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

As is often the case at the mid-point of these sorts of projects, a combination of unforeseen circumstances and new issues raised by ongoing grant work has required some reworking of the project plan. Meanwhile, additional financial support from other funding sources has enabled us to digitize more directories than we originally expected. The implementation of the Fedora repository system at Tufts in winter 2003-spring 2004 and the conversion from TEI P3 SGML to TEI P4 XML standards required a revision of procedures and reworking of already processed documents. Although these events set us back from our original timeline, the resulting content will be consistent with currently evolving digital library standards and best practices as adopted at Tufts and other institutions. We felt that the time spent bringing the data up to these standards was well worth the effort involved. We also discovered that there were few guidelines for tagging data intensive texts and we elected to develop a tagging scheme within the TEI P4 standard and at the same time develop automated text mark up procedures which will be documented on our web site and in future reports.

With additional financial support from DOCEMA, the DCA has secured funding for conversion to digital format of seven additional city directories (1845, 1849-50, 1855, 1860, 1870, 1872, and 1872). These directories have been digitized, with the A-Z entry sections having undergone double-key data entry. DOCEMA sponsored directories (which are also being used on the DOCEMA website, available at URL: http://boston1872.com/design/index.html) increase the total number of directories included in the project at this time to twelve, increasing by two the original number of directories to be included in the project, and offering a broader temporal coverage for the project.

Eleven of the directories are presently being structured with elements from the TEI P4 XML Guidelines. Digital Divide Data continues work on the text conversion for the 1945 city directory. Adopting the P4 XML guidelines mid-project has led to a revision of existing procedures designed for TEI P3 SGML elements and a period of conversion from a number of the directories already tagged prior to the transition. Using a number of regular expressions, macros and global search and replace techniques, a considerable amount of the tagging work has been systematically automated.
The well-formed XML documents and accompanying METS metadata is currently scheduled to be added to the Tufts Fedora system and made accessible via Tufts Digital Library in February 2004. In the meantime, development versions of the documents are available, with limited functionality from the existing Boston Streets web portal at URL: http://yngve.lib.tufts.edu/BostonStreets.

A revised interim user interface is also available at the site, anticipating future interaction points using FEDORA disseminators this spring. Research is currently underway looking at methods of providing "search-on-tag" functionality and other search and browse features. An additional 600 historic images of Boston from The Bostonian Society's collections have been added to Tufts Digital Library over the last six months. Templates are being designed to output METS files for the 1751 images currently included in Boston Streets. This information is required for migration of the images to the FEDORA system.

In addition to eight overview maps, 103 plates from Hopkins and Bromley atlases of the Boston and its surroundings (1874, 1898, and 1928) have undergone conservation efforts, Mylar encasements, and digitization. Tiff images captured from these plates are being geo-referenced and projected into appropriate coordinate systems while geographic point data for each address identified on the maps is being harvested and used to generate GIS shape files. This vector and raster data is being used to model the geo-spatial information and access tool component of the project. A functioning prototype of this system is scheduled for completion in April 2004, with revisions continuing through out the summer based on focus group feed back and testing. An online survey/feedback tool is also being designed for inclusion in the new web resource made available for external review this spring.

Developing a system to automatically generate connections from one set of data to another and the ability to visualize that information on a geographic space, in an attempt to further contextualize the material topically and temporally, has been identified as the primary research area of the project. Project team members are also researching methods of linking coordinate data to individual directory entry addresses and geographic terms found in images. Determining the levels of accuracy and scalability using a variety of methods, including database joins and manual plotting also continues to be a focus or research. We will attempt to find a level of automation and accuracy that provides usable results that are scaleable and sustainable within the context of our resources.

Following a well-received presentation in April to members of Boston Library Consortium at Brandeis University, project staff also showcased the project's development and archival considerations associated with the GIS data and structured documents at the Fall meeting of the New England Archivists at the University of Connecticut and the Digital Resources for the Humanities conference at the University of Gloucestershire, Cheltenham, England, UK.