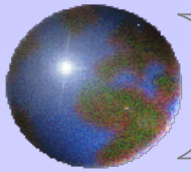
Sonoma County GIS

The Good, The Bad & The Developer Kit

By Michelle Mestrovich & Michael Hansen

September 13, 2006



Sonoma County GIS

Part 1
What is Geocoding?

Part 2
Scrubbing the Data

Part 3
Tricking the Geocoder

Part 4
The Geocoding Developer Kit

Part 1: What is Geocoding?

Part 1: What is Geocoding

- A definition
- Components of the Geocoding Process

Part 1: What is Geocoding?

Geocoding:

The process of identifying the coordinates of a location given its address or unique identifier.

Part 1: What is Geocoding?

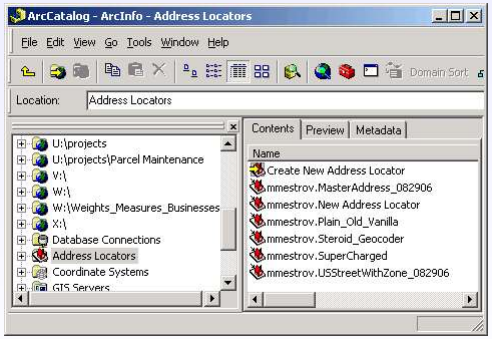
Geocoding Components

- Reference Data

- Address Data

ST_NUM	PRE_DIR	ST_NAME	ST_TYPE	SUF_DIR	ZIPCODE	COMMUNITY
133-144		Healdsburg	Ave.		95448	Cloverdale
1440		Guerneville	Road		95403	Santa Rosa
16124		Drake	Rd.		95446	Guerneville
980		Hopper	Street		95403	Santa Rosa
201	West	Sierra	Ave.		94931	Cotati
6750		Commerce	Blvd.		94928	Rohnert Park
7120		Bodega	Ave.		95472	Sebastopol

- Software.



Part 2: Scrubbing the data

- Initial Geocode
- Make copies of tables
- Inspect existing data.
- Add necessary fields and concatenate.
- Eliminate addresses that won't geocode.
- Eliminate unwanted characters & spaces.
- Find and replace or calculate (fine tuning).

Part 2: Scrubbing the Data

Try an Initial Geocode

Go For It!!!!

You might get lucky!!! (yeah, right ☹)

Part 2: Scrubbing the Data

Make Copies of Tables

- This is a must, make copies of all the tables you plan to manipulate in case you need to get back to the original. Bad things can happen to good tables.

- Also, don't forget to work in an edit session when manipulating your addresses.

Part 2: Scrubbing the Data

Inspect the existing data

- Out of area addresses.
- PO Box Numbers.
- Text designating the Apartment, Unit, Space, etc.
- Formatting issues.
- Misspellings.
- Missing or incomplete addresses.
- Other text in the address field
- Extra spaces inside the text string

Part 2: Scrubbing the Data

Add Necessary Fields & Concatenate

Name	Location
Joes Dinner	1235 W 3rd St, Santa Rosa, 95403

Name	Location	ADDRESS	COM	ZIPCODE
Joes Dinner	1235 W 3rd St, Santa Rosa, 95403			

Location	House_no	PRE_DIR	ST_NAME	ST_TYPE	SUF_TYPE	ZIPCODE	
Martys	12345	W	9th	St		95403	

Location	House_no	PRE_DIR	ST_NAME	ST_TYPE	SUF_TYPE	ZIPCODE	ADDRESS
Martys	12345	W	9th	St		95403	

Part 2: Scrubbing the Data

Eliminate Addresses that will not Geocode

- Removing addresses that cannot be geocoded saves time and improves your match score (e.g PO Boxes, <nulls>, etc).
- Summarize on a field to see if you have any misspellings.
- Maintain a list of valid zip codes and communities in order to eliminate out of area addresses.

Part 2: Scrubbing the Data

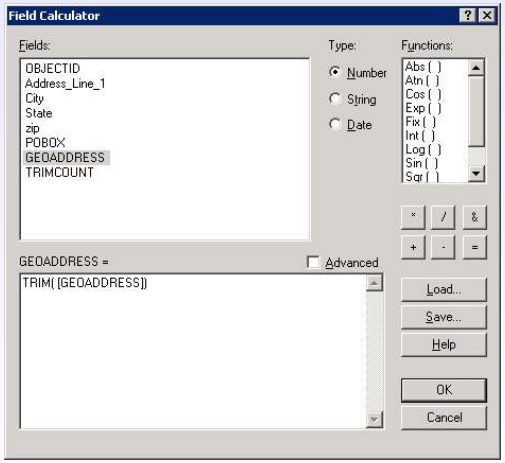
Eliminate unwanted characters & spaces

- Use Trim() string function to remove spaces (Front and Back)

TRIM([FieldName])

- Use Replace() string function to remove unwanted characters.

REPLACE([FieldName], "*OLDSTUFF*", "*NEWSTUFF*")



Part 2: Scrubbing the Data

Eliminate unwanted characters & spaces (cont.)

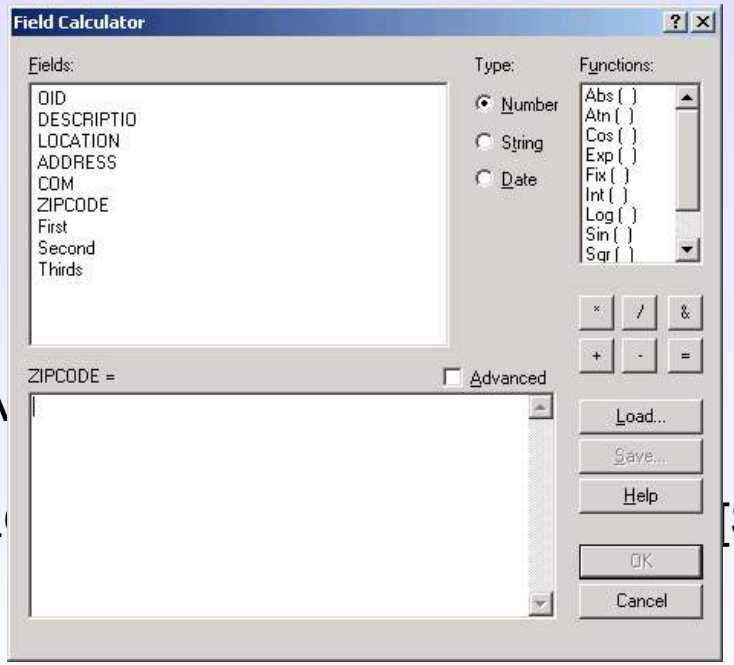
Note: You are not limited to the functions found in the "Field Calculator" dialog. You can utilize other VB functions which are not listed.

Part 2: Scrubbing the Data

Fine Tune the addresses

- Break down the address & clean it up.

Name	Location	ADDRESS	COM	ZIPCODE	FIRST	SECOND
Joes Dinner	1235 W 3rd St, Santa Rosa, 95403	1235 W 3rd St	Santa Rosa	95403	14	26



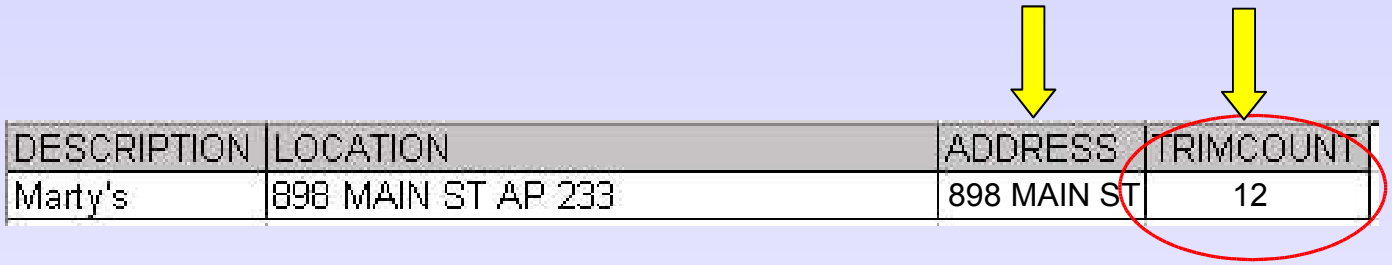
Mid([LOCA [First]) - 1))

Right ([L [Second])

Part 2: Scrubbing the Data

Fine Tune the addresses (cont.)

- Stripping out unwanted data



DESCRIPTION	LOCATION	ADDRESS	TRIMCOUNT
Marty's	898 MAIN ST AP 233	898 MAIN ST	12

InStr([LOCATION], “ AP ")

Trim(Left([LOCATION], ([TRIMCOUNT]-1)))

Part 2: Scrubbing the Data

Fine Tune the addresses

- Concatenate into “Address” field.

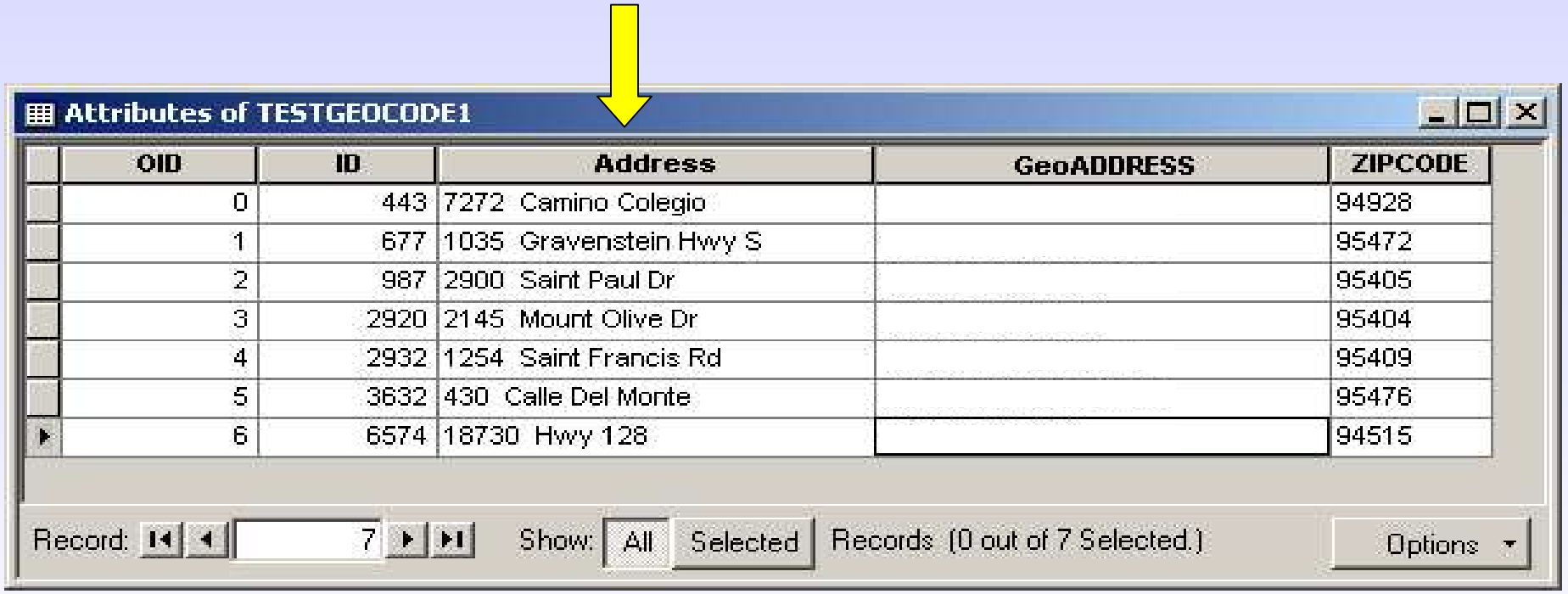
Location	House_no	PRE_DIR	ST_NAME	ST_TYPE	SUF_TYPE	ZIPCODE	ADDRESS
Martys	12345	W	9th	St		95403	12345 W 9TH St

```
TRIM([HOUSE_NO] & “ “ & [PRE_DIR]
        & “ “ & [ST_NAME] & “ “& [ST_TYPE] & “ “ & [SUF_TYPE])
```

Part 3: Tricking the Geocoder

Part 3: Tricking the Geocoder

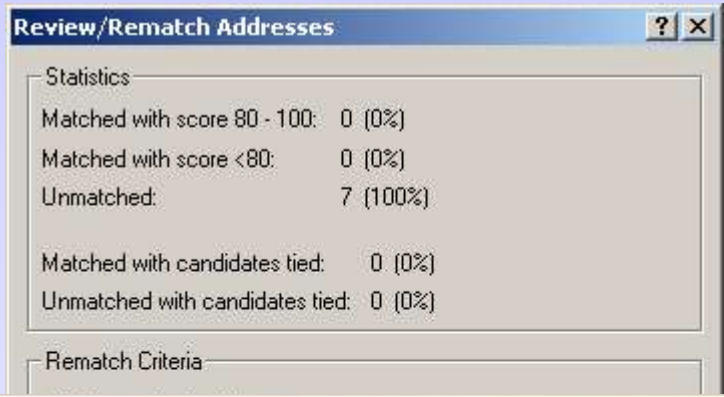
Original Address

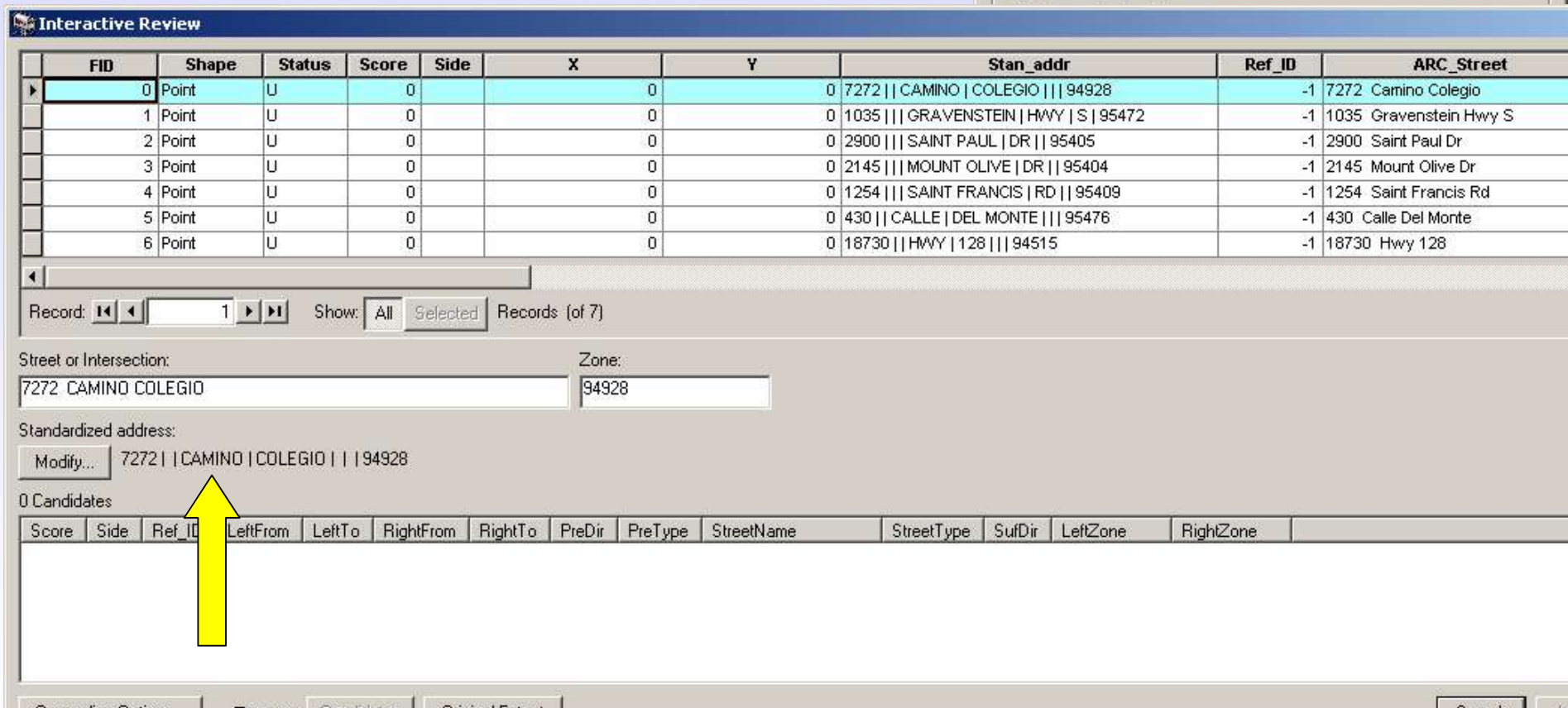


Part 3: Tricking the Geocoder

Geocode results from original address:

Match = 0%





Part 3: Tricking the Geocoder

Sonoma Sonoma County Street idiosyncrasies

Puts “Camino” into PreType

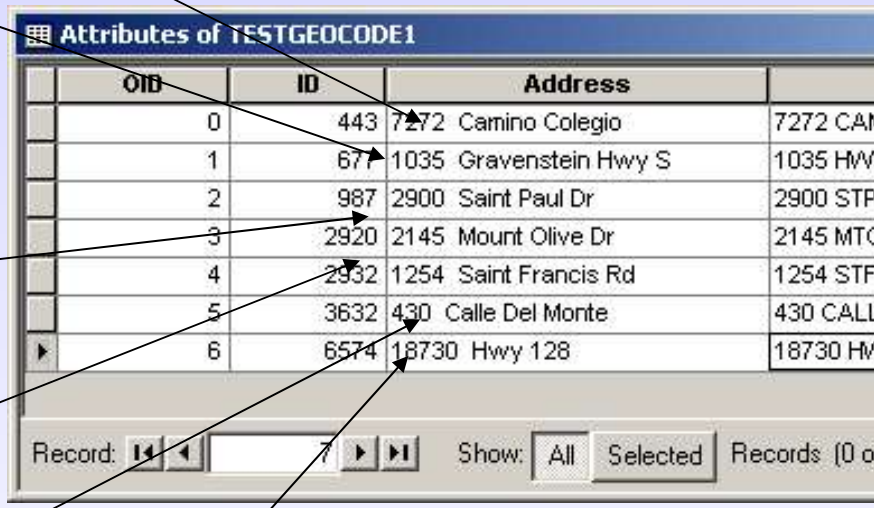
Standardizes Gravenstein Hwy S correctly but street is called Hwy 116 in street data.

“Saint Paul” is “St Paul” in data.

“Mount Olive” is “Mt Olive” in data.

Puts “Calle” into PreType

Put “Hwy” of “Hwy 128” into PreType



Part 3: Tricking the Geocoder

Before Modification – “Camino Colegio”

After Modification – “CaminoColegio”

Modification – Remove Spaces

REPLACE ([ADDRESS], “CAMINO “, “CAMINO”)

Part 3: Tricking the Geocoder

Before Modification – “Gravenstein Hwy S”

After Modification – Hwy Hwy116 S

Modification –
Change “Gravenstein Hwy” to “Hwy Hwy116” with no spaces

(Remember to keep Suffix Direction otherwise you get a tie)

REPLACE ([ADDRESS], “GRAVESTIEN HWY“, “HWY HWY116”)

Part 3: Tricking the Geocoder

Before Modification – “Saint Paul Dr”

After Modification – “StPaul Dr”

Modification –
Change “Saint” to “St” and remove space between “St” and “Paul”

REPLACE ([ADDRESS], “SAINT PAUL “, “STPAUL”)

Part 3: Tricking the Geocoder

Before Modification – “Mount Olive”

After Modification – “MtOlive”

Modification –
Change “Mount” to “Mt” and remove space between “Mt” and “Olive”

REPLACE ([ADDRESS], “MOUNT OLIVE “, “MTOLIVE”)

Part 3: Tricking the Geocoder

Before Modification – “Calle Del Monte”

After Modification – “CalleDelMonte”

Modification – Remove Spaces

REPLACE ([ADDRESS], “CALLE DEL “, “CALLEDEL”)

Part 3: Tricking the Geocoder

Before Modification – “Hwy 128”

After Modification – “Hwy128”

Modification – Remove Spaces

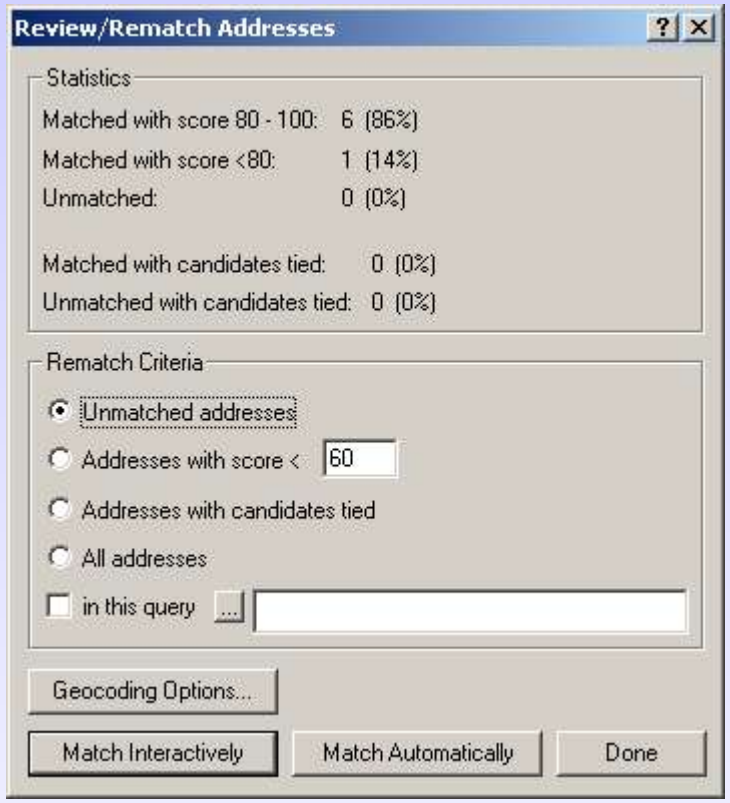
REPLACE ([ADDRESS], “HWY “, “HWY”)

Part 3: Tricking the Geocoder

Geocode results from modified address:

Match = 86%

All scores are less than 100,
but all but one over 90



Attributes of Geocoding Result: Geocoding_Result_14

FID	Shape	Status	Score	Side	X	Y	Stan_addr	Ref_ID	ARC_Street	ARC_Zone
0	Point	M	96	R	6364035.55613	1884626.76393	7272 ||| CAMINOCOLEGIO ||| 94928	37035	7272 CAMINOCOLEGIO	94928
1	Point	M	78	L	6328005.00241	1904659.00276	1035 || HWY | HWY116 || S | 95472	44504	1035 HWY HWY116 S	95472
2	Point	M	93	L	6371515.60451	1917673.275	2900 ||| STPAUL | DR || 95405	44770	2900 STPAUL DR	95405
3	Point	M	94	L	6364097.85809	1921788.6335	2145 ||| MTOLIVE | DR || 95404	4628	2145 MTOLIVE DR	95404
4	Point	M	95	R	6376524.53749	1932531.32472	1254 ||| STFRANCIS | RD || 95409	44895	1254 STFRANCIS RD	95409
5	Point	M	93	R	6424479.80622	1876422.07559	430 ||| CALLEDELMONTE ||| 95476	18428	430 CALLEDELMONTE	95476
6	Point	M	93	R	6375210.35821	1983457.66314	18730 ||| HWY128 ||| 94515	43823	18730 HWY128	94515

Record: 1 Show: All Selected Records (0 out of 7 Selected.) Options

Part 3: The Geocoding Developer Kit

Part 3: The Geocoding Developer Kit

- Getting the Developer Kit
- General Process of Gecoding
- Files involved with the geocoder
- Customizing the Geocoder

Where to find the developer kit

- Geocoding Developer Kit

http://edn.esri.com/index.cfm?fa=downloads.detail&downloadId=22

The kit contains the software program, documentation and samples that are used for creating or customizing geocoding rule bases.

Very, Very, Very important. Read all the documentation first before you mess with the kit.

Part 3: The Geocoding Developer Kit

General Process of Geocoding

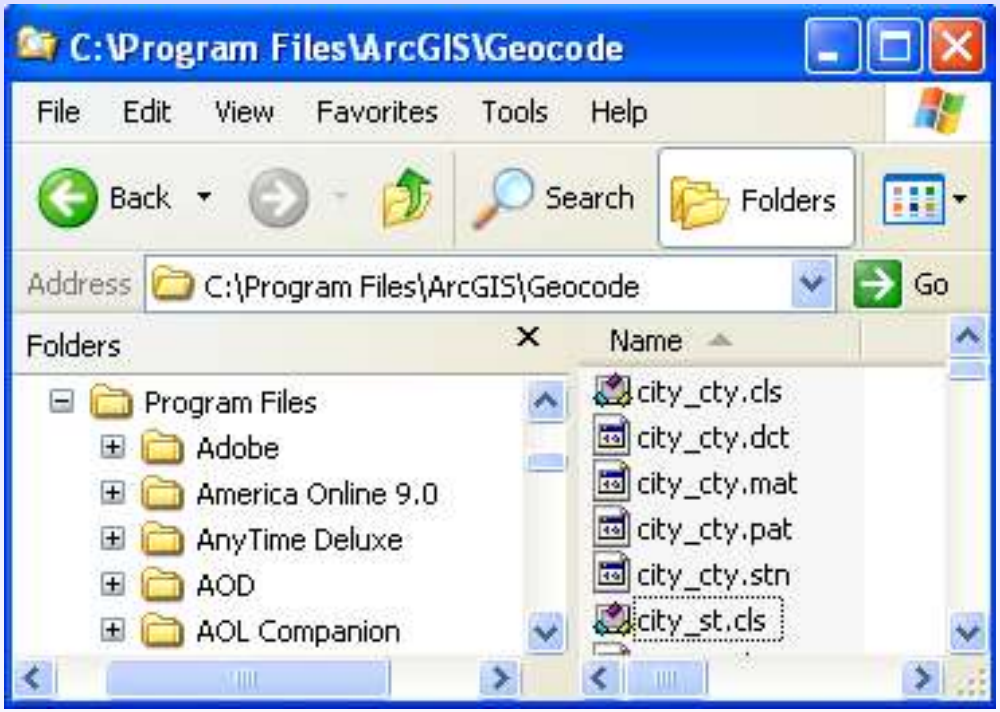
- Define Matching Strategies
- Standardization
- Blocking
- Matching
- Review/Edit

Part 3: The Geocoding Developer Kit

Files involved with the Geocoder

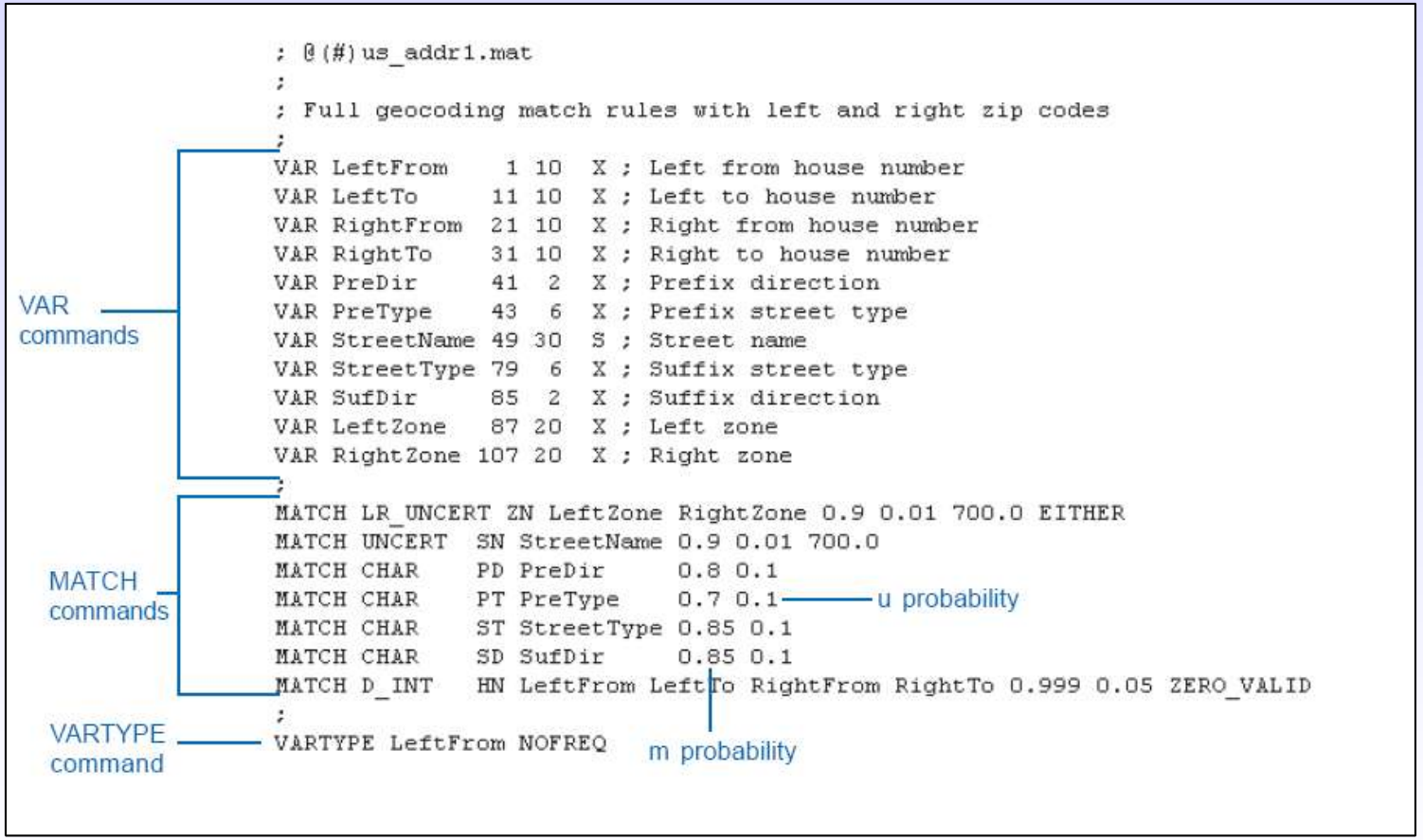
The geocoder is rule based. The rules tell the geocoder how to standardize and match the data to the related location in the reference data.

- Match Rules
  - ------.mat
- Standardization Process
  - ------.stn
  - ------.dct
  - ------.cls
  - ------.pat



Files involved with the Geocoder (cont.)

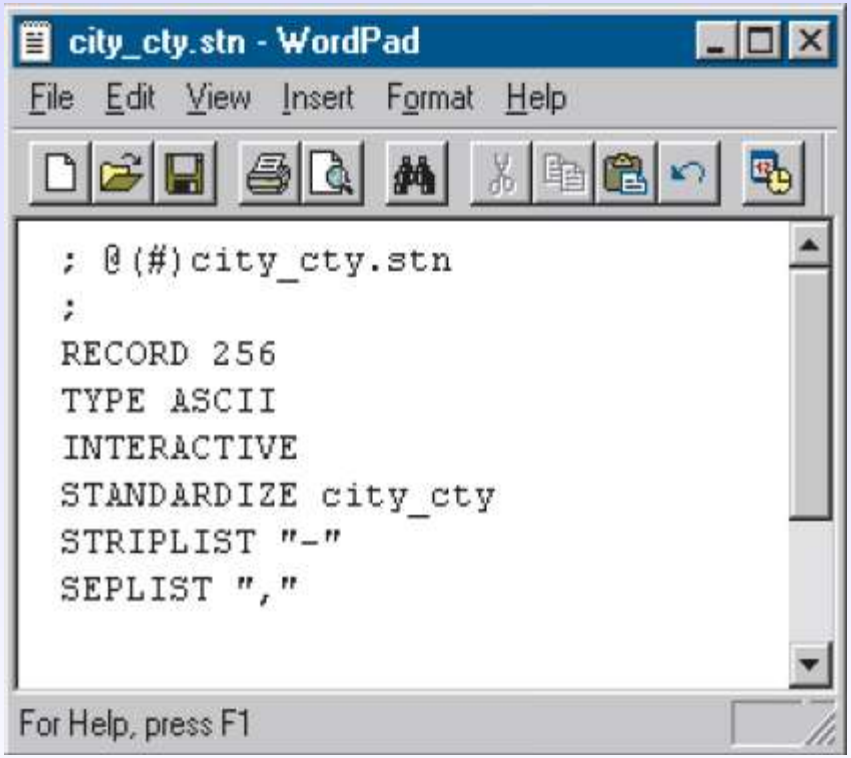
.mat



Files involved with the Geocoder (cont.)

.stn

- Record size
- Input file type
- SEPLIST & STRIPLIST



Files involved with the Geocoder (cont.)

.dct ~ match key dictionary

- Defines which field a portion of an address should be parsed to.

```
\FORMAT\  SORT=N
;  @(#)us_addr.dct
;
;   Street address match key
;
HN    N     10    X; House Number
PD    C     2     X; Pre-direction
PT    C     6     X; Pre-type
SN    C     30    S; Street Name
ST    C     6     X; Suffix type
SD    C     2     X; Suffix direction
```

Part 3: The Geocoding Developer Kit

Files involved with the Geocoder (cont.)

.cls ~ Classification Table

- Used to identify classified keywords that appear in the address.

us_addr.cls - WordPad

File Edit View Insert Format Help

```
\FORMAT\ SORT=Y
; @(#)us_addr.cls
;
; Explanation of classes
;
; O = NULL word (THE, OF)
; B = Box  (BOX)
; Q = Post  (POST)
; Y = Office  (OFFICE)
; K = FPO APO GENDEL
; L = OLD
; M = Multiunit (APT)
; E = Building type
; F = Floor
; G = Directional modifier (END POINT VIEW) for  E
; C = Cardinal number (ONE TWO)
; O = Ordinal number (FIRST SECOND)
; D = Direction (NORTH)
; T = Street type (ST AV)
; R = Rural route (RR)
; X = Route modifier (US, STATE)
; S = St
; N = Number which may be followed by either an or
;     (FIFTY, SIXTY, etc.)
; Z = Number suffix (TH, ND)
; H = Mile
; J = RURAL, STAR
; I = Comppany suffix (INC., AGENCY)
; A = Abbreviations to expand
; V = State names or abbreviations
; P = used internally
```

Explanation of classes in .cls file

us_addr.cls - WordPad

File Edit View Insert Format Help

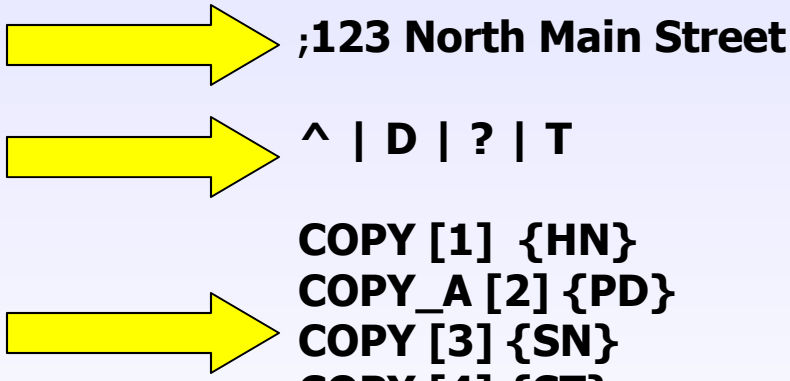
THREE	3	C
FOUR	4	C
FIVE	5	C
SIX	6	C
SEVEN	7	C
EIGHT	8	C
NINE	9	C
TEN	10	C
ELEVEN	11	C
TWELVE	12	C
THIRTEEN	13	C
FOURTEEN	14	C
FIFTEEN	15	C
SIXTEEN	16	C
SEVENTEEN	17	C
EIGHTEEN	18	C
NINETEEN	19	C
E	E	D
EAST	E	D
N	N	D
NO	N	D
NORTH	N	D
NE	NE	D
NORTHEAST	NE	D
NORTHWEST	NW	D
NW	NW	D
S	S	D
SO	S	D
SOU	S	D
SOUTH	S	D
SOUTN	S	D
SE	SE	D

Example of some columns as they appear in the classification table

Files involved with the Geocoder (cont.)

.pat ~ Pattern File (binary)
.xat ~ editable pattern file

The .pat is a binary file which is used to define pattern rules and actions for an address. The .xat if the file you can use to edit and then recompile to create a new .pat



Customizing the Geocoder

IMPORTANT:

BEFORE TWEAKING ANY GEOCODING FILES, MAKE BACKUKPS OF YOUR ORIGINALS.

This is easily done by making a backup copy of the c:\program files\ArcGIS\Geocode directory.

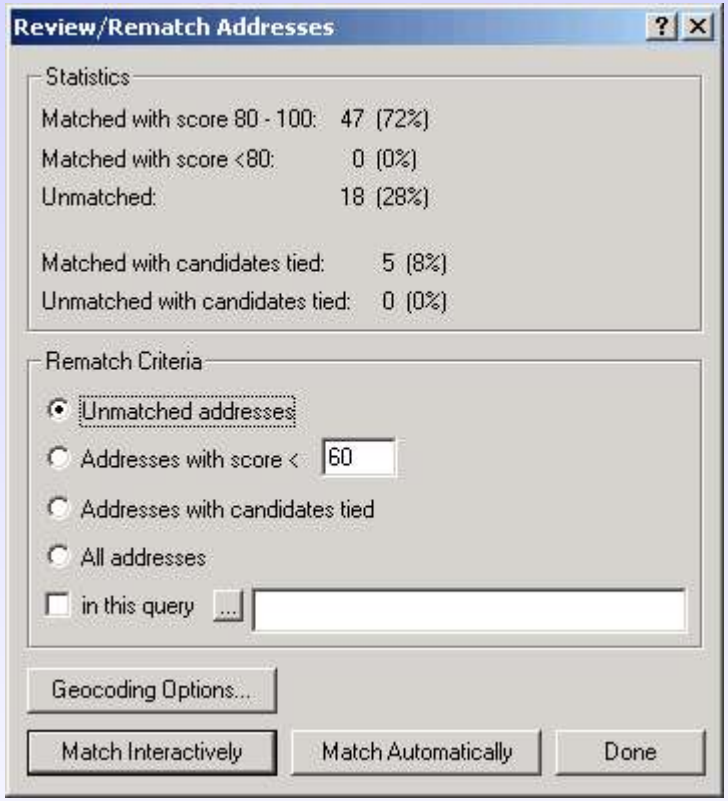
Customizing the Geocoder

Dealing with Sonoma County Street Issues:

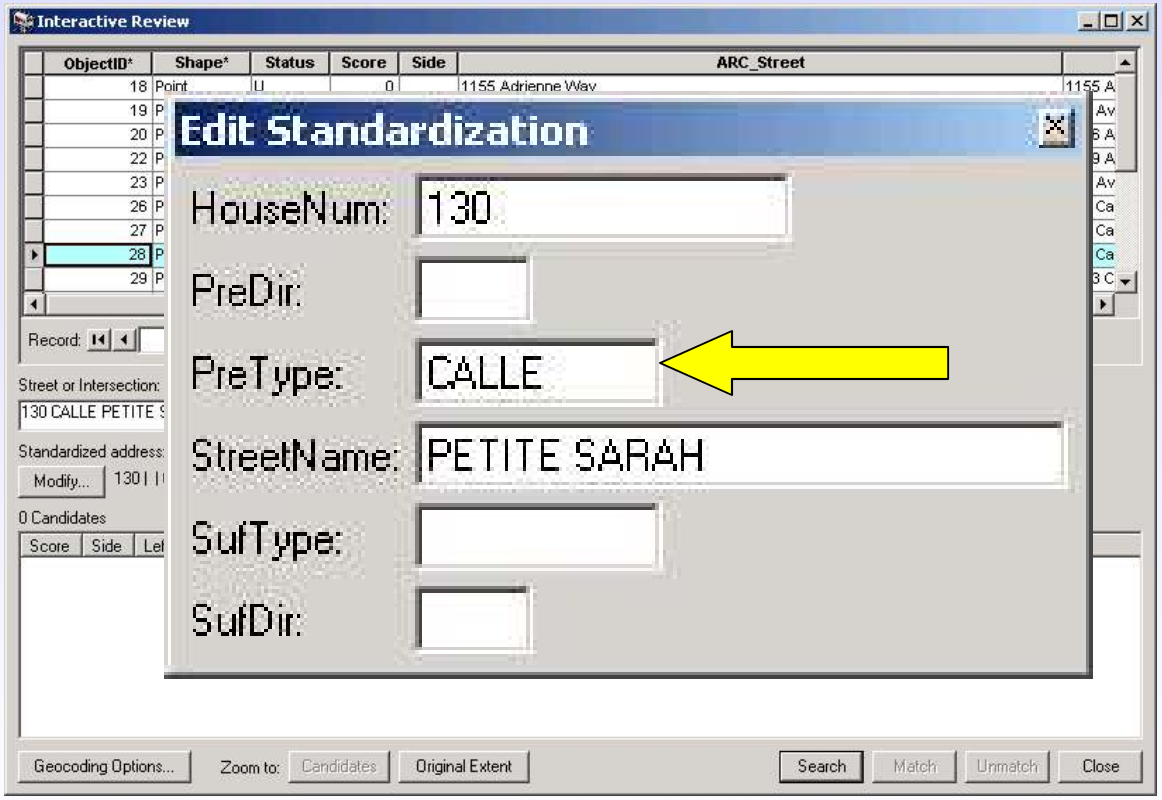
- Dealing with Spanish names.
- Dealing with “SAINT”.
- Dealing with “HWY” in the street name.

Part 3: The Geocoding Developer Kit

Plain Old Vanilla Geocoder



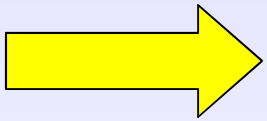
Match = 72%



Part 3: The Geocoding Developer Kit

Modify .cls file to deal with Spanish Street Types

Edit: C:\Program Files\ArcGIS\Geocode*.cls

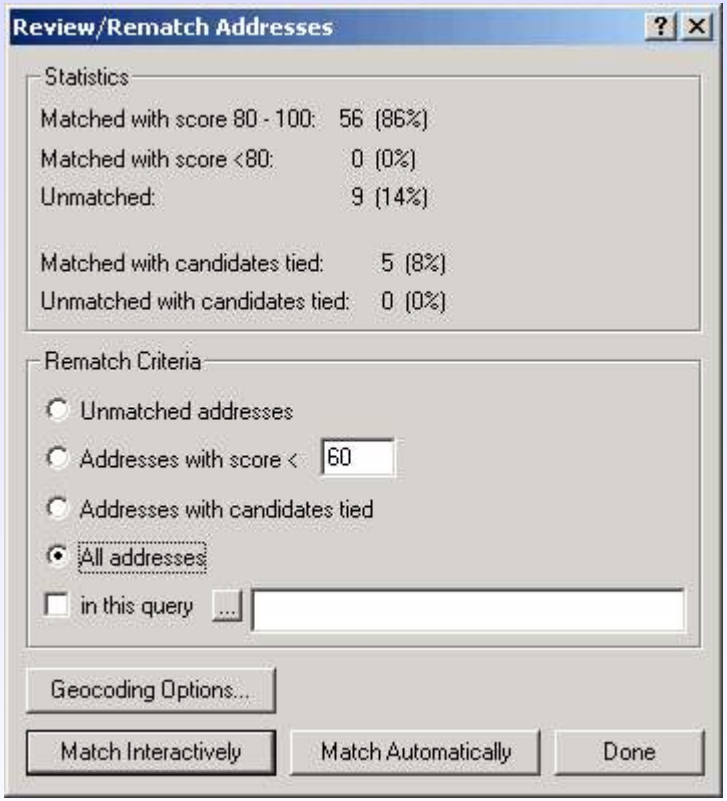


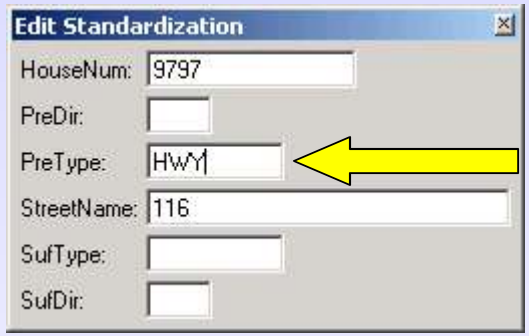
```
BLFS        BLFS      T
BYP         BYP       T
BYPA     BYPA    BYP       T
BYPAS       BYP       T
BYPASS      BYP       T
;CAMI       CAMINO    T
;CAMINO       CAMINO    T
CI          CIR       T
CIR         CIR       T
CIRC        CIR       T
CIRCLE       CIR       T       850.0
CORCLE       CIR       T
CR          CIR       T
```

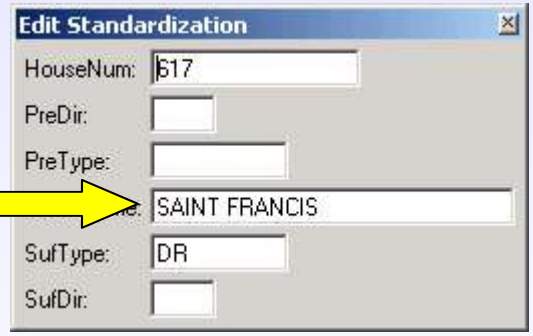
Part 3: The Geocoding Developer Kit

Super Charged Geocoder

Match = 86%



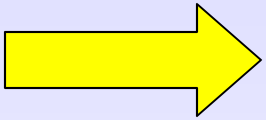




Part 3: The Geocoding Developer Kit

Modify .cls file to deal with "Saint"

Edit: C:\Program Files\ArcGIS\Geocode*.cls



STA	STATION	A
STRM	STREAM	A
SAINT	ST	A
TOWNHSE	TOWNHOUSE	A
TWNHOUSE	TOWNHOUSE	A
TWNHSE	TOWNHOUSE	A

Part 3: The Geocoding Developer Kit

Modifying the .xat file

- Install the Geocoding Developer Kit
- Modify the files found in c:\Program Files\ArcGIS\GDK.
- Use the Standardize Editor to verify your modifications.
- Once modified and compiled, move the new .pat to c:\Program Files\ArcGIS\Geocode

Modifying the .xat file (cont.)

- Add DEBUG & OUTFILE to the .stn file in order to use the STANDARDIZER EDITOR.

Modifying the .xat file (cont.)

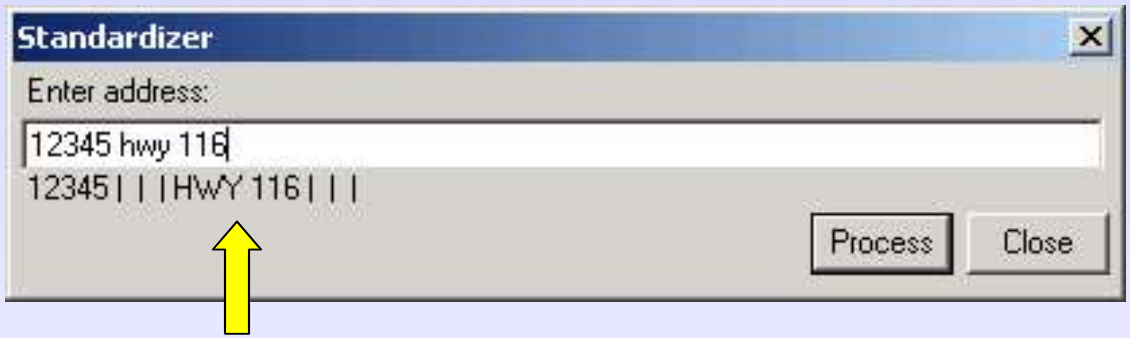
- Open the .xat file using a text editor and add your modifications.
- May require you to comment out conflicting patterns.
- Use ENCODE.PAT to recompile binary .pat.

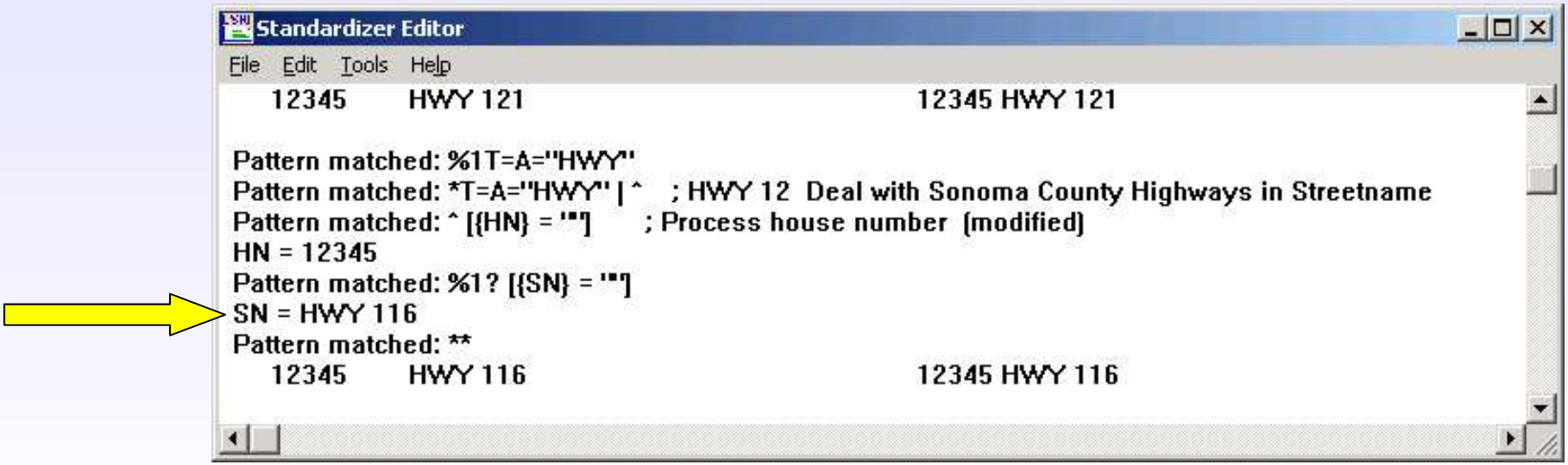
```
\SUB HIGHWAY2
*T=A="HWY" | ^    ; HWY 12  Deal with Sonoma County Highways in Streetname
COPY_A [1] temp
CONCAT " " temp
CONCAT [2] temp
RETYPE [1] ? temp
RETYPE [2] 0
RETURN
```

```
*T=A="HWY" | ^ | D   ; HWY 116 N  Deal with Sonoma County Highways in
Streetname
COPY_A [1] temp
CONCAT " " temp
CONCAT [2] temp
COPY_A [3] {SD}
RETYPE [1] ? temp
RETYPE [2] 0
RETYPE [3] 0
RETURN
```

Modifying the .xat file (cont.)

- Use Standardizer Editor to verify correct syntax.



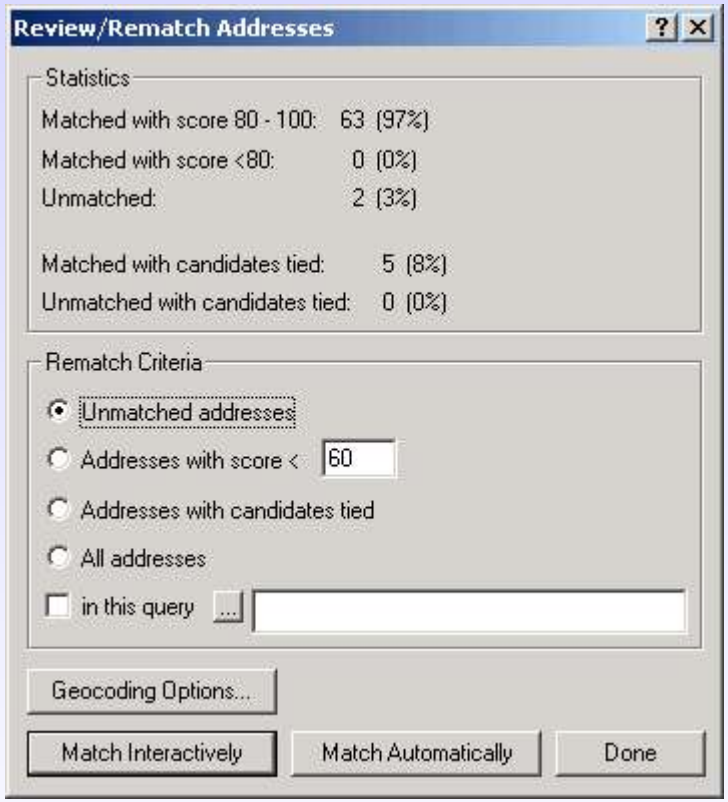


Modifying the .xat file (cont.)

- Move the newly compiled .pat to the geocoding directory c:\program files\arcgis\geocode and overwrite existing .pat
- Re-geocode and see what you get.

Part 3: The Geocoding Developer Kit

Geocoder On Steroids



Match = 97%

Conclusion

- Try and see if your client can give you the address data in a parsed format.
- Make a list of repeating issues.
- Find easy ways to clean up these issues by experimenting with VB Functions and the “Field Calculator” dialog box.
- Download the GDK and read through the documentation. See if it might help you resolve recurring issues.
- Cruise the forums or ESRI support for help & hints.
- Feel free to contact either Michael or myself if you have questions.

Michelle Mestrovich mmestrov@sonoma-county.org
Michael Hansen mhansen@sonoma-county.org