Boston Streets: Mapping Directory Data
Eighteen-month report
IMLS Project Grant #LG-02-02-0072-02
Submitted by: Gregory C. Colati, Project Director and Jessica M. Branco, Project Coordinator
June 24, 2004

New directions and new energies were the watchwords for the past several months. We finally solved the vexing problem of how to automate the linking of cataloged images, addresses listed in the directories, and geospatial points, and then how to display this information at the appropriate locations on the historical and modern maps of Boston included in the project. This was the key relationship that we hoped to demonstrate in the project and we are very excited about the potential of the results. By regularizing and manipulating the data using regular expressions and PERL scripts, instead of hand-encoding, harvesting, and plotting each piece of information on the maps by hand, workflow has become efficient, accurate, and highly automated. In addition, the long-awaited release of the newest version of the Tufts Digital Library system, sitting on Tufts' FEDORA repository has lead to the directories being searchable and viewable as text documents, as well as being datasets for the mapping tool. These new directions in the data preparation and display areas have created a new momentum carrying forward the project, project staff, and interested end-users.

With the ability to join together, filter and deliver sizeable amounts of information on demand, we can show that tying things to geographic points can lead to increased knowledge and the ability to discover information about something even if you know nothing else but the spot on the earth. Building layers of context around a point, a picture, or a person benefits the cataloger, the researcher, and the casual user. The key to building these relationships - like all data driven projects - is in regularization and development of centralized authority lists that are human-understandable and accessible. This truth illustrates the continuation of proven library and archival practices into the digital environment.

To create the linkages from one digital form of collection materials to the next, we began with previously created authority lists, studied their construction, augmented them with additional terms to facilitate connections, and then used them as the source of regularization models for the geographic address point data collected from historic atlas plates, images, and addresses listed in the city directories. The source authority lists for street and place names resulted from the detailed cataloguing work The Bostonian Society's Library and Special Collections has applied to their visual works collections which are included in the project.

The editors of the city directories were more concerned with fitting things into limited space than they were with consistency, understanding that, given the right context, a human could decipher even arcane and irregular abbreviations. As we well know, what a human does easily, a computer finds impossible to do without significant intervention.

Project staff used a combination of regular and replacement expressions in a text editor to find recognizable street names in the directories and regularize them to the standard, i.e. 36 Washington regularized as 36 Washington Street. Other passes through the data were done to regularize typical abbreviations, such as C'bridge to Cambridge Street. Batch processing of city directories also incorporated these authority terms into PERL scripts, allowing for the regularizations to happen quickly and in an objective manner. Current in-house standards for the city directories require that 90% of all personal and organizational entries in the TEI XML-encoded include a name and at least one valid address, either commercial or residential. The regularization of attribute values to authority terms within the address element (tag) allows for linkages and querying using modern, unabbreviated formats, while the digital library patron reading the directory continues to see the original format of the addresses as presented by the 19th and early 20th century editors and printers.

Once addresses were given regularized values, copies of the XML documents were "exploded" using another PERL script to sort directory entries into complete, acceptable, and additional batches and to harvest the desired attribute values from elements used to encode personal and organizational names, occupations, and addresses. The purpose of these lists is two-fold - they isolate and allow for the joining of commercial or residential addresses with geographic points taken from the historical atlas, and they provide a means of measuring the percentage of directory entries which have been successfully encoded. One list created in this process identifies which entries needs attention, aiding project staff in refining the encoding processes at the directory and project levels.

Once isolated, an address from a commercial listing, residential listing, or image's metadata can be connected, using Microsoft Access, with corresponding X and Y coordinates for the geographic point for that address taken from the atlases. This new database file is exported from MS Access, and is used to populate a database of geospatial information for the project and also imported into the ESRI software suite and transformed into a shapefile of symbols, points, and tables, for use in geospatial visualization and mapping tools.

The geospatial mapping tool developed for the Boston Streets project has been dubbed "Cowpaths," after the urban legend that the streets of Boston were laid out over the paths cows used to walk when the city was young and now accounts for the general disorder and "crookedness". Cowpaths allows users to define their own data layers, query the data and have their results mapped on any number of historical maps. This allows users to ask a whole range of questions. Please see example experiences below:

Example 1:

Cowpaths: plotting occupations
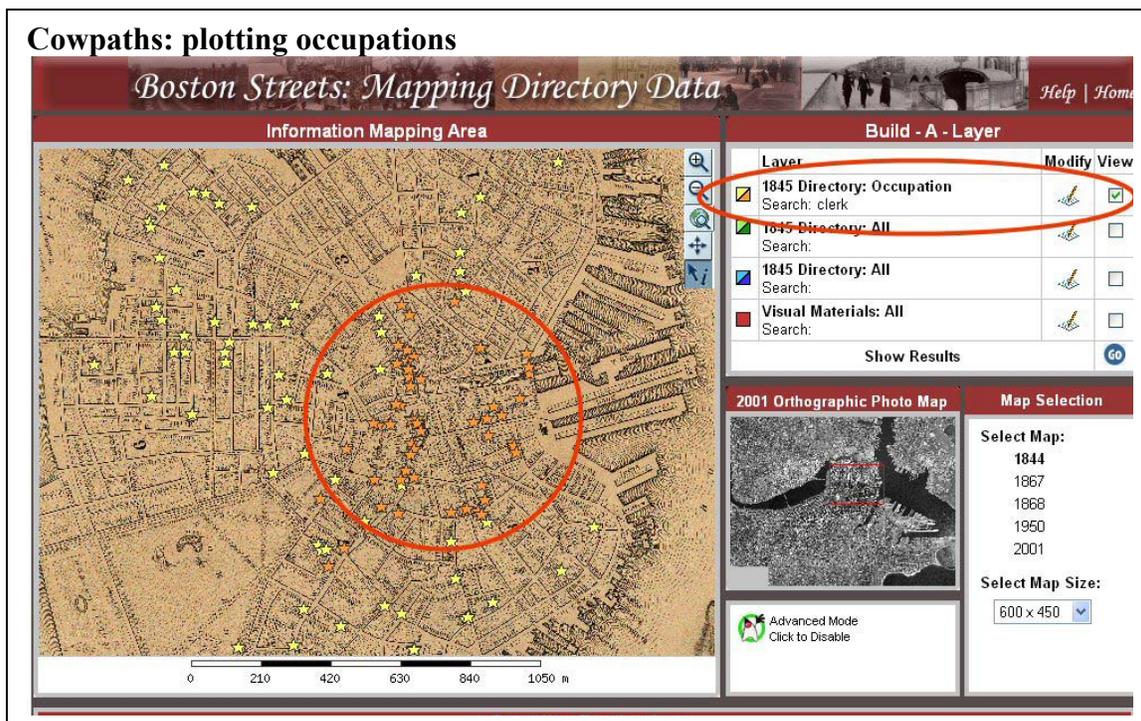
In the first example, the 1844 Dickerson map is used as a base, and the 1845 city directory data has been filtered for "Occupation." "Clerk" was the word used in the search of occupations. Dark stars indicate commercial addresses, while light colored stars indicate residential addresses of persons with the occupation of clerk. As you can see from the distribution, in general, clerks worked in the commercial center of the city and lived in the surrounding neighborhoods.

Example 2:

In this next example, the 1925 directory, again filtered for Occupation and searched for "clerk" is overlaid upon the 1950 city map. This time however, the Visual Materials layer is also visible. It is also possible to use the "Visual Materials" layer to plot photographs from the collections onto the maps and associate them with directory information about that same place.

The queried point in the center of the screen has both a star and a box, indicating that there is both a photograph of this point and information about the address.

Code	Full_Name	Occupation	Commercial_Address	Residential_Address	Date
R.	Rawstron, Stanley R	clerk	67 Milk Street	97 Pinckney Street	1925
R.	Rawstron James jr	clerk	67 Milk Street	97 Pinckney Street	1925

The Information Display Area is now active and shows results from the two layers. There is one image (from an earlier period) associated with this spot, and two clerks who live in one of the houses in the picture. Their commercial address is listed as well, giving another visualization of the relationship between working address and home address. In this case the two clerks have the same last name and work together at 67 Milk Street, deep in the commercial center of the city.

The image thumbnail is linked the same photograph in the Tufts Digital Library (TDL) collections, pictured below:

Tufts Digital Library: image display

Clicking on the preview image thumbnail in Cowpaths executes a search in the TDL and provides full metadata for the image, including an extensive description about the photograph provided by The Bostonian Society.

Cowpaths builds on information from various sources to contextualize a single piece of information using a series of linkages made possible via the text and metadata regularization processes. By linking these unifying, identical terms, it becomes possible to visualize on a map and in one or more photographs the lives of real people who lived and worked in the buildings pictured here, and throughout Boston. Cowpaths, in conjunction with the TDL also links the images to the people who worked in and inhabited the city. The TDL catalog provides searching and linking to a host of other information resources, including full-text searching of the directories, metadata searching of the photographs and other, more traditional ways to discover information. Using the two resources simultaneously allow the patron to instantly explore the history of a person or place in new ways.

The programming behind the Cowpaths tool was done by Peter Wilkerson, of Asheville, North Carolina. Peter was brought onto the project using DCA funds to help us work through address linking difficulties, visualization tool construction, and to help create a more generally understandable user interface. Additional programming support was provided by Robert Dockins, a graduate student in Computer Science at Tufts.

Cowpaths and the entire Boston Streets project are scheduled to be the centerpiece of a session at the annual meeting of the Society of American Archivists in Boston this August. We expect that Cowpaths, the documents in the TDL, and the larger web site of supporting and contextualizing information will be available to the public by that date. We will then continue to refine the user interface and begin testing the interface with both students at Tufts and patrons at the Bostonian Society in order to measure its effectiveness in both an academic and public history settings, as well as for the reference services provider and the general user.