XML Directory to ESRI Shapefile Conversion (Geospatial Data Join) Procedures
Created by: Jessica Branco
Revised: June 23, 2004

Overview:

These procedures provide detailed instructions for creating regularized directories, tab-delimited files, joins with geospatial data and shapefile visualizations of data found in the XML-encoded Boston city directories. These procedures offer a step-by-step guide to using TextPad, CygWin, MS Access, and Esri's ArcMap program for the conversion.

Procedures:

1. In TextPad, parse directory, verify that there are no errors.
2. Select File>Save as, and save copy of directory as YYYY.xml (YYYY represents directory year) in directory/network drive.

Note: You must be on one of these two machines in the downstairs work area (on mctldca23 or mctldca24) to run these next steps in the conversion process.

3. Copy the YYYY.xml file from the directory/network drive to mctldca23 or mctldca24's c:AutomatedProcessing directory. Overwrite an existing earlier version of the file is you are re-running this process.
4. Open Cygwin by clicking on desktop shortcut found on mctldca23 or mctldca24.
5. At prompt, type: cd /cygdrive/c
6. Press Enter
7. At prompt, type: cd AutomatedProcessing
8. Press Enter
9. At prompt, type: ./crosswalk.pl YYYY.xml > YYYYr.xml
10. Press Enter (this process may take several minutes)
11. When the script is finished regularizing street names, it will return a prompt at the command line.
12. After crosswalk.pl is finished, at the prompt, type: ./dir2tab.sh YYYYr (do not include .xml file extension)
13. Press Enter
14. This command will create five xml files: YYYYr-intermixed, YYYYr-complete, YYYYr-acceptable, YYYYr-needs-attn, and YYYYr-out-of-scope. These files will be located in the c: AutomatedProcessing dircectory, where the regularized source YYYYr.xml file was originally placed. "Done" will appear at the prompt when the process is complete.

15. Review lists

a. To compare needs-attentions to complete or acceptable lists. Looking for less than 10% needs-attention
    i. Open all five files in Textpad
    ii. Determine the number of entries included in each of the complete, acceptable, needs-attn, and out-of-scope files by selecting each file's last line and recording the first number in the lower right corner of the window. This is the total number of lines in the document. Subtract 1 to account for the field labels in the first line.
    iii. Total the Complete, Acceptable, and Out-of-Scope entries. If this total is less than 90% of the number of lines in the Intermixed file, continue tagging the directory to increase the level of completeness. Use the needs-attn list as a point of reference for what types of content need to be worked on (occupations, addresses, etc.)

b. For regularization corrections: If there is a significant number of abbreviation or missing streets, augment regularization source list of authority terms and re-run regularization and list making (see augmenting regularization list and script creation below)

16. If 90% or more of the entries are viable and/or out-of-scope, combine complete and acceptable lists using the following process:
    a. Open YYYY-complete.xml and YYYY-acceptable.xml in TextPad
    b. With the YYYY-acceptable.xml document window active, press ctrl-A
    c. With all the text higligheted, press ctrl-c
    d. Switch to the YYYY-complete.xml document window
    e. Scroll to the end of the document
    f. At the end of the last line, press return
    g. With the cursor on the new line, press ctrl-v
    h. Find and delete the line that has the column labels from YYYY-acceptable.
17. Change the column labels in YYYY-complete.xml to the following:
    ID = DIR_ID
    NAME = FULL_NAME
    LASTNAME = SURNAME
    OCCUPATION = OCCUPATION (no change)
    COMMADDR = COMM_ADDR
    RESADDR = RES_ADDR
    DATE = DIR_DATE

18. Add the directory date to end of list entries using the following regular expression:
    a. Find: $
    b. Replace with: \tYYYY
    c. Delete the added YYYY from the end of the column label line (first line)

19. Save file as YYYYjoin.txt

Procedures for Importing, Joining and Exporting Data in MSAccess
Importing:

20. Select MSAccess from the Start Menu and/or Programs
21. Select: File>Open from the top bar menu
22. Navigate to network drive/directory and select 1898addrpts.mdb
23. Select: File>Get External Data>Import from the top bar menu
24. Navigate to YYYYjoin.txt file (change file type to Text files), click Import
25. Select "Delimited", click Next
26. Select "Tab" as delimiter to separate fields
27. Check "First row contains field names"
28. Select {none} as text qualifier, click Next
29. Select "In a New Table" to store data, click Next
30. Click Next again to skip over specific field information
31. Chose no primary Key, click Next
32. Enter YYYYjoin, if it is not automatically generated, in Import to Table field, click Finish

Joining addresses with X and Y coordinate information:
Joining Commercial Addresses:

33. Make the database menu (smaller, non-tabular menu) the active menu.
34. Select Tools>Relationships from the top bar menu
35. Right click in window that appears and select "Show Table" and select the Table tab
36. Highlight the appropriate YYYYjoin table and click Add. (1898addreg should already be visible with previous joins connected).
37. Click Close
38. Select the Comm_Addr field in the menu you've just added.
39. Click and drag the Comm_Addr field (it will turn into a bar) and drop it on top of the Primary_Addr field in the 1898addrpts menu. A line should appear between the two table menus.
40. When Edit Relationship window appears, click Create
41. Close the relationships windows by clicking on the X in the upper right corner. Save changes when prompted.
42. Returning to the database menu (smaller, non-tabular menu) the active menu, select Queries from the list on the left.
43. Select Create Query by using wizard
44. From Tables/Queries pull-down menu, select Table: YYYYjoin
45. Use > button to add fields from the left side list to he right side list, in the following area (select fields on left by clicking on the field name):
    a. Full_Name
    b. Surname
    c. Occupation

   d. Com_Addr
   e. Res_Addr
   f. Dir_ID
   g. Dir_Date
46. Return to the Table/Queries pull down menu and select Table:1898Addreg
47. On the right side list, highlight Res_Add
48. On the left side list, highlight Primary_Addr
49. Add Primary_Addr to the right side list using the > button. It should appear below the Res_Addr
50. Next, add Street_Num, Street_Name, Build_Name using the > button
51. Click on Dir_Date on the right side list and add Plot_Date, X_Coord, and Y_Coord below Dir_Date
52. Click Next
53. Name query YYYY_commercial
54. Select "open query to view information" and click Finish

Exporting Commercial Addresses

55. With YYYY_commercial query as the active window, select File>Export from the top bar menu
56. Navigate to the appropriate networked drive and directory to save the data in, name file YYYY_comercial.txt and export as file type Text Files, and click Export All
57. Select "Delimited" as export type and click next
58. Choose Tab as delimiter
59. Check "Includes field names in first row"
60. Select {none} as text qualifier, click Next
61. Verify that "Export to file" path and filename is pointing to the appropriate directory and has YYYY_comercial.txt as the filename
62. Click Finish
63. Open YYYY_commercial.txt file using TextPad
64. Append a new first column label – "Data_Code" - and value - "COM" - for each entry using the following regular expression:
    a. Find: ^
    b. Replace with: COM\t

    Note: You must change the COM on the first line (the field labels line) to DATA_CODE and make certain there continue to be one Tab between DATA_CODE and FULL_NAME
65. Save changes to YYYY_commercial.txt

Joining Residential Addresses

66. Return to MSAccess, make the database menu (smaller, non-tabular menu) the active menu.
67. Select Tools>Relationships from the top bar menu

68. Right click on the line connecting the 1898Addreg's Primary_Addr and YYYYjoin's relationship, select Edit relationship
69. In the Join window, click on Comm_Addr and change, using pull down menu, to Res_Addr
70. Click OK
71. Close the relationships windows by clicking on the X in the upper right corner. Save changes when prompted.
72. Returning to the database menu (smaller, non-tabular menu) the active menu, select Queries from the list on the left.
73. Select Create Query by using wizard
74. From Tables/Queries pull-down menu, select Table: YYYYjoin
75. Use > button to add fields from the left side list to the right side list, in the following area (select fields on left by clicking on the field name):
    a. Full_Name
    b. Surname
    c. Occupation
    d. Com_Addr
    e. Res_Addr
    f. Dir_ID
    g. Dir_Date
76. Return to the Table/Queries pull down menu and select Table:1898Addrpts
77. On the right side list, highlight Res_Add
78. On the left side list, highlight Primary_Addr
79. Add Primary_Addr to the right side list using the > button. It should appear below the Res_Addr
80. Next, add Street_Num, Street_Name, Build_Name using the > button
81. Click on Dir_Date on the right side list and add Plot_Date, X_Coord, and Y_Coord below Dir_Date
82. Click Next
83. Name query YYYY_residential
84. Select "open query to view information" and click Finish

Exporting Residential Addresses

85. With YYYY_residential query as the active window, select File>Export from the top bar menu
86. Navigate to the appropriate networked drive and directory to save the data in, name file YYYY_residential.txt and export as file type Text Files, and click Export All
87. Select "Delimited" as export type and click next
88. Choose Tab as delimiter
89. Check "Includes field names in first row"
90. Select {none} as text qualifier, click Next
91. Verify that "Export to file" path and filename is pointing the appropriate directory and has YYYY_residential.txt as the filename
92. Click Finish

93. Open YYYY_residential.txt file using TextPad
94. Append a new first column label – "Data_Code" - and value - "RES" - for each entry using the following regular expression:
    a. Find: ^
    b. Replace with: RES\t
    Note: You must change the RES on the first line (the field labels line) to DATA_CODE and make certain there continue to be a single Tab between DATA_CODE and FULL_NAME
95. Save changes to YYYY_residential.txt

Combining Commercial and Residential Address Data

96. Re-open the correspondingYYYY_comercial.txt
97. Select all of the YYYY_residential file, except for the first line (field labels)
98. Copy the selected lines
99. Paste the lines at the next available line at the end of the YYYY_commercial.txt file
100. Select File>Save As
101. Save file as YYYY_combined.txt in the appropriate directory
102. Close all files. Do not save changes to YYYY_commercial.txt

Procedures for creating Shapefiles in ESRI's ArcMap

103. Open ArcMap from the Start Menu on mctldca13 (currently the Reading Room computer, logged in as Archives Pub).
104. Select Start ArcMap with a new, empty map. This is the first option listed.
105. Check "Immediately add data"
106. Cick OK
107. When the add files windows appears, navigate to the appropriate network drive and folder.
108. Select the YYYY_combined.txt file.
109. Click Add
110. The Table of Content window should appear on the white area. If nothing is visible confirm that Table of Contents is selected under the View menu on the top bar.
111. In the Table of Content window, select the source tab at the bottom.
112. Highlight with the mouse the appropriate YYYY_combine.txt file listed.
113. Right click on the selected filename and select Display XY Data
114. Map the X Field to X_Coord, map the Y Field to Y Coord
115. Click OK
116. In Table of Contents, right click on displayed data layer with symbol (file name with Event appended to end usually appears above source .txt file.)
117. Select Data
118. Select Export Data
119. Select Export all features
120. Select Same Coordinate system as this layer's source data

121. Rename Export_Output.shp (at end of file path) to YYYY_combined.shp
122. Click OK
123. Add shapefile to view to verify it works (automatically prompted to choose to add)
124. Select Exit from File Menu to close ArcMap. There is no need to save changes to Untitled map document.