Putting People In Places:
a semi-automated approach to parsing and geocoding city directories

Gary Spikberg, Department of Computer Science
Ankitha Pille, Department of Computer Science
Elijah Pass, Department of Computer Science
Robert Pastel, Department of Computer Science
Don Lafreniere, Department of Social Sciences

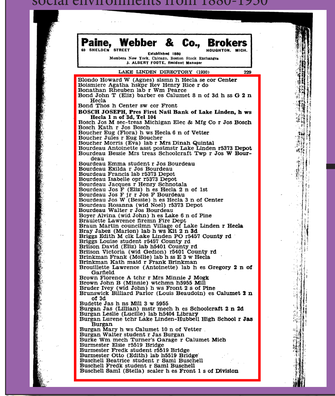
Michigan Tech

Keweenaw TIME TRAVELER
map your history!

Historical Environments Spatial Analytics Lab

Step One: Digitize

The aim of this project was to create a semi-automated approach for converting text based city directories into geographically referenced points in a GIS. This project contributes to the larger Keweenaw Time Traveler project which is creating a 3D geospatial model of the industrial mining communities in Michigan's Copper Country complete with a digital reconstruction of the built and social environments from 1880-1950



Paine, Webber & Co., Brokers
M. SHELDICK STREET — Established 1880 — HOUGHTON, MICH.
Members New York Cotton Assoc from Hancock.
J. ALBERT FOOTE, Resident Manager

LAKE LINDEN DIRECTORY. (139)

Blondo Howard W (Agnes) slsmn h Hecla se cor Center
Boismiere Agatha hskpr Rev Henry Rice r do
Bonathan Rheuben lab r Wm Pearce
Bond John T (Eliz) barber es Calumet 8 n of 3d h ss G 2 n Hecla
Bond Thos h Center sw cor Front
BOSCH JOSEPH, Pres First Natl Bank of Lake Linden, h ws Hecla 1 n of 3d, Tel 104
Bosch Jos M sec-treas Michigan Elec & Mfg Co r Jos Bosch
Bosch Kath r Jos Bosch
Boucher Eug (Flora) h ws Hecla 6 n of Vetter
Boucher Jules r Eug Boucher
Boucher Morris (Eva) lab r Mrs Dinah Quintal
Bourdeau Antoinette asst postmstr Lake Linden r5373 Depot
Bourdeau Bessie Mrs treas Schoolcraft Twp r Jos W Bourdeau
Bourdeau Emma student r Jos Bourdeau
Bourdeau Exilda r Jos Bourdeau
Bourdeau Francis lab r5373 Depot
Bourdeau Isabelle opr r5373 Depot
Bourdeau Jacques r Schnotala
Bourdeau Jos F (Eliz) h es Hecla 2 n of 1st
Bourdeau Jos F jr r Jos F Bourdeau
Bourdeau Jos W (Bessie) h es Hecla 3 n of Center
Bourdeau Rosanna (wid Noel) r5373 Depot
Bourdeau Walter r Jos Bourdeau
Boyer Alvina (wid John) h es Lake 6 n of Pine
Braulette Lawrence firemn Fire Dept
Braun Martin councilmn Village of Lake Linden r Hecla
Bray Jabez (Marion) lab h ws Kit 2 n 3d
Briggs Edith M clk Lake Linden PO r5457 County rd
Briggs Louise student r5457 County rd
Brilson David (Eliz) lab h5401 County rd
Brilson Victoria (wid Gedion) r5401 County rd
Brinkman Frank (Mollie) lab h ss E 3 w Hecla
Brinkman Kath maid r Frank Brinkman
Brouillette Lawrence (Antoinette) lab h es Gregory 2 n of Garfield
Brown Florence A tchr r Mrs Minnie J Mogk
Brown John B (Minnie) wtchmn h5955 Mill
Bruder Ivey (wid John) h ws Front 2 s of Pine
Brunswick Billiard Parlor (Louis Beaudoin) es Calumet 2 n of 3d
Budette Jas h ns Mill 3 w 5955
Burgan Jas (Lillian) matr mech h es Schoolcraft 2 n 2d
Burgan Leslie (Lucille) lab h5404 Library
Burgan Lurene tchr Lake Linden-Hubbell High School r Jas Burgan
Burgan Mary h ws Calumet 10 n of Vetter
Burgan Walter student r Jas Burgan
Burke Wm mech Turner's Garage r Calumet Mich
Burmester Elsie r5519 Bridge
Burmester Fredk student r5519 Bridge
Burmester Otto (Edith) lab h5519 Bridge
Buschell Beatrice student r Saml Buschell
Buschell Fredk student r Saml Buschell
Buschell Saml (Stella) sealer h es Front 1 s of Division

After scanning the directories we use ABBYY Finereader Pro to crop to the names, and read them using the Optical Character Recognition (OCR) function. From here we save a text only copy of the directory, as well as the cropped image. Both will be used later.

Step Two: Clean Data

To clean the text file before it can be parsed all of the multiline entries need to be placed on one line. A manual cleaning of OCR errors must be complete. For example, 'W's tend to be recognized as 'VV' or' \V' due to faded print. This is especially difficult when working from microfilm copies of city directories.

Blondo Howard W (Agnes) slsmn h Hecla se cor Center
Boismiere Agatha hskpr Rev Henry Rice r do
Bonathan Rheuben lab r Wm Pearce
Bond John T (Eliz) barber es Calumet 8 n of 3d h ss G 2 n Hecla
Bond Thos h Center sw cor Front
BOSCH JOSEPH, Pres First Natl Bank of Lake Linden, h ws Hecla 1 n of 3d, Tel 104
Bosch Jos M sec-treas Michigan Elec & Mfg Co r Jos Bosch
Bosch Kath r Jos Bosch
Boucher Eug (Flora) h ws Hecla 6 n of Vetter
Boucher Jules r Eug Boucher
Boucher Morris (Eva) lab r Mrs Dinah Quintal
Bourdeau Antoinette asst postmstr Lake Linden r5373 Depot
Bourdeau Bessie Mrs treas Schoolcraft Twp r Jos W Bourdeau
Bourdeau Emma student r Jos Bourdeau
Bourdeau Exilda r Jos Bourdeau
Bourdeau Francis lab r5373 Depot
Bourdeau Isabelle opr r5373 Depot
Bourdeau Jacques r Henry Schnotala
Bourdeau Jos F (Eliz) h es Hecla 2 n of 1st
Bourdeau Jos F jr r Jos F Bourdeau
Bourdeau Jos W (Bessie) h es Hecla 3 n of Center
Bourdeau Rosanna (wid Noel) r5373 Depot
Bourdeau Walter r Jos Bourdeau
Boyer Alvina (wid John) h es Lake 6 n of Pine
Braulette Lawrence firemn Fire Dept
Braun Martin councilmn Village of Lake Linden r Hecla
Bray Jabez (Marion) lab h ws Kit 2 n 3d
Briggs Edith M clk Lake Linden PO r5457 County rd
Briggs Louise student r5457 County rd
Brilson David (Eliz) lab h5401 County rd
Brilson Victoria (wid Gedion) r5401 County rd
Brinkman Frank (Mollie) lab h ss E 3 w Hecla
Brinkman Kath maid r Frank Brinkman
Brouillette Lawrence (Antoinette) lab h es Gregory 2 n of Garfield
Brown Florence A tchr r Mrs Minnie J Mogk
Brown John B (Minnie) wtchmn h5955 Mill
Bruder Ivey (wid John) h ws Front 2 s of Pine
Brunswick Billiard Parlor (Louis Beaudoin) es Calumet 2 n of 3d
Budette Jas h ns Mill 3 w 5955
Burgan Jas (Lillian) matr mech h es Schoolcraft 2 n 2d
Burgan Leslie (Lucille) lab h5404 Library
Burgan Lurene tchr Lake Linden-Hubbell High School r Jas Burgan
Burgan Mary h ws Calumet 10 n of Vetter
Burgan Walter student r Jas Burgan
Burke Wm mech Turner's Garage r Calumet Mich
Burmester Elsie r5519 Bridge
Burmester Fredk student r5519 Bridge
Burmester Otto (Edith) lab h5519 Bridge
Buschell Beatrice student r Saml Buschell
Buschell Fredk student r Saml Buschell
Buschell Saml (Stella) sealer h es Front 1 s of Division

Step Three: Parse Directory Data

Beerling Hjalmar (Clara E) whsemn I E Swift Co h15 Albion
Beesley Geo r147 Montezuma
Beesley Harry S h147 Montezuma
Beiring Anthony eng Copper Range R R r Hancock
Beiring Francis atndt Goodell Service Station r Hancock
Beiring Jos H bkpr Paine Webber & Co r Hancock
Bekia Pritz student r131 Agate
Bekola J F student Mich College of M & T r132 Blanche
Bellenbaum Edw A student Mich College of M & T r181 Ruby
Belling Jas H carmn Cudahy Pkg Co r Hancock
Belltari Edna maid 47 College av
Belyan Michl driver Copper Range Motor Bus Co r South Range
BENALLACK RICHARD T (Mamie), Cashier Houghton National Bank, h182 Albion, Tel 1017-J
Benallack Vernon C mech Universal Auto Co r182 Albion
Bendry Rhoda clk Mich College of M & T r9 Jasper
Benedict C Harry lecturer Mich College of M & T
Benge Alf mach Copper Range R R r73 Dodge
Benjamin Fran S student Mich College of M & T r181 Ruby
Bennets Felix (Doris) lab h295 Edwards
Bennett Donald C student Mich College of M & T r125 East
Bennett Frank H (Beatrice) mgr h105 Blanche
Bennett Mary W (wid Thos C) h79 Prospect
Bennett Olive J r79 Prospect
Bennetts Arth student r311 Albion
Bennetts Evelyn S Mrs slswn Levine Morris r206 Albion
Bennetts Geo r137 Hubbell av
Bennetts Thos (Evelyn) mstr 195 Montezuma h206 Albion
Bennetts Wm J (Mae) landscape gdnr 211 Diamond h do
Bentley Frances F probate register Houghton County r Hancock
Berg Josephine C slswn The French Shop r5 Pilgrim ter
Berger Bertha ironer Needham Bros Laudry
Bergeron Frank (Laura) cook h209 Montezuma
Bergeron Raymond Rev asst pastor StIgnatius Ch r107 Portage
Bergh Anna r87 Horatio
Bergh Anton miner h87 Horatio
Bergh John r87 Horatio
Bergh Jos lab r87 Horatio
Bergrath Henry (Adolphine) opr Houghton County Traction Co r Laurium
Bertell Trathen cash J C Penney Co Inc r330 Calverley
Bertram Carl r328 South
Bertram Florence r328 South
Bertram Lillian C sten Jas T Healy Agcy r328 South
Bertram Martin (Clara) lab h328 South
Bernard W Ben (Cora Y) supvr h36 Isle Royale

To allow for rapid importation and searchability of the city directory data, we needed to parse information about each resident from the directory into individual segments.

Occupations, workplaces, residential addresses, and housing tenure are captured as well. The blue underlined areas noted at left are occupations.

In addition to exact civic addresses, there are also relative locations for communities that had not yet received home mail delivery . Those were captured by directions (n, e, s, or w) as well as by the name of who they lived

We created a Groovy script, using regular expressions, to parse the entries automatically. We first isolated words starting with a capital letter (to capture proper names) and recorded middle or first initials were possible.

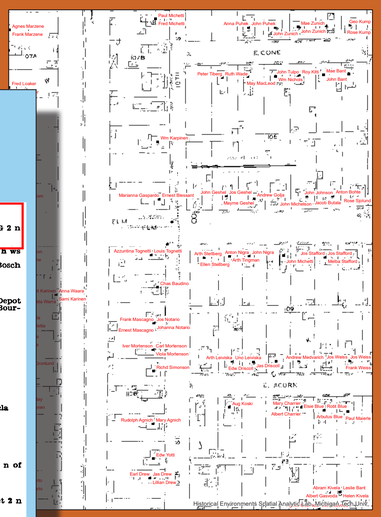
This process was evolved by and tailored to each different set of directories, mostly due to the entry formats changing in both different cities, and different years. The directory to the right contains punctuation after the name, occupation, address. This made it easier to isolate each segment.

A small output sample:

Barkell Chas, sawyer, res Lake st W Hancock.
Barkell Saml, carpenter, for J N Mitchell.
Barrett, Electrician, res Smeltingville.
Bartle Harry, carpenter, res Franklin tp.
Bastrin Herman, carpenter, res Chas Bleise.
BATES S M, a res of Houghton Co since 1861, proprr Hancock Heavy Truck line, No. 114 Franklin, res the same.
Baudette George, lab, bds 210 Ravine.
Baum, Mrs. E, wid, res 312 Ravine.
Baumau Frank W, molder, bds Pacific House.
BAWDEN FRED J, commission merchant, res 903 Water.
Bawden Mrs Richard, res 600 Hancock.
Beauchamp Camil, lab, res Quincy nr cemetery, W Hancock.
Beauchamp D, lab, res Quincy nr cemetery, W Hancock.
Beauchemin C, lab, bds 702 Hancock.
Beauchemin Louis, carpenter, res 702 Hancock.
Beauchesne Z, wks saw mill, res 201 Ravine.
BEDELL GEORGE, a res since 1868, head machinist Quincy Mills, Quincy township.
BEDELL CORNELIUS, a res since 1861, machinist Quincy Mills, res Quincy township.
Beckley Alfred, head sawyer and filer.

outLineNum	outFirstNam	outLastNam	outMI	outMrs	outWid	outSpouse	outJob	outWorkPlac	outT
2	Aesche	Mary	null	Mrs	wid	Jos	null	null	r
3	Allan	Alex	null	null	null	Emma	lab	null	h
4	Allwin	Leo	null	null	null	null	lab	null	h
5	Allwin	Victor	null	null	null	Ida	null	null	h
7	Amos	Nancy	null	null	wid	John	null	null	h
8	Anderle	Harry	J	null	null	null	tchr	Lake Linden Hr	
10	Annabelle	Barrette	null	null	null	null	null	null	r
11	Archambeau	Ida	null	null	null	null	null	null	h
12	Asselin	Albert	J	null	null	Lake Linden	null	null	h
13	Aubin	Hormisdas	null	null	null	Anna	null	null	h
14	Aubin	Jos	H	null	null	Mary	lab	null	h
15	Aubin	Lawrence	null	null	null	Pauline	lab	null	h
17	Aubin	Nelson	null	null	null	Rebecca	null	null	h
19	Audette	Raymond	null	null	null	null	bkpr	First Natl Barr	
21	Ayouette	Alphonse	null	null	null	Alvina	lab	null	h
23	Backman	Amanda	null	null	null	null	elk	A P Tea Co	r
25	Baillargeron	Jos	null	null	null	Mary	lab	null	h
26	Barbeau	Alex	J	null	null	null	pres	Village of Lakh	
28	Baril	Alphonse	null	null	null	Genevieve	lab	null	h
30	Baril	Arth	null	null	null	Mary L	slsmn	Nelson Cantish	
32	Baril	Henry	null	null	null	null	null	null	h
34	Baril	Jas	null	null	null	null	councilmn	Village of Lakr	
35	Baril	Jos	null	null	null	Victoria	brkmn	null	h
39	Baril	Mose	null	null	null	Virginia	lab	null	h
40	Baril	Oliver	null	null	null	Alexina	justice of the	Schoolcroft Th	
41	Baril	Pulcherie	null	null	wid	Diedonne	null	null	r
43	Baril	Walter	null	null	null	Delia	null	null	h
45	Barrette	Uric	null	null	null	Mary	lab	null	h
46	Bawden	John	null	null	null	Eliza	lab	null	h
47	Beauchaine	Jos	null	null	null	Clara	lab	null	h

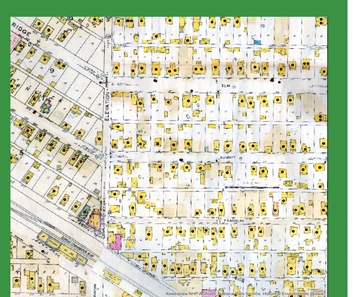
Step Four: Automated Geocoding

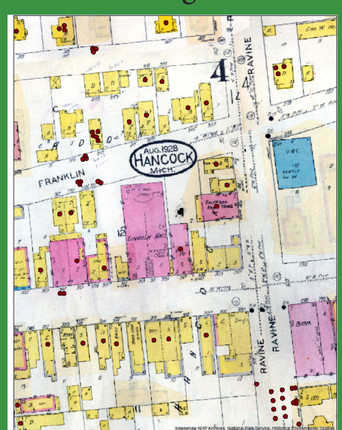


After parsing, we geocode the parsed addresses with a previously completed GIS feature class created by digitizing all of the buildings noted on Sanborn Fire Insurance Plans. The feature class includes the civic addresses and is updated every 10 years roughly coinciding with the decennial census. Several passes through the data are required adjusting the script to account for different tolerances for spellings of streets and street name changes.

Step Five: Manual Geocoding

The final step requires researchers to review matches to ensure accuracy and to manually place relative locations on the map.





City and Year	# of Directory Entries	Total Parsed	% Parsed	Auto Geocoded	Manual Geocoded	% Geocoded
Hancock 1939	4504	4138	91.87%	2494	1095	86.73%
Hancock 1930	3788	3450	91.08%	2069	288	68.14%
Laurium 1930	2575	2459	95.49%	2162	In Progress	87.92%
Laurium 1916	4529	4259	94.04%	1827	In Progress	42.90%
Houghton 1930	3397	3116	91.73%	1457	In Progress	46.76%
Houghton 1916	3606	3142	87.12%	889	In Progress	28.29%