Advanced Information Delivery System for the Abraham Lincoln Writings

Michal Ondrejcek, Rob Kooper, and Peter Bajcsy
National Center for Supercomputing Applications, University of Illinois, Urbana, IL 61801
ondrejce@illinois.edu, kooper@ncsa.uiuc.edu, pbajcsy@ncsa.uiuc.edu

Abstract

This work presents a virtual observatory of the Abraham Lincoln writings. Many of the Lincoln correspondence papers and legal documents have been digitized, pre-processed and annotated for public viewing. However, there are challenges in delivering the information to the general public due to the large number of documents, very large volume of digital files, lack of metadata, heterogeneous pieces of information and need for services such as searching and transcription. We have prototyped an advanced information system that delivers Lincoln writings together with historical maps, contemporary Google maps, and Lincoln Log metadata, and other types of information to provide a multi-dimensional view of Lincoln's life. The novelty of our work is in designing the system that provides such a unique multi-dimensional view of Lincoln's life by combining technologies needed for building virtual observatories.

Introduction

With the upcoming bicentennial celebration of the birth of Abraham Lincoln in 2009, our work is motivated by delivering the information about Lincoln's life to scientific and educational communities. Many Lincoln documents have already been studied and made available to the public through books, monographs, and initiatives, some of which are available online [1-4]. The existing virtual spaces accessible via the Internet usually do not provide a comprehensive view of the fast growing amounts of digital information about Lincoln's life. Our objective is to integrate heterogeneous data sources in a virtual observatory and provide access to temporal, spatial and contextual dimensions of the underlying large volume of data.

The overall project to digitize, store, and make publicly available all Lincoln writings is a joint effort of multiple institutions, including NCSA, the Illinois Historic Preservation Agency, and the Abraham Lincoln Presidential Library and Museum. As part of this project, the design of the virtual observatory has two major objectives. The scientific objective reflects our broader interest in manipulation and image processing of terabytes of data. The educational objective aims at making the documents accessible to the general public, students and scholars through the web-based interface. The advanced information delivery system supports data browsing, text query-based searching, geospatial data retrieval and visualization, and transcription services for transcribing image scans of handwriting to text.

Technical overview

Image scans: The original documents represent a large collection of the incoming and outgoing correspondence of Abraham Lincoln. The average size of one scanned Lincoln paper is about 150 MB. Currently, there are about 39,000 scanned documents (5.9TB) in the repository. Unfortunately, not all of them can be made accessible to the general public due to the restrictions imposed by the donors. In general, documents from the National Archives can be studied without any major restrictions, as in the case of the documents from the Library of Congress that are cross-linked from our website.

Image cropping: In order to preserve the color scale, brightness, and contrast, the sheets are scanned together with the color scale bar. The color bars and the background areas have to be removed before the web-based dissemination. Thus, we have designed an image-cropping algorithm using four classification schemes with and without training. The details of the image-cropping algorithm are described in [5].

Data management: Metadata about the Lincoln writings were obtained from The Lincoln Log [6] and The Papers of Abraham Lincoln in Springfield, Illinois. We have extracted, cleaned, and inserted the metadata into a MySQL database. The metadata correspond to the timeline, places visited by Abraham Lincoln, latitude and longitude of each place visited by Lincoln, etc. The design of the database supports any future expansions to accommodate new fields and expected 200,000+ pages.

Information delivery: The multidimensional web-based interface is based on the Google Maps API for visualization of the documents path and places relevant to Lincoln's life. Additional features are historical maps that can be overlaid on Google Maps, selected songs of the period, and examples of Lincoln's legal work. The front end consists of an HTML file with Google Map loaded, a search form and pre-defined data sets, and a JavaScript script. The client-side HTML and JavaScript files make requests to the server. The server-side consists of PHP files and MySQL database. The result is returned as an XML response to the Ajax engine. A user can search the database, view the results, view individual documents, edit or transcribe them, delete existing postings, or add new ones.

Historical maps: Maps have to be processed in order to overlay them on the Google Maps. It can be a tedious process since the geodetic coordinate system (a datum, a projection, an origin, a unit System and two axes) is not always available. Google Maps uses WGS84, Mercator projection and a pixel unit system [7]. Most of the historical maps of the United States are in (American) polyconic [8] or in Molweide pseudocylindrical projections, later in conic, Lambert conformal conic and Albers equal area projections. In our case, the projection transformation does not have to be exact. For small areas resembling Molweide projection, for example a simple perspective correction was sufficient.

Multi-dimensional view of the data

The graphical design of the virtual observatory features multi-dimensional view of the heterogeneous data. The dimensions are temporal, spatial, and content-related. The temporal (years 1809-1865) information is being retrieved from the Lincoln Log and the spatial information appears within the map itself in the form of markers and lines. A user can search for and edit documents. In the case of letters, he/she can visualize its path in the Google map pane. Additional information includes period music and links to selected legal cases of Abraham Lincoln. One of a shortcoming of the Observatory prototype is uneven representation of the data. There are ten of thousands of scanned documents compare to only few maps, songs and links to the legal cases. However, the problem is the lack of data, rather than the design.

Summary

We have built a prototype of a virtual observatory for the Lincoln writings. The scanned documents and corresponding metadata have been pre-processed and the queries of the data content have been visualized using multi-dimensional web-based interface. The underlying architecture is based on the Google Map API and Database Ajax/Javascript requests using PHP and MySQL. The target audience for the virtual observatory is the general public, students and scholars. This work is part of the University of Illinois Bicentennial Celebration project [9].

Acknowledgment

The authors would like to thank the UIUC Provost Office, Instituto Technologico de Costa Rica, the Goolge Summer of Code, the NSF TeraGrid program and to NCSA for providing funding and resources.

References

POC: Michal Ondrejcek, ondrejce@illinois.edu

[1] Library of Congress. Abraham Lincoln Papers at the Library of Congress. 2008.
URL: http://memory.loc.gov/ammem/alhtml/malhome.html
[2] Abraham Lincoln Presidential Library. 2008. URL: http://www.alplm.org/home.html
[3] The Papers of Abraham Lincoln. 2008.
URL: http: //www.papersofabrahamlincoln.org/
[4] The Collected Works of Abraham Lincoln. University of Michigan. 2006, URL: http://quod.lib.umich.edu/l/lincoln/
[5] M. Casares, K. Hard and P. Bajcsy, Image analyses of very large size collections of scanned documents, NCSA PSP 2007; URL: http://isda.ncsa.uiuc.edu/publications.html
[6] A life chronology compiled by the Lincoln Sesquicentennial Commission;
URL: http://www.thelincolnlog.org/view
[7] The details can be found in Google Map API documentation; URL: http://code.google.com/apis/maps/documentation/index.html
[8] John P. Snyder, Map Projections: A Working Manual, Professional Paper Report Number 1395, Publisher: U.S. Geological Survey, United States Government Printing Office, Washington, D.C 1982
[9] URL: http://www.uillinois.edu/lincoln/resources.cfm.
The prototype is available at URL:
http://isda.ncsa.uiuc.edu/lpapers/search.html