The Historical Survey Database

by Jeffrey Remling

A thesis submitted in partial fulfillment
of the requirements for the degree of Master of Science in the
Department of Computer Science
Pace University
May, 2005
Thesis written by

Jeffrey J. Remling

B.A. East Stroudsburg University 1993

M.A. East Stroudsburg University 1996

M.S. Pace University 2005

Approved by

_______________________________, Thesis Chairman

_______________________________, Member, Advisory Committee

_______________________________, Member, Advisory Committee
# Table of Contents

Table of Contents.................................................................................................................. iii

List of Illustrations................................................................................................................ iv

Introduction: The Historical Survey Database................................................................. 1

Background: Computers and History................................................................................. 3

The Data......................................................................................................................... 6

The Entity – Relationship Model..................................................................................... 10

The Data Entry................................................................................................................ 13

Tweaking the Data........................................................................................................... 19

Geographical Information System.................................................................................. 20

The Research................................................................................................................... 25

Bibliography.................................................................................................................... 41
List of Illustrations

Image 1, page 485 of Longworth's 1840-41 New York City Directory ................. 8
The person table ...................................................................................... 11
The address table .................................................................................... 11
The ethnicity table ................................................................................... 12
The widow table ...................................................................................... 12
The wife table ......................................................................................... 12
The misc table ......................................................................................... 13
Image 2, page 485 after run through OCR program ..................................... 15
Image 3, the same text file corrected and delimited ...................................... 16
Image 4. The tgr36061blk00 base map ...................................................... 22
Map 1. Location of blacksmith shops in New York City, 1840 ...................... 27
Map 2. Blacksmith shops, coal merchants and coal yards as listed in Longworth's 1840-41 NYC Directory ........................................................... 29
Map 3. Locations of address marked with a 'd' on the far right side of the directory's page ......................................................................................... 31
Map 4. Locations of address marked with a 'dd' on the far right side of the directory's page ..................................................................................... 32
Map 5. Locations of address marked with a 'b' (red circles), 'x' (blue hexagons) and 'nr'(green triangles) on the far right side of the directory's page ................................................. 32
Map 6. Locations of address marked with a 'u' on the far right side of the directory's page ......................................................................................... 33
Map 7. Location of banks in New York City in 1840 .................................... 34
Map 8. Location of wool dealers (red pentagons) and cotton dealers (blue triangles) New York City in 1840 ........................................................................... 34
Map 9. Location of drygoods dealers New York City in 1840..................35
Map 10. Public houses, taverns and bars in NYC in 1840.....................37
Map 11. Residences of shipmasters in NYC in 1840..........................38
Map 12. Offices (green) and residences (purple) of attorneys in NYC in 1840.....39

The Historical Survey Database

Introduction: The Historical Survey Database
This project is an attempt at an interdisciplinary approach to merge computer science and history. Although other attempts have been made before, they usually omit the historian's absolute need to be able to have a citation back to the primary source, and to be able to enter a near real time research-interpretation-research cycle.

The Historical Survey Database (HSD) incorporates all of the entries in Longworth's Directory of New York City, 1840-41 into multiple tables of a MySQL database. The directory was a listing of over 35,000 inhabitants and over 45,000 addresses in Manhattan, and included information such as occupation, marital status, and ethnicity.

The database can be queried using standard SQL commands. The HSD also allows the data to be displayed not only in quantitative table forms, but also as a geographic information system (GIS) on an 1840 map of New York City. For this reason, it is often necessary to store the query results in a separate table for future processing by the GIS program (ArcView).

In time, multiple other sources of information, for example additional directories, census data, department of health records, and banking records, will be added to the database, creating a mountain of information which can then be data-mined.

There is a seeming divide between those who have a proficiency in the field of science and those who are proficient in the artistic fields. This divide is only seeming because those who are good in the sciences tend to be well versed in the arts as well, but concentrate their professional careers in the more financially profitable sectors.
They tend to use the arts for their entertainment and personal growth. Thus, they may be well read in history and art, yet lack professional training and experience in the actual skills needed to create works of art or write histories.

By contrast, those who are seemingly good in the arts might also have great abilities in math and science. History seems to indicate that Leonardo de Vinci was an artist above all else; even his scientific notebooks are works of art. More often, thinking of artists of various media living and working in contemporary New York City, it is hard to think of one who does not incorporate math or science into his/her work. Much art is based upon mathematical proportion, and concepts such as fractals are finding their own niche in the art world. Surely the New York skyline itself would look vastly different if architects, who tend to view themselves as artists, did not have a solid understanding of math and physics.

Yet although in both fields there are a number of people who appreciate interdisciplinary approaches – and perhaps it tends to be the brightest, or at least most successful, people in both fields who incorporate interdisciplinary approaches into their work – they are not the majority. Far more often then not, an individual may state that they have chosen one specific field in life because "I'm not good in math," or "the science is too hard." It is possible that these people might have brains geared towards their chosen field, they might be more motivated to contribute to society in these ways, or they simply might just be bad at the sciences and sticking to what they can get by with.

Disputes over why this rift exists are beyond the scope of this work; it is enough to acknowledge that a certain rift does seem to exist. This rift creates a
problem. The liberal arts centered individuals have a hard time understanding how the possibilities and limitations of science and technology can contribute to their field, while the science centered population lacks the experience and skills needed to see artistic ventures carried through. This thesis project, tentatively named the *Historical Survey Database*, is an interdisciplinary tool designed to bridge the rift between history and the scientific field(s) of computer technology.

**Background: Computers and History**

From the author's point of view, it is odd that the two fields of history and computer science are not utilized together more often. Historians are taught three skills – researching data, interpretation of the data, and presentation of the interpretation, usually in written form. Computer scientists are essentially concerned with methods of storing data, interpreting data quickly and presenting data easier. When placed in this light, a potential symbiotic relationship seems obvious.

This is not to imply that no historian has ever attempted to use computer technology before. It is just that computer technology is vastly underutilized in the field of history, and when it is used, it seems to be in one of three limited forms.

The first main use of computers in history was an explosion of the sub-field of quantitative history for a brief period in 1972-1973. The concept was simple enough. By adding large tables of data into mainframe computers and then applying statistical algorithms, certain underlying patterns should emerge. In October 2003, Dr. Roy Rosenzweig of George Mason University described this period to me, when he was in graduate school. He stated that he remembered standing in line with a stack of punch
cards that he had made, containing data from every one out of two hundred names from a nineteenth century directory of New York City, and loading them into a mainframe. According to Dr. Rosenzweig, this was a period when quantitative history was an exciting field, because the promise of computing power was about to open a new door, and possibly point to definitive histories. Unfortunately, no new discernable results appeared. Dr. Rosenzweig concluded that after about a year or eighteen months, the promise evaporated, and quantitative history was relegated back to small-scale studies of easily obtainable, and easier analyzed, data.

The reason why no new major patterns were discovered are probably varied, but probably include the limited number of records that could have been added into a mainframe at the time due to the size restrictions of memory, the lack of many statistical algorithms that have come to light in that past 30 years, the limited processing speed and costs of the mainframes, and, perhaps most damming, a possible misunderstanding of the data itself. As will be shown in the next section, the Longworth's Directory of New York City is only a ten percent sampling of the city's population.

The second main use of computers in history is an attempt to make primary sources held in libraries, archives, and museums, more accessible to the public. This generally means some form of digitization project, in which the document is at least scanned (what the author calls image-based digitization), and/or converted into a text

---

1 In history, a primary source in an actual historical document, such as a copy of the Declaration of Independence, or far more commonly, private letters, journals, orders, etc. A secondary source is a document, often a book or article, based upon primary sources, but reflecting its author's interpretation of the primary sources it is based upon. Valid secondary sources must cite their primary sources, thus creating a paper trail back to the original documents.
file (what the author calls text-based digitization) and the digitized file then made available on the internet.

The author has two main criticisms of digitization projects. First, the text-based digitization files seldom, if ever, refer back to an image of the primary source to allow the researcher to verify what is in the text file. Second, there is no one master guide for digitization projects, meaning researchers have difficulty in finding these records placed on the web, and also the institutions performing the digitization projects run the risk of wasting resources if two or more institutions digitized the same documents at the same time.

The third, and last, use of computers in history is to create a new media for publication. The afore-mentioned Dr. Rosenzweig is the director of the Center for History and New Media at George Mason University, the only known institution which gives PhDs in history that include a focus on technology. The New Media Center at the CUNY Graduate Center is also doing some excellent work in this field.

The problem with new media publications is that they seldom contain citations back to the primary source. Further, there is no way to guarantee that the website will remain up years into the future, and therefore websites should not be cited in future works. Any contribution they make to history must therefore be suspect and cannot be built upon.

It should be noted that all methods historians have attempted to use computer systems mentioned above are designed to help with researching specific questions, the first of the historian's skills mentioned above, or to aid in publication, the third skill. Interpretation seems to have been left alone. Yet in practice, historical
interpretation is linked to research in a cycle. Some fact is found, an interpretation is presented, and additional facts are sought after to confirm or rebuke the interpretation.

What is therefore needed is a vast, preferably *all-inclusive database*,\(^2\) which will allow researchers to search through the data in any feasible way. The results should be available in a reasonable time, allowing the results to be queried again and again, to perpetuate the research-interpretation cycle. Multiple forms of output should be allowed, both to aid the researcher in interpretation, and to facilitate the final publication. Finally, all results must be able to be traced back to their primary source, thus allowing further scholarship to be built on the work. This is the idea behind the *Historical Survey Database*.

### The Data

The Historical Survey Database takes *Longworth's Directory* of New York City for the year between 1840-41\(^3\), and places the textual content into a MySQL database (server 4.1). This directory contains listings of over 35,000 residents of New York\(^4\) and over 43,000 addresses. Additional information included in this source may include the resident's occupation, distinctions between home address and working address, ethnicity, and marital status.

The exact nature of how the publisher assembled and compiled his data is not known. However, it must be acknowledged that the directory itself must only be

---

\(^2\) By "all-inclusive database" I mean one single database or network of databases containing ALL historical records. All nations, all times, all fields.

\(^3\) In New York, during the 19th Century, leases traditionally ended on May 1\(^\text{st}\), also known as "moving day." Therefore, the dates of this directory start on May 1, 1840, and last until April 30, 1841.

\(^4\) It should be noted that the term "New York City" in 1840 consisted only of the borough of Manhattan, and not the outer boroughs.
considered a sampling of the New York’s population. Contemporary estimates show New York City’s population as 312,710$^5$ in 1840, meaning the sampling found in the directory represents a little more than ten percent (10%) of the actual population.

Exactly how this sampling was achieved is in itself unclear although some expected causes of exclusion seem evident. For example, the term "col'd" was included in 73 entries, out of the 35,696 entries listed in the directory. It seems highly unlikely that the black population of Manhattan was only 73 at this time, especially since eight "African" churches were listed. It is far more likely that the racism of the day excluded the vast majority of blacks from being included. Further, 19th Century racism tended to dump all non-whites into the classification of "colored," meaning that of the 73 listed a possibility exists that some might be from Southeast Asia, Native American, or members of any of a number of non-Caucasian and non-African ethnic groups. New York in 1840 was the nation’s largest seaport and it seems probable that Polynesians, Hawaiians and others spent at least some time living there.

---

This lack of representation was also not just limited to non-whites. Archeologist Diana Dizerega Wall has commented to the author in December, 2004, that when using the same directory for her dissertation's research, she found the population of the city's Five-points district, then one of the most notorious neighborhoods in the Western Hemisphere, also lacked representation. Although data
from the HSD has not yet confirmed this lack in the Five-points district, it is probable that the city's poor, more precisely the city's unskilled labor, were not adequately represented in the directory.

Aside from being a biased sampling, the data has other drawbacks. In many cases, entries are missing information. Often first and middle names are missing, or just represented by the first initial. Occupations and addresses are often missing as well. In other cases, part of an entry exists, such as a street number with no corresponding street name. Since the type the book was printed with was hand set, undoubtedly some entries are mistaken due to manual errors such as a '6' being put in upside down or a '3' being used instead of an '8' or a '5'.

In addition to missing data, non-standard abbreviations also have the potential for misinterpretation. In obvious attempts to save time in type setting and save space in the book, the publisher and his typesetters often abbreviated occupations and street names. "Broadway," for example, was often spelled out. More likely, however, it was abbreviated such as "Broadw." or "B'way." Naturally, these abbreviations had to be removed and the entire word(s) entered in the database wherever possible, to allow for a greater ease in text searching.

From the standpoint of a historian, these drawbacks in the data exist in most sources of the period, and have to be accounted for. Further, although a certain percentage of the entries may be inaccurate, on the whole the source is still a wealth of information to be mined.

From a computer science standpoint, the data has a few other benefits. For one, it is static. No "new" listings will be added into the 1840-41 directory, nor will
any be deleted. The static nature of the data eliminates the need to update the
MySQL database, making coding functions easier. Since the tables will not be added
to, considerations such as semaphores and other applications for locking the database
prior to altering it need not be addressed.

It should be noted at this point that the data entry into the HSD is by no means
complete. Ideally many additional directories will be added, as will tax records,
currently held at the NYC Municipal Archives, and banking records. However, with
the completion of one directory, the HSD is far enough along to demonstrate results.

The Entity – Relationship Model

There are numerous ways in which the data in the directory could be placed
into a relational database. The most obvious was that the database could simply
consist of one table, containing all the fields. However, the sparse nature of some of
the data makes this approach inefficient. For example, most of the entries only have
one address associated with them. Other entries may have two, three or four
addresses. To use the one-table model, a table would have to be constructed with
four fields to contain addresses, when fewer then 2% of the population use a third
and/or fourth address. Space would likewise be wasted in fields for ethnicity, marital
status, and a few other categories.

To better utilize system resources, a multiple table format was constructed and
is outlined below.
### person table

<table>
<thead>
<tr>
<th>field name</th>
<th>description</th>
<th>attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>entity</td>
<td>a unique key for a specific individual or business</td>
<td>primary key</td>
</tr>
<tr>
<td>lastname</td>
<td>individual's last name or a business' name</td>
<td></td>
</tr>
<tr>
<td>firstname</td>
<td>individual's first name</td>
<td></td>
</tr>
<tr>
<td>middlename</td>
<td>individual's middle name</td>
<td></td>
</tr>
</tbody>
</table>

### address table

<table>
<thead>
<tr>
<th>field name</th>
<th>description</th>
<th>attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>entity</td>
<td>a unique key for a specific individual or business</td>
<td>primary key</td>
</tr>
<tr>
<td>occupation</td>
<td>the vocation or profession of the individual or business</td>
<td>primary key</td>
</tr>
<tr>
<td>streetlocation</td>
<td>the entities address in a <em>streetnumber streetname</em> format.</td>
<td>primary key</td>
</tr>
<tr>
<td>city</td>
<td>Manhattan by default, other cities such as Brooklyn, Williamsburg, and even London and Havana get mentioned</td>
<td></td>
</tr>
</tbody>
</table>

Note that the *entity* field in the *person* table is the sole primary key, but in the *address* table it is part of a composite primary key. This allows many-to-many relationships between individuals, occupations, and buildings, that is, one building at a particular address can contain more than one business owned and/or operated by

---

* A detailed examination of the *entity* field is given in the data entry section, below.
different individuals. It also allows specific individuals to own businesses at one or
more addresses, or for one individual to own different types of businesses at different
addresses.

**ethnicity table**

<table>
<thead>
<tr>
<th>field name</th>
<th>description</th>
<th>attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>entity</td>
<td>a unique key for a specific individual or business</td>
<td>primary key</td>
</tr>
<tr>
<td>value</td>
<td>either &quot;col'd&quot; or &quot;colored&quot; depending on how it appeared in the directory</td>
<td></td>
</tr>
</tbody>
</table>

**widow table**

<table>
<thead>
<tr>
<th>field name</th>
<th>description</th>
<th>attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>entity</td>
<td>a unique key for a specific individual or business</td>
<td>primary key</td>
</tr>
<tr>
<td>spouseLastName</td>
<td>last name of deceased spouse</td>
<td></td>
</tr>
<tr>
<td>spouseFirstName</td>
<td>first name of deceased spouse</td>
<td></td>
</tr>
<tr>
<td>spouseOccupation</td>
<td>deceased spouse's occupation – probably, but not yet determined, being continued by widow</td>
<td></td>
</tr>
</tbody>
</table>

**wife table**

<table>
<thead>
<tr>
<th>field name</th>
<th>description</th>
<th>attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>entity</td>
<td>a unique key for a specific individual or business</td>
<td>primary key</td>
</tr>
<tr>
<td>spouseLastName</td>
<td>last name of husband</td>
<td></td>
</tr>
<tr>
<td>spouseFirstName</td>
<td>first name of husband</td>
<td></td>
</tr>
<tr>
<td>spouseOccupation</td>
<td>husband's (or family's) occupation</td>
<td></td>
</tr>
<tr>
<td>field name</td>
<td>description</td>
<td>attribute</td>
</tr>
<tr>
<td>------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>-----------</td>
</tr>
<tr>
<td>entity</td>
<td>a unique key for a specific individual or business</td>
<td>primary key</td>
</tr>
<tr>
<td>value</td>
<td>Unknown what this value is.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Rarely throughout the directory a one or two letter code will appear at the</td>
<td></td>
</tr>
<tr>
<td></td>
<td>far right of the page. The letters are either &quot;d, x, nr, r, dd, u, m, b&quot;</td>
<td></td>
</tr>
<tr>
<td></td>
<td>and are associated with an entry.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>It is hoped that by either plotting these addresses, or performing</td>
<td></td>
</tr>
<tr>
<td></td>
<td>quantitative analysis on the data, their purpose will become clear.</td>
<td></td>
</tr>
</tbody>
</table>

The Data Entry

To enter the data from the directory into the database took over five months at roughly four hours per day. Initially, a section of the directory, usually corresponding to the first letter of the entries' last name's (for example all of the As, Bs, etc.) was scanned into an Adobe Photoshop formatted file at 300 dpi (image 1, above). Optical character recognition software (PaperPort) was then used to convert the image to a text (.txt) file. The high resolution of the scan was found necessary to decrease the amount of errors in the OCR process. By no means did it remove all the errors.

After the OCR was done, each line was manually compared to the book and retyped. At minimum, the OCR left at least 3-4 characters per line incorrectly scanned. More often whole lines had to be retyped. From a data entry standpoint, the only real advantage of scanning the pages first was to ensure, or at least maximize the probability, that no lines in the directory were left out of the text file. A second advantage, that of tracing data back to its primary source, will be addressed later.
While correcting the OCR generated text file, commas and metadata were inserted to delimit the various fields into the following format:

last name, first name, middle name (initial), occupation, street number, street name, marker that next address is a residence ("h.") or second business address ("2nd") or third business address ("3rd"), etc., the corresponding alternate address street number, the corresponding alternate address' street name, and possibly one or more bracketed (metadata) sets denoting additional information such as ethnicity or marital status.

In the early stages of data entry, one other minor problem arouse which had to be dealt with. Prior to the data entry, the author viewed the types of font to be used as a purely esthetic question. While attempting to correct the OCR text files, however, it quickly became apparent that the OCR program converted some lower-case Ls to the number one, and made several other similar mistakes. These inconsistencies wreaked havoc on text searching. Luckily, this problem was caught early in the data entry stage and courier new, a font which seems to make clear distinctions between all standard ASCII characters, was used.
erand Edward, bootmaker 73 Avenue 9th 'erhiaer Abraham, builder 233 Seventh rerhulse Isaac, carpenter 139
Laurena - 'aria George P. whips 32ST Broadway h. 191 Greene rents & Archer, whips 328k Broadway rering
James, 3 State
rertoni Charles, carpenter Lewis h. 315 Third rerLon Malatiah, platf. scales 12 Allen 6 rertop Richard C. 22
Nassau h. 12 Allen
mertois William C. sawfiler 127 Ridge carton & Main c. h. brokers 22 Nassau
riaglon Henry A. broker:1 Hanover ha. Eleventh c. As’. 6th siimgtomm James, notary 34 Wall h. 128 Macdougal
ringlon & Bobinaon, brokers 1 Hanover c. Wall wenBenjamin, chairgilder 37 Burton
wets Benjamin, printer 30 Vandewater wen Edward H. attorney 11 Pine h. 175 Greene wan Edward W. clerk 135
Cedar
sea Ferris, confectons 1 Avenue 6th wait George B. tavern 3 MoLt loan Henry, furniture Grand hess John,
dockhuildcr 5 Avenue D heen John, tailor 90 Orange jwen rev. John 3. 183 Orchard
Lwen & Co. Ora, graniteyard 612 Greenwich h. 13 Morton beaSt. John, carter 485 Grand
Owen Philip, mason 160 Thirtteenth Owa* Samuel H. millinery 235k Division Owaa Thomas, merchant 209 Front
h. 24 Oliver Owen William, tavern 10 Centre-rr. Oaeit Mary widow of Jehiel, 158 flowery Oweai Edward, 153
Hudson
Owens Francis, carpenter rear 128 Leonard Owene James, coal 191 Mulberry
Owens Philemon S. grocer 27 Burton h. 26 Avenue 6th Owens Robert, rigg-r 5 Birmingham
--Owens Thomas, carter 342 Greenwich Owens Olive widow of Ebenezer, 9 Goerck Owene Rebecca widow of
Thomas, 375 Grand
Owens Rosanna widow of Charles, milliner 441 Pearl On William, exchange 197 Gree-swich

Image 2, page 485 after run through OCR program
Note that this particular page is worse the most, but not all, after OCR. Approximately 25% of the pages scanned resembled this one. The "p485#" was added immediately after the scan.
Overand, Edward, bootmaker, 73, Avenue 9th
Overhiser, Abraham, builder, 233, Seventh
Overhulse, Isaac, carpenter, 139, Laurens
Overin, George P., whips, 328-1/2, Broadway, h., 191, Greene
Overin & Archer, whips, 328-1/2, Broadway
Oving, James, 3, State
Ovington, Charles, carpenter, Lewis, h., 315, Third
Ovington, Malatiah, platf. scales, 12, Allen {d}
Ovington, Richard C., 22, Nassau, h., 12, Allen
Ovington, William C., saw filer, 127, Ridge
Ovington & Main, c. h. brokers, 22, Nassau
Ovington, Henry A., broker, 1, Hanover, h., Eleventh c. Av. 6th
Ovington, James, notary, 34, Wall, h., 128, Macdougal
Ovington & Robinson, brokers, 1, Hanover c. Wall
Owen, Benjamin, chair gilder, 37, Burton
Owen, Benjamin, printer, 30, Vandewater
Owen, Edward H., attorney, 11, Pine, h., 175, Greene
Owen, Edward W., clerk, 135, Cedar
Owen, Ferris, confectons, 1, Avenue 6th
Owen, George B., tavern, 3, Mott
Owen, Henry, furniture, Grand
Owen, John, dockbuilder, 5, Avenue D
Owen, John, tailor, 90, Orange
Owen, John J., rev., 183, Orchard
Owen & Co., Ora, granite yard, 612, Greenwich, h., 13, Morton
Owen, St. John, carter, 485, Grand
Owen, Philip, mason, 160, Thirteenth
Owen, Samuel H., millinery, 235-1/2, Division
Owen, Thomas, merchant, 209, Front, h., 24, Oliver
Owen, William, tavern, 10, Centre-m.
Owen, Mary, 158, Bowery {widow of Jehiel}
Owens, Edward, 153, Hudson
Owens, Francis, carpenter, 128 {rear}, Leonard
Owens, James, coal, 191, Mulberry
Owens, Philemon S., grocer, 27, Burton, h., 26, Avenue 6th
Owens, Robert, rigger, 5, Birmingham
Owens, Thomas, carter, 342, Greenwich
Owens, Olive, 9, Goerck {widow of Ebenezer}
Owens, Rebecca, 375, Grand {widow of Thomas}
Owens, Rosanna, milliner, 441, Pearl {widow of Charles}
Oxx, William, exchange, 197, Greenwich

Image 3, the same text file corrected and delimited.

Note: the part starting the letter P has been left out of Image 3 for space considerations.

Once the data is in the comma-delimited form, a Java encoded parser, written by the author, was then used to read the text file and transfer it into proper SQL insert statements. First, the parser read the page number metadata, such as p485# shown above, which was manually added immediately after the OCR conversion. Next the
parser read each line of code on that corresponding directory page and generated an
entity value.

The entity value is far more then just a sequentially incremented integer. It is based on a formula that will, one, allow each piece of data to be easily traced back to the primary source and, two, will allow additional directories or other primary sources to be added to the database.

The first two characters are letters, corresponding to the publisher of the primary source. In all cases so far, "LO" is used, referring to "Longworth," the publisher of the 1840-41 directory. In other years, other publishers compiled the directory, for example Poole and Doggett.

The next six characters are the year(s) of publication, in all cases so far 184041. Naturally, other directories, will have different years and other primary sources, such as tax records, might only have a four-digit value for their year.

The next two-to-three characters are the page number the entry is found on. As mentioned above and shown in images 2 and 3, the '#' was used as a metacharacter after each page number. This allowed the parser to know when it had moved to the text on a new page.

The next two characters are the first two letters of the person's last name or business' name. This is followed by a single character representing the person or business' first name, or a '?' if the information is missing.

The last one-to-two characters are the line number of the entry, counted down from the top of the page.

Thus LO18404197BJJ46 corresponds to Longworth's 1840-41 directory, page
97, the 46th entry down from the top, which, when looked up is a Mr. John Blair.

The entry's initials allow for confirmation that this is indeed the right entry.

This form of the entity field allows for ease in tracing each entry back to its original source. When directories from other years are entered, naturally the six characters corresponding to the year can be changed, as can the first two characters if the publisher had changed. For the few years when different directories from competing publishers were available at the same time, the first two characters only need to be changed and the six-year characters could remain identical. In the future, when either directories from different cities are entered, or different types of data from the same city, such as the New York State Census data, a few other columns might be added to this field; this is a simple process and can be done automatically once a new format is established.

Although good in theory, one problem appeared immediately when the parser was run. Occasionally in the occupation field, a more detailed discussion of what the individual did and/or office hours were also listed. For example, Hoe & Co. listed their occupation as "printing-press makers, machinists in general, sawmakers, &c.; a complete printing-office supplied at short notice." The commas located in this description wreaked havoc with the parser, which was designed to use commas as a delimiter. A search was then done of the text to determine what, if any, characters were not to be found in it. Afterwards, it was decided to change the comma, as a delimiter, to an exclamation mark. This was quickly done using the "find and replace" function of the text editor, one entry at a time, and manually making sure that the desired commas used in the text were not changed.
Once the parser was run on the correctly delimited, multiple text files containing SQL commands were generated. Specifically each entry in the directory was sorted into zero or more entries per table and written as an INSERT statement into the table's corresponding text file. The files were then loaded into their corresponding tables in the HSD database.

Tweaking the Data

After the data was loaded, initial examinations of the data were undertaken. As expected, the abbreviations within the data were causing a major problem. For example, the number of distinct occupations was initially over 8,000. The database was queried for distinct occupations, and then abbreviated occupations in the original text file were sought out and unified. Using "boardinghouses" for example, abbreviations such as "b.house", "board.", "boardh.", boardingh.", etc., were all unified as "boarding house" with a space between the words.

Many occupations listed in the directory are composite words, hyphenated words, or two (or more) words. Further, these terms might have changed over the past 160 years. Therefore, a unified approach of separating all composite words and removing hyphens was accepted in an attempt to create a uniformed method to facilitate text searches of the data.

When all of the occupational abbreviations were spelled out in the uniformed method described above, a total of 1979 distinct occupations remained in the database. Similar abbreviations were corrected for street names, first names, and occasionally other fields.
Naturally, the data was altered in the text files containing the SQL statements, rather than just the database itself, to allow for easier rebuilding of the database in the event of a system crash, and to update the text file for possible additional experimentation.

**Geographical Information System**

As mentioned, the HSD is an interdisciplinary project, initially designed to merge history and computer science. Upon attempting to add a geographic information system (GIS) to the project, it was soon discovered that another discipline, tentatively called *geography-GIS*, needed to be added. Geography-GIS is in itself an interdisciplinary field, combining the geography with many facets of computer science, especially computer graphics.

For the GIS software, two different products were examined, *Map Maker* and *ArcView 9.0*. Basically, the two products do the same thing. They maintain a base map of an area, the map containing a scale grid which can be associated to points on the map. This grid can then be used to associate street addresses with their corresponding points on the map. These programs also allow additional maps to be overlaid onto the base map, and allow databases containing address information to be read in.

Among GIS software, ArcView dominates the market and is considered the standard. ArcView is a suite of incredible tools, including ArcMap – which displays the actual map, and ArcCatalog – which helps in linking databases with ArcMap. The ArcView suite has far more features in it than have been utilized in this thesis.
Nonetheless, it has at least one major draw back. ArcView's shapefile format, in which the base map and the base map's grid are associated, cannot be made with ArcView. Creation of shapefiles, seems to be a corporate secret. The sales staff at ESRI, ArcView's makers, stated that a shapefile cannot be made (despite the fact that ESRI must make them all of the time), and exhaustive internet searches and inquiries to various government employees and assorted university professors have been unsuccessful in determining how to make a shapefile. In fact, it is a little disturbing how little information on this subject is actually on the web. More than one link to government websites which should have contained information to create a shapefile, or a similar file structure, were found out to have been cut. When considering how terrorist groups could use this information, it may not be far fetched to assume the Department of Homeland Security has been at work.

Map Maker has the useful feature of allowing the user to link a map with any grid type and scale the user wishes. Further, these base maps can then be exported in a range of formats, including ESRI's shapefile format. The bad news is that these files can only read data from a Microsoft Access database. The system is not designed to work with more powerful databases, including MySQL.

The issue of the grid's scale is an important one. In 1951, the United States Army developed the Universal Transverse Mercator (UTM) spatial coordinate system. In the UTM system, the entire surface of the Earth is mapped out in meters. Using a UTM system in ArcView would allow the HSD to sort within a one-meter accuracy. Therefore, queries such as "What is the average distance between a blacksmith's shop and a cooper's shop?" would be possible to determine.
Since shapefiles have to be pre-made for ArcView, the user does not always get a choice of the scale. Among searches of New York City, the only shapefile found was a map of present day New York City with a grid in latitude and longitude, known as the tgr36061blk00 base map. Although latitude and longitude is more inline with the author's personal cartographic experience, the lack of detail obtainable was regrettable. Further, websites were found which told that the NYC municipal government does have a more detailed basemap for GIS systems, in UTM, but that the map is restricted to a few city departments⁷, presumably due to security concerns.

Image 4. The tgr36061blk00 base map

⁷ This UTM map is called the NYCMAP. The author's contacts at the New York City Rent Guidelines Board, and the New York City Economic Development Corporations, two city agencies that could clearly benefit from the NYCMAP and the demographics it could provide, have never even heard of it.
After spending considerable time in an effort to get the best of both worlds, i.e. a UTM map grid and the ability to use the MySQL database, and failing, a decision was finally made to keep the MySQL database at the cost of a more detailed map. The GIS map is only one facet of how the Historical Survey Database can be used, and a good SQL-based database will probably prove more useful then a better map on an Access database.

With the decision to use ArcView made, an Object Linking and Embedding Database provider (OLE DB) needed to be acquired, downloaded, and added to ArcCatalog. Once the OLE DB was correctly installed, a link to the HSD database was established.

Upon establishing the link, a small test table was constructed in HSD, containing 4 distinct addresses. One thing that became apparent immediately was that the streetnumber field and the streetname field, as they were separated in the comma-delimiting phase of data entry, needed to be combined so that the numbers and address were part of the same field. Therefore, the address table was changed, merging these two fields into one field called streetlocation. Passing the small test table to ArcMap through ArcCatalog, the system plotted three of four points in about two minutes.

By comparison, it took four and one half days to plot all 45,000+ addresses in the address table. After the address table was run through ArcView, the following result table was obtained:

Matched with a score of 80-100: 2988 (7%)
Matched with score < 80%: 30670 (70%)
Unmatched: 9984 (23%)
As shown above, only 33,658 addresses were actually plotted. The 23% that could not be plotted were due to various reasons. First, different types of abbreviations used in the directory could not easily be converted into a plottable form. For example some address entries did not have a street number, but were written as "Fulton c. Nassau," (the "c." being an abbreviation of "corner of") or "Broadway n. Fulton" (where the "n." is for "near"). Although an address near an intersection is too ambiguous to be plotted, this was not so for corner of. By replacing the abbreviation "c." with "corner of" it was found that the ArcView software could plot the points of the intersection on the map and gave the following result table.

<table>
<thead>
<tr>
<th>Matched with a score of 80-100:</th>
<th>3054 (7%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matched with score &lt; 80%:</td>
<td>30710 (70%)</td>
</tr>
<tr>
<td>Unmatched:</td>
<td>9878 (23%)</td>
</tr>
<tr>
<td>Matched with candidates tied:</td>
<td>21808 (50%)</td>
</tr>
<tr>
<td>Unmatched with candidates tied:</td>
<td>0 (0%)</td>
</tr>
</tbody>
</table>

This yields a total of 33,764 addresses plotted, an improvement of 106 addresses.

A second reason that the data could not be plotted is that the address listed in the directory, although valid, is not currently in use, and therefore is not part of the tgr36061blk00 base map. This problem can be broken down into two distinct sub-problems.

The first sub-problem is that a street name in the directory was replaced with a specific building name. This is almost exclusively a problem with the markets of the
day. For example, a butcher's address of "27 Fulton-market" cannot be plotted. Although the location "Fulton-market" could be replaced with "1 Fulton," and the location would come up on the map, this would require altering the database in a way that would limit the usefulness of potential searches. A better solution is to assign an alternate address for the place names of the buildings. It is believed this will increase plottable addresses by another 350.

The second sub-problem is that the valid 1840 street address is no longer valid, and therefore not showing up on the tgr36061bk00 base map. The most obvious example of this on Image 4, above, is the footprint of the World Trade Center. In 1840 this footprint was a continuation of cross streets, most notably, Fulton. Fulton Street was the only through island route in lower Manhattan, and as such was a main thoroughfare. This problem involves the shapefile which links the street address data to the base map, and as already stated, cannot easily be changed. Luckily, for the most part lower Manhattan streets have not changed much, and this type of error is not too common in the HSD. As the author gets more experience in GIS and better ingratiated into the NYC government's community, it his hoped that either a better base map will appear or a way made apparent to update the corresponding lots on the tgr36061bk00.

The Research

At present the database is accessed strictly through SQL statements at a DOS command prompt. Queries and tables used to hold the results are first entered in a

---

8 On April 20, Larry Silverstein, who holds the lease on the World Trade Center site, stated that the new plan for the site may allow for several cross streets, most notably Fulton Street, to pass through the site again. If this plan comes to pass, then future shapefiles of the area will correct this problem.
plain text file and saved. At the DOS prompt, the table is first created, and the query then run and its results placed in the table. For example, suppose we wish to study the blacksmiths in New York City during 1840. First, the "\testblacksmith.txt" is entered at the DOS prompt. This command fetches and executes a file containing the following SQL command:

```
create table testblacksmith
(
    entity varchar(20) not null,
    streetlocation varchar(30),
    primary key (entity, streetlocation)
) type=InnoDB;
```

Next, the query itself is typed into the DOS prompt, assuming it is relatively short, or can be first written in a text file and then run. Continuing to use blacksmiths as an example, the following query,

```
INSERT INTO testblacksmith SELECT address.entity, address.streetlocation from address where address.occupation = "blacksmith";
```

This places 319 entities and addresses into the testblacksmith table.

In ArcCatalog, the testblacksmith table is accessed via the OLE DB and dragged into ArcMap. Once the table is processed by the geoprocessing function of ArcMap, the following map emerges.
The results of the search show the locations of 235 blacksmiths, out of the city's reported 319 blacksmiths. The missing 86 entries are the result of insufficient data in the directory. One of the major problems is that many addresses not included do not have a street number, but are listed with entries such as "Twelfth near Wooster." Although this can give a researcher a general idea of where the shop is, as stated above, it is too ambiguous a point to plot. Another entry has a street number of "68" but no street name. Although there is a blacksmith shop at 68 Wooster, this is inconclusive evidence to associate these two entities as being at the same location.

It should also be pointed out that the addresses mapped do not add up to 235. The software reported that 116 entries were duplicate addresses and, naturally, are only represented by one point per address.
By looking at Map 1, the user can instantly see a cluster of blacksmith shops, centered on Broome Street, plus or minus two blocks, and stretching from the East River to the Hudson. Prior to this query, this clustering was probably unknown.

The data, represented in this GIS form, immediately suggests three possible interpretations for this cluster. One, since blacksmiths were dependent on a source of fuel, usually coal, perhaps the city's coal suppliers were located in this area?

Two, during the mid-late 1830s, the industrial development in the port, and perhaps the world, might have reached a phase when a greater number of blacksmiths per capita were required. A larger number of smiths could have been trained to fill the need. They might then have purchased property where it was then readily available for a reasonable price, in this case, the then outskirts of the city.

Three, there could have been a greater number of smiths initially scattered throughout the city south of this cluster, prior to the Great Conflagration of 1835. The conflagration, a massive fire that swept 17 blocks of lower Manhattan, reportedly destroyed over 700 houses, and burnt many shipping, wholesale drygoods, and grocery houses. This area could also have housed a number of blacksmiths. During the rebuilding could the smiths have decided, and/or were guided, to make their shops on the then northern boundary of the city in an attempt to decrease the risk of another major urban fire?

Without the data represented in this form, the knowledge of the location of the blacksmith shops would likely not have been noticed, and the three potential interpretations would not have been made. We can now further use the data to start eliminating these possible interpretations.
Map 2 still shows the blacksmiths, still as red circles, but this time people and businesses listed as either coal merchants or coal yards have been added as blue triangles. The database contains 71 entries listed as coal, presumably coal merchants, and 2 listed as coal yards, totaling 73 entries. Of these 73, only 50 (68%) are matched and plotted on the map. In reality, only 23 triangles appear on the map, the remaining 27 entries are duplicated at these address, presumably business partners and/or employees of these yards. As observed, no clustering of coal dealers corresponds to the clustering of blacksmiths. This suggests that the first proposed interpretation, mentioned above, is invalid.
When the HSD expands to include the NYC directories from 1834-1840, it should be possible to run similar searches to determine if blacksmiths were more common in lower Manhattan prior to the fire. Further, without even bothering with plotting the points on the map, the database itself can be used for standard quantitative history to determine what percentage of the population was made up of blacksmiths during the different years.

Even without the expansion of the database, however, a large number of patterns can be discerned and the output might be used to establish a better understanding of the data itself. For example, as mentioned above when discussing the entity-relationship of the misc table, the Longworth's Directory occasionally had a character on the far right hand side of the page. Enquiries at the New York Public Library and the New York Municipal Archives showed that no one knew what these markings were. It was hoped that by analyzing the data, a pattern would emerge which would help to decode the markings.

Map 3, below, shows the addresses associated with a 'd' plotted onto the map. This character was the most commonly used code in the directory. As Map 3 shows, the locations associated with the 'd' code are distributed throughout Manhattan. However, towards the southern edge of the map, a large cluster is visible.
Map 3. Locations of address marked with a 'd' on the far right side of the directory's page.

Map 4, which shows the nine entries associated with the code 'dd,' places about one half the entries within this same area. Map 5, shows that the entries marked 'b,' 'x,' and 'nr' are also clustered in this area. The only exception to this is the location of items marked 'u,' and the sparseness of this value might have something to do with that.
Map 4. Locations of address marked with a 'dd' on the far right side of the directory's page.

Map 5. Locations of address marked with a 'b' (red circles), 'x' (blue hexagons) and 'nr' (green triangles) on the far right side of the directory's page.
We therefore can see a pattern emerging in this area of New York. Further, when we look at various businesses and trades, for example banks (Map 7), wool and cotton dealers (Map 8) and drygoods dealers (Map 9), we see similar clustering in this same area.
Map 7. Location of banks in New York City in 1840.

Map 8. Location of wool dealers (red pentagons) and cotton dealers (blue triangles) New York City in 1840.
For many of these occupations, this clustering is no surprise. During 1840, the stretch of South Street, just to the southeast of this area of clustering, was the primary docks for the nation's largest seaport. Goods such as southern cotton were first brought to New York prior to export to Europe. Wool was often purchased in England and Continental Europe and imported to the United States through these docks.\(^9\) Drygoods also came into the port through South Street, and naturally a good number of banks placed their offices within this area to make money off of the shipping interests.

New York City's South Street Seaport Museum has long maintained that Wall Street came out of South Street, meaning that the port activities of the city made it the

---

\(^9\) The first actual packet line -- that is, a line of shipping that had a predetermined route and a predetermined time of departure -- the Black Ball Line, was started by Quakers seeking a regular shipping schedule to facilitate their wool importation business.
financial capital of the world. These clusters of businesses and financial institutions in this area of the port suggest that the museum's claim is valid.

Although these clusters do not tell us what the codes in the misc table stand for, since the clustering does center on what must have been some of the most commercially useful real estate in Manhattan, this suggests that the entries might be related to the real estate part of the entry in the directory and not necessarily the people. This property must have been expensive, and therefore it would seem likely that multiple businesses would share the same building. Therefore, 'd' could be for "downstairs," 'u' for "upstairs," 'b' could be for "back," and 'nr' for "near."

Although seemingly sound based on common business practices this theory is not conclusive due to several problems with it. One problem is that some of the entries in the directory actually state "rear" in the text, meaning that at least one other method to designate such locations already existed. Also it is known that in the 1840s, it was common for buildings to be sub-divided to allow businesses to be placed on different floors of the same building. Therefore, if 'u' does stand for "upstairs" there should be a far greater number of entries containing it.

Aside from studying the business side of life in 1840 New York, the HSD can also be used to give a better understanding of social life. For example, Map 10 shows the locations of public houses, taverns and bars. Again the data shows a cluster, this time a little farther north and East of the previous cluster. More specifically, the southern half of this cluster practically outlines the location of New York's infamous Five-Points district, one of the most notorious areas of New York in its day. It is interesting to note that the Fire-Points district typically is considered to have ended in
about the center of this cluster. If a correlation exists between the amount of public
drinking establishments and the criminal activity of an area, it may well be possible to
prove that the reputation of Five-points should have been extended beyond the semi-
oficial borders of the neighborhood.

By contrast, by looking at the homes and offices of some of the professionals,
the more genteel areas of the city show up. Map 11 locates the homes of sail
shipmasters. Note the concentration on the east side, just north of the port area. This
area is not only situated a short walk from the primary port location, but also near
some of the city's shipyards, or directly across the East River from some of the yards
in Brooklyn.
Map 12 shows both the homes and offices of the cities 751 attorneys. What is interesting here is that out of the 751 attorneys, only 379 listed their home addresses in the directory. However, Map 12 suggests that the residences outnumber the offices. In truth, although the number of attorney offices listed seems to outnumber the homes by a near 2:1 ratio, the fact is that there is a very high rate of repetition of office addresses in the data, thus making fewer plottable points. The reason for this is most likely the fact that in 1840, as now, lawyers tended to form law firms rather than work as individuals. We also notice that the vast majority of these law offices are in the same business cluster that was noticed earlier. It is interesting to note that the majority of law firms seem closer to the port than to city hall and the courthouses.
Map 12. Offices (green) and residences (purple) of attorneys in NYC in 1840

It is also interesting that the cluster of bars, taverns and counting houses shown in Map 10 seems to be only sparsely populated with the representatives of these professions. In short, it would therefore seem that this was an undesirable section of the city.

It has been demonstrated that with only the one directory loaded in the HSD, patterns such as clusters of craftsmen, business centers, and the location of fashionable and undesirable neighborhoods can be determined. However, one point is not sufficient to show a pattern, and although the HSD does allow more then one point per map, each map is only one point in time. Additional directories must now be added to continue the potential this project offers.
Further, in addition to additional directories, additional types of data entered will enhance the project. The city's nineteenth century tax records, for example, stored in the New York City Municipal Archives will allow a far greater detail of comparison when attempting to find the wealthier and poorer areas of the city. Best of all, the surviving records of many of the banks listed in the directory have already been located. Information such as savings accounts and mortgages will allow users to "follow the money" on a scale never before possible. Business and personal relationships that have not survived in other written forms will become apparent and new interpretations, as well as a better understanding, of life in nineteenth century New York will emerge.
Bibliography
