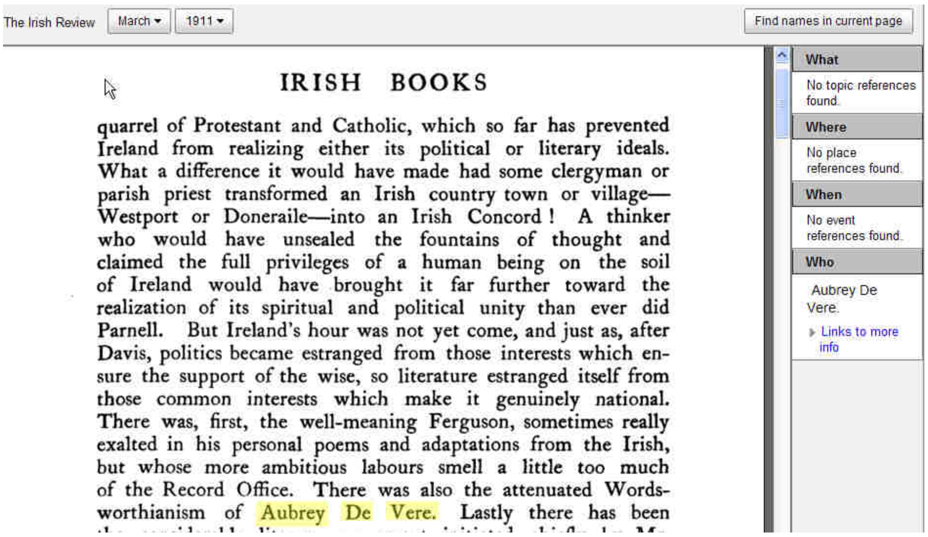
Figure 2: Automatically create a mapping interface to a biographical timeline

4 Markup Options for Newly Digitized Scholarly Materials

Scholarly materials are currently being digitized at a prodigious rate in the areas of digital humanities. Advanced digitization technology offers the opportunity for interactive interfaces to enable knowledgeable scholars and editors to supply their

[3] http://en.wikipedia.org/wiki/Atom
[4] http://sunsite.berkeley.edu/goldman/
[5] http://gray.ischool.berkeley.edu/omeka/

own markup which can then be interfaced to external related contextual resources. Our collaborative work with the Centre for Data Digitisation and Analysis at Queens University, Belfast has made available a large number of scanned documents on Irish history, culture, and scholarship. Because the digital technology can supply horizontal and vertical coordinates for each word in the scanned text, we can allow editors and scholars (and users of the materials) the opportunity to highlight segments of text and identify them as names or places (for example). Figure 3 shows a screenshot where a page from The Irish Review for March 1911 has a user-highlighted segment to identify the Irish poet Aubrey De Vere. Links are automatically made to external resources which related to this person.



IRISH BOOKS

quarrel of Protestant and Catholic, which so far has prevented Ireland from realizing either its political or literary ideals. What a difference it would have made had some clergyman or parish priest transformed an Irish country town or village—Westport or Doneraile—into an Irish Concord ! A thinker who would have unsealed the fountains of thought and claimed the full privileges of a human being on the soil of Ireland would have brought it far further toward the realization of its spiritual and political unity than ever did Parnell. But Ireland's hour was not yet come, and just as, after Davis, politics became estranged from those interests which ensure the support of the wise, so literature estranged itself from those common interests which make it genuinely national. There was, first, the well-meaning Ferguson, sometimes really exalted in his personal poems and adaptations from the Irish, but whose more ambitious labours smell a little too much of the Record Office. There was also the attenuated Wordsworthianism of Aubrey De Vere. Lastly there has been

What
No topic references found.

Where
No place references found.

When
No event references found.

Who
Aubrey De Vere.
▸ Links to more info

Figure 3: Manual editorial markup of scanned scholarly materials for a personal name

Note that the viewer can add markup on four facets: what, where, when and who, not just where and when (geography and time).

References

1. Cunningham, H, Maynard, D, Bontcheva, K, Tablan, V: GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*, Philadelphia, US.
2. Sautter, G, Padberg, K, Tichy, W: Empirical Evaluation of Semi-Automated XML Annotation of Text Documents with the GoldenGATE Editor. In: L Kovács, N Fuhr, C Meghini (eds.) ECDL 2007. LNCS, vol. 4675, pp. 357--367. Springer, Heidelberg (2007).
3. Smith, D, Crane, G: Disambiguating Geographic Names in a Historical Digital Library. In P Constantopoulos, I Sølvberg (eds.). ECDL 2001. LNCS, vol.2163, pp. 127 – 136. Springer, Heidelberg (2001)