

A Geospatial Information Integration Approach to Building High-Resolution Gazetteers

Presented to: 10th Annual IMSC Student Conference
Los Angeles, CA, April 14th 2006

Daniel W. Goldberg
GIS Research Laboratory
Department of Computer Science
University of Southern California
<http://uscgislab.net>

Outline

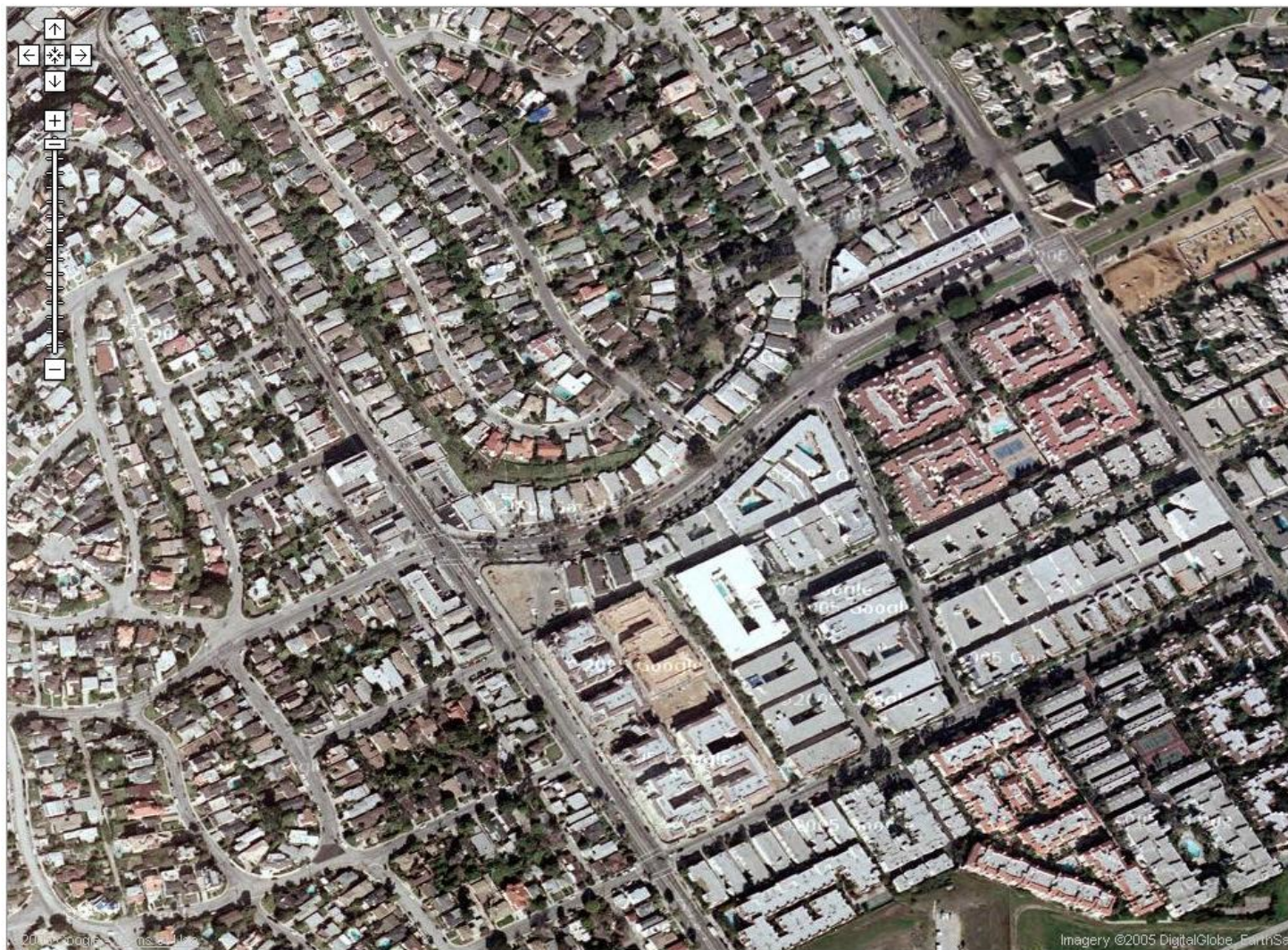
- Motivation and Problem Description
- Data Sources & Methods
- Results
- Future Work

Motivation

- Consider this image
- El Segundo, CA 90245
- What do we see?

Urban Area
Buildings

Need detailed
spatial models



Gazetteers

Geographic data structures specifically designed to maintain this information

- Name(s)
- Type(s)
- Footprint(s)

Problem

- There is a need for highly detailed gazetteers
 - Feature identification in spatial models
 - Emergency management
 - Named entity recognition (NER) in text
 - Georeferencing
 - Access to information by placename
 - Digital libraries
- But there are few around, and they are not highly detailed
 - USGS GNIS, NGA GeoNames, UCSB ADL



"The suspect entered *McDonalds* on 3rd Street at 5:00 pm"

"All books and images about the Roosevelt Hotel"

The Information is Available

- The information to make these things possible is out there
 - Online Phone books
 - Online Property Tax sites
 - Etc.
- **We can extract and integrate information from multiple sources to build detailed regional gazetteers**

Feature Generation



Data Source – Zip+4 Files

- Published by the USPS
 - List street segments
 - List valid addresses per segment

ID, Pre, Name, Suffix, Post, Start, End, ...
100, E, Maple, Ave, N, 700, 798, Even, ...
101, E, Maple, Ave, N, 701, 899, Odd, ...
102, E, Maple, Ave, N, 800, 898, Even, ...
103, E, Maple, Ave, N, 801, 899, Odd, ...
104, E, Maple, Ave, N, 900, 950, Even, ...

Data Source – Assessor

Los Angeles County
Office of the Assessor

Records for this property are kept at the West District Office

5055014012

ELLENDALE
LOS ANGELES
ORCHARD

Copyright - LA Assessor

108ft

Property Information

Assessor's Id. Number	5055-014-012
Site Address	2652 ELLENDALE PL LOS ANGELES CA 90007
Property Type	Multi-Family Residential
Region / Cluster	09 / 09441
Tax Rate Area (TRA)	00210

[Click Here to View Assessor's Map](#)

Recent Sale Information

Latest Sale Date	
Indicated Sale Price	

[Search For Recent Sales](#)

2004 Roll Values

Recording Date	05/13/1987
Land	\$335,057
Improvements	\$508,083
Personal Property	\$1,600
Fixtures	\$0
Homeowners' Exemption	\$0
Real Estate Exemption	\$0
Personal Property Exemption	\$0
Fixture Exemption	\$0

[Click Here for 2004 Annual Taxes](#)

Legal Description

LOT 41 ELLENDALE PLACE AND LOT 41 1/2 BOWKER'S ADD TO ELLENDALE PLACE

Building Description(s)

Improvement 1	
Square Footage	10,402
Year Built / Effective Year Built	1963 / 1963
Bedrooms / Bathrooms	24 / 24
Units	12

Provides:

- Information about features
 - Verification of existence
 - Type (Residential/Commercial)

- Interface is queryable by
 - Address

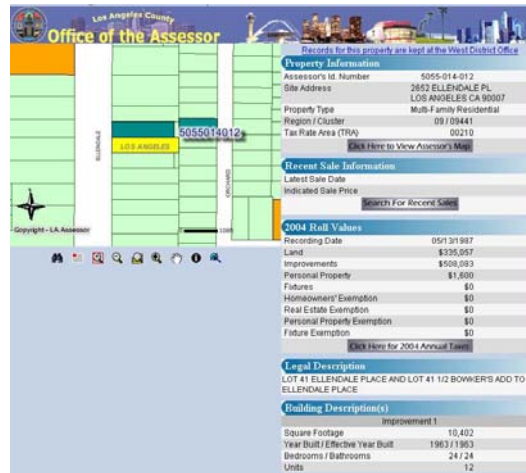
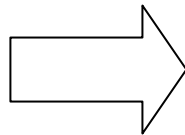
<http://assessormap.co.la.ca.us/mapping/viewer.asp>

Feature Generation

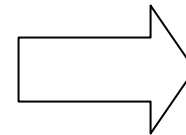
- USPS Zip+4 files
 - Can generate all addressable addresses
- Los Angeles County Assessor (LACA) Site
 - Can verify addresses which do exist

id, 12231, Maple,Ave ,
 id,
 id,
 id,
 id,
 id,
 id,
 id,
 id,
 id,
 id,

Zip+4



LACA



id, 12231, Maple,Ave ,

 id,

 id,
 id,
 id,

 id,

 id,
 id,

Real Addresses

Feature Types

- Now that we have features
 - What are their names and types?
- All we know so far about a feature
 - It is a building
 - Commercial
 - Residential
- Can we get more information?

Feature Types Cont.

- Supplement with phone book information
 - Yellow Pages (Commercial)
 - White Pages (Residential)

Data Source – Superpages

Yellow Pages

[Home](#) → [Category Browser](#) → Home & Garden

Home & Garden

<p>Appliances Dealers, Service & Repair, ...</p> <p>Cleaning Equipment & Supplies Carpet Cleaning Supplies, Cleaning Supplies, ...</p> <p>Cleaning Services Housecleaning, Laundries, Sewer & Drain Cleaning, ...</p> <p>Domestic Services Child Care, Maids & Butlers, ...</p> <p>Home Furnishings Carpet & Rug Dealers, Furniture Stores, ...</p> <p>Home Improvement & Maintenance Interior Designers, Kitchen & Bathroom Remodeling, Pest Control, ...</p> <p>Lawn & Garden Garden Centers, Landscaping, ...</p> <p>Packaging & Shipping Materials</p> <p>Retail</p> <p>Pets & Animals Animal Hospitals, Pet Grooming & Boarding, Pet Shops, ...</p>	<p>Plants</p> <p>Security Systems & Services Burglar Alarms, Keymakers, Locksmiths, ...</p> <p>Television Service Providers Cable TV, Satellite TV, ...</p> <p>Utilities Electric Companies, Garbage Removal, Gas Companies, ...</p> <p>See Also: Construction & Contractors Home & Garden Stores Moving & Storage Office Furniture, Equipment, & Supplies</p>
--	---

[Up to top level](#)

Can't find it? Try the [Alphabetical List](#)

Provides:

- Information about features
 - Name
 - Category
 - Address
 - Phone Number
- Interface is queryable via
 - Type Category
 - Location (Zip Codes)

<http://www.superpages.com>





Data Source – Switchboard

FIND A BUSINESS | FIND A PERSON | SEARCH BY PHONE | WEB SEARCH | AREA & ZIP CODES





First Name: Last Name City State List
 smith el segundo ca

smith in el segundo, ca 90245
 28 people found (1-10 shown) | [Search Public Records](#) [Help](#)





Smith, A [Email, Maps and What's Nearby](#)SM
 El Segundo, CA 90245
 (310) 322-0836
[Update/Remove this listing](#)
 Updated Address & Phone Number Found for A Smith
[Instant Background Check for A Smith](#)

 [Public Records](#)  [Address History](#)  [Search by SSN](#)  [Background Check](#)





Smith, Alfred [Email, Maps and What's Nearby](#)SM
 311 W Palm Ave,
 El Segundo,
 CA 90245-2267
 (310) 322-6621
[Update/Remove this listing](#)
 Updated Address & Phone Number Found for Alfred Smith
[Instant Background Check for Alfred Smith](#)

 [Public Records](#)  [Address History](#)  [Search by SSN](#)  [Background Check](#)

Smith, B [Email, Maps and What's Nearby](#)SM
 El Segundo, CA 90245
 (310) 640-8567
[Update/Remove this listing](#)
 Updated Address & Phone Number Found for B Smith
[Instant Background Check for B Smith](#)

 [Public Records](#)  [Address History](#)  [Search by SSN](#)  [Background Check](#)

Smith, Brandy [Email, Maps and What's Nearby](#)SM
 770 W Imperial Ave,
 El Segundo,
 CA 90245-2053
 (310) 615-1944
[Update/Remove this listing](#)
 Updated Address & Phone Number Found for Brandy Smith
[Instant Background Check for Brandy Smith](#)

 [Public Records](#)  [Address History](#)  [Search by SSN](#)  [Background Check](#)

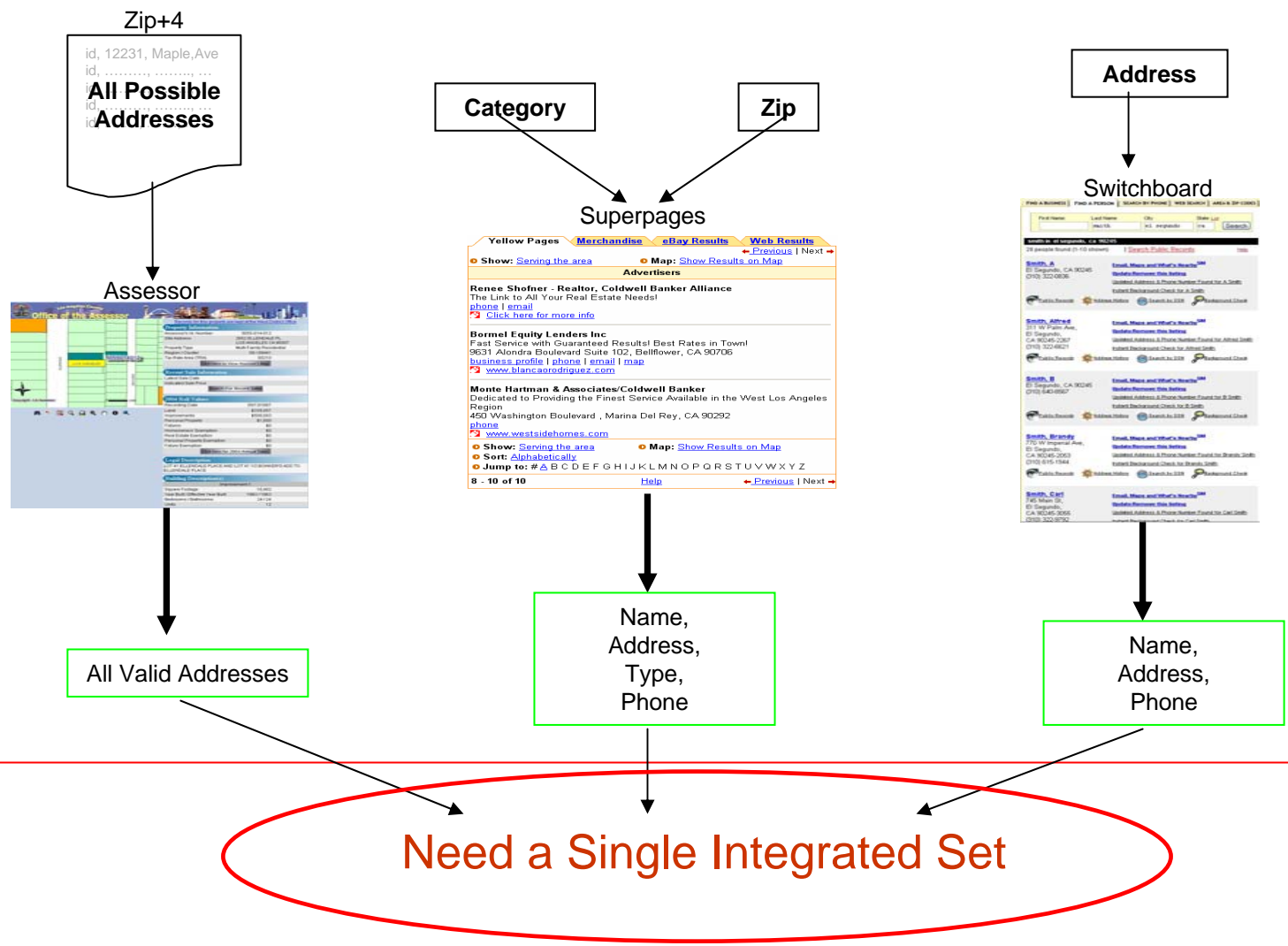
Smith, Carl [Email, Maps and What's Nearby](#)SM
 745 Main St,
 El Segundo,
 CA 90245-3055
 (310) 322-9792
[Update/Remove this listing](#)
 Updated Address & Phone Number Found for Carl Smith
[Instant Background Check for Carl Smith](#)

<http://www.switchboard.com>

Provides:

- Information about features
 - Type (Residential)
 - Owner Name
 - Address
 - Phone Number
- Interface is queryable by
 - Address

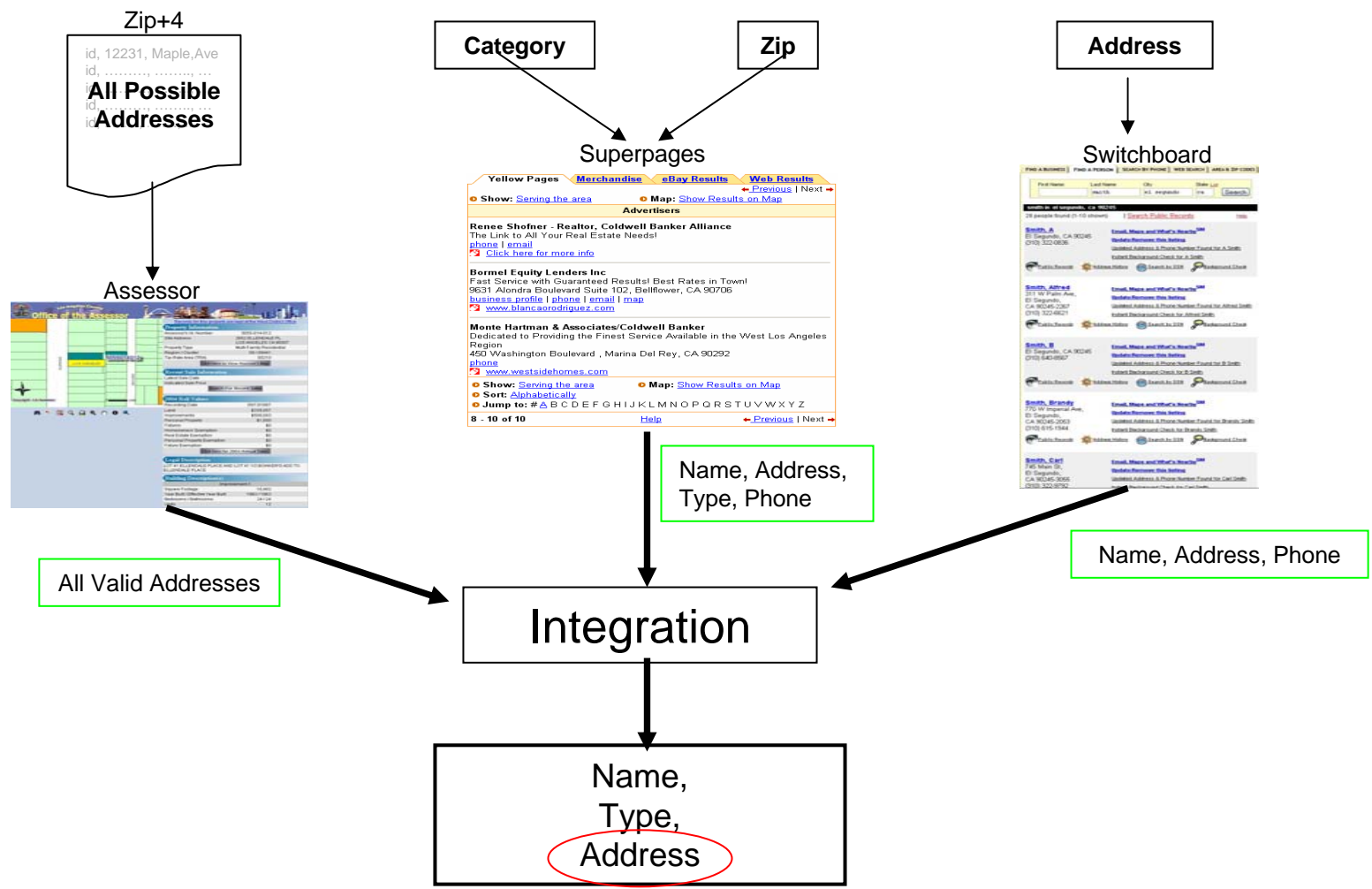
Data Sources & Methods So Far



Integration Process

- Normalization
 - Put all records into a consistent format
- Record Linkage
 - Eliminate duplicate records
 - Merge attributes from multiple records

Method So Far



Need a footprint

Footprint Generation

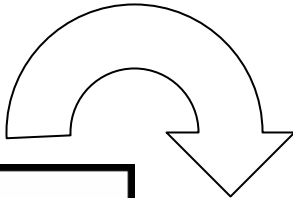
- Generate points as first approximation

Geocode

Test

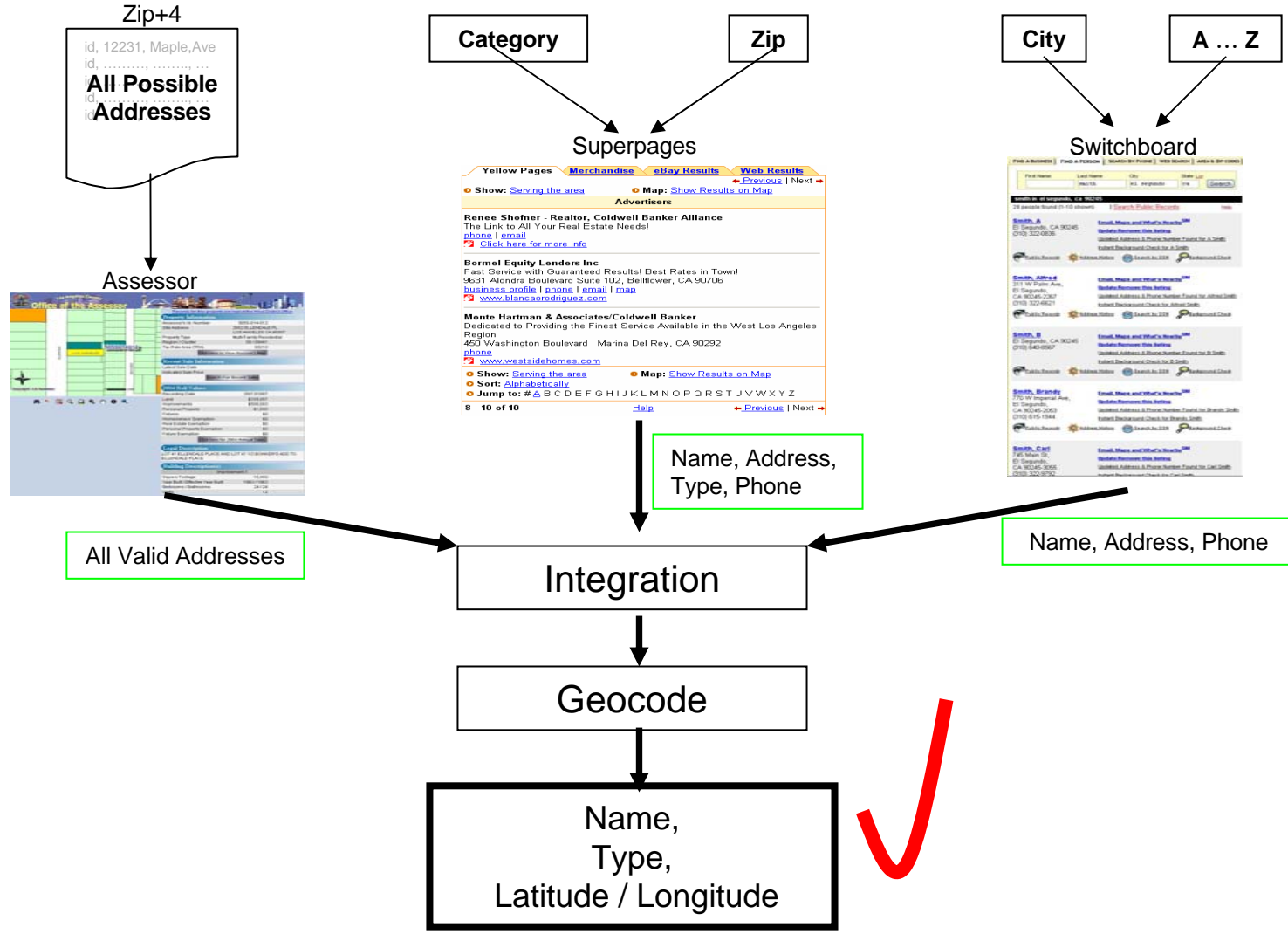
To test the operation using the HTTP GET protocol, click the 'Invoke' button.

Parameter	Value
streetaddress:	<input type="text" value="2652 ellendale place"/>
city:	<input type="text" value="los angeles"/>
state:	<input type="text" value="ca"/>
zipstr:	<input type="text" value="90007"/>

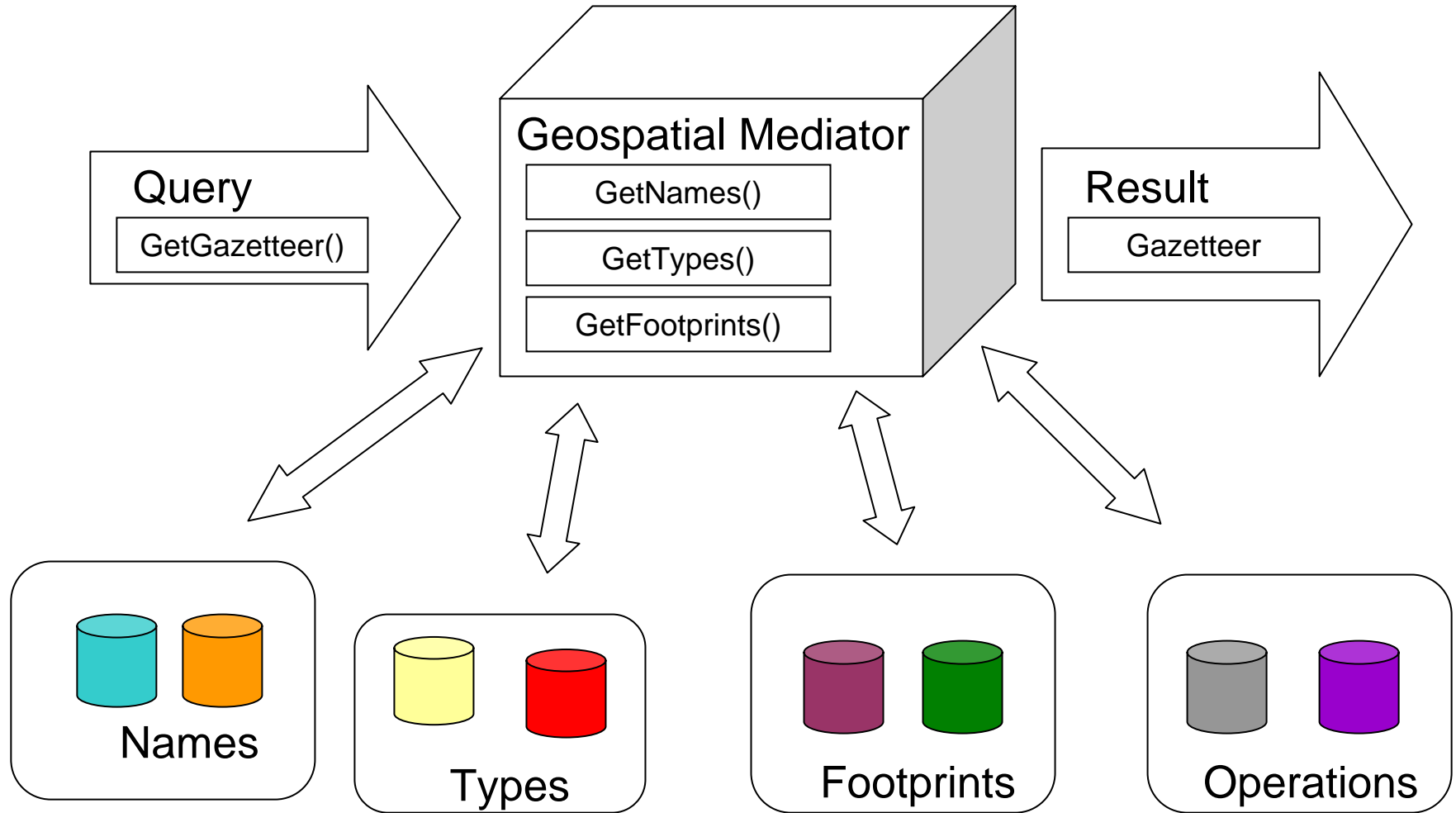


```
<geopoint>
  <lat>34.005206532674805</lat>
  <lon>-118.47605547145031</lon>
  <geoerror/>
  <errorbounds>0.0015020408163265305</errorbounds>
  <errorboundsunit>Decimal Degrees</errorboundsunit>
</geopoint>
```

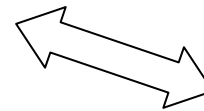
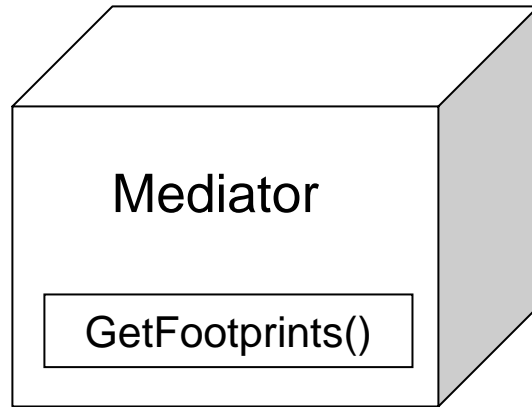
Complete Gazetteer



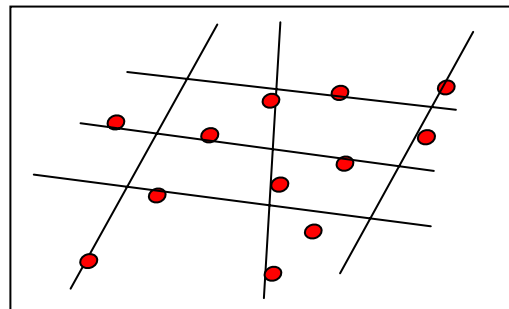
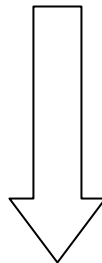
Iterative Improvement



Example: Improve Footprints



Linear Interpolation
Geocode Operator



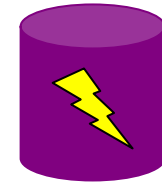
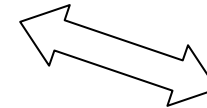
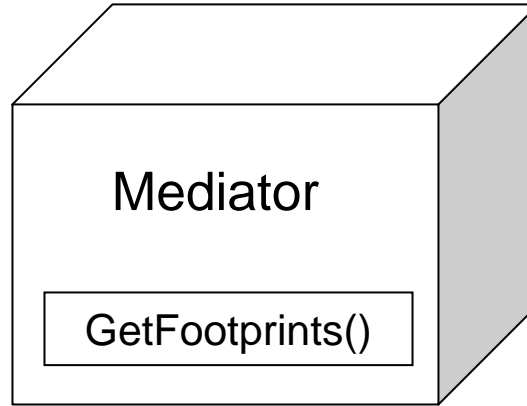
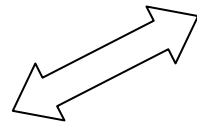
Point
Footprints

Relation
Constraints:
Extent=US
Accuracy=0

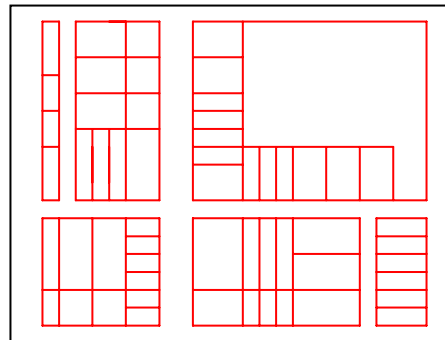
Example: Improve Footprints



Raster Parcel Source



Vector Extraction Operator



Vector Footprints

Relation
Constraints:
Extent=LA
Accuracy=1

Example: Improve Footprints



Results

- Comparison to Existing Gazetteers
- Ground Truth Evaluation

Comparison – Existing Gazetteers

- UCSB Alexandria Digital Library (ADL): GNIS + GeoNames + TGN
- Los Angeles Comprehensive Bibliographical Database (LACBD)

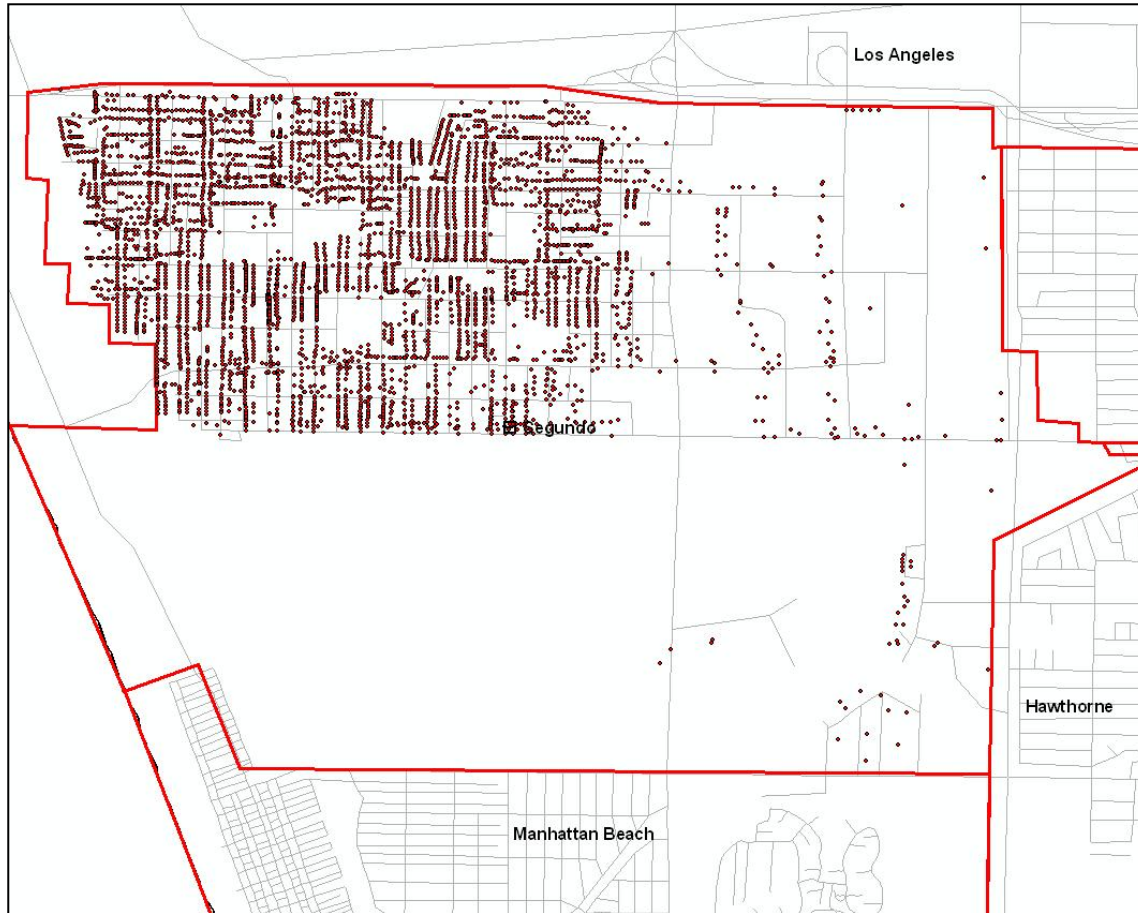


ADL (22 Features)



LACBD (11 Features)

Comparison – Automatically Generated Gazetteer



AGG (5,046 Features)

Feature Comparison

Feature Type	Feature Name	ADL	LACBD	AGG
B	El Segundo City Hall	X	X	X
T	Old Town Music Hall		X	
E	Center Street Elementary	X	X	X
E	St Johns Luth Child Dev	X		
E	St Anthony Elementary		X	X
E	Richmond Street Elem		X	X
E	Webster University		X	X
E	Arena High	X		
E	El Segundo High	X	X	X
E	El Segundo Mid	X	X	X
H	Chevron Refinery	X		
H	Airport Towers Number 1	X		
L	El Segundo Public	X	X	X
PK	Holly Valley	X		
PK	Candy Cane	X		
PK	El Segundo		X	
PK	Dockweiler Beach St	X		
PK	Kansas	X		
PO	El Segundo	X		X
R	Pacific Baptist	X		X
R	St Andrews Church	X		X
R	New Mt Calv Mis. Bap	X		
R	Temple Rodeph Shalom	X		X
R	El Segundo Christian	X		X
R	El Segundo Foursquare	X		X
S	El Segundo Golf Course	X		X
	Ratio of total (superset)	21/26	10/26	15/26
	Recall % (superset)	81%	38%	58.00%
	Ratio of total (buildings)	14/18	9/18	14/18
	Recall % (buildings)	78%	50%	78%

B building, T theater, E educational facility, H heliport, L library, PK park, PO post office, Religious site, and S sports facility

Ground Truth

- Surveyed 43 street segments of different types by walking the blocks
 - Residential (R)
 - Commercial (C)
 - Industrial (I)
 - R/C
 - I/C
- Compared
 - What the AGG said was there
 - What was actually there

Accuracy

- Location

Precision:

Of the number of features we extracted, how many actually existed?

Recall:

Of the number of features in existence, how many did we get?

- Name / Type

Precision:

Of the names/types we extracted, how many were correct?

Recall:

Of the possible names/types in existence, how many did we get?

Superpages

Type	Name/Type Precision	Name/Type Recall
I	1.00	0.68
I/C	1.00	0.83
C	0.95	0.88
C/R	0.92	0.87
R	N/A	N/A
Overall	0.96	0.81

Type	Location Precision	Location Recall
I	0.90	0.39
I/C	0.88	0.78
C	0.75	0.99
C/R	0.94	0.63
R	N/A	N/A
Overall	0.72	0.58

- Very accurate name/type
- Overestimated features (shopping centers, downtown blocks, etc.)

Switchboard

Type	Precision	Recall
I	0.10	0.08
I/C	0.00	0.00
C	0.67	0.17
C/R	0.78	0.45
R	1.00	0.44
Overall	0.49	0.22

- Underestimated Features (cell phones, unlisted numbers, etc.)

Assessor

Type	Precision	Recall
I	0.86	0.71
I/C	0.94	0.84
C	0.84	0.72
C/R	0.86	0.89
R	1.00	0.91
Overall	0.89	0.81

- Very accurate features
- Not much type information

Automatically Generated Gazetteer

After integrating
the three sources

Type	Precision	Recall
I	0.94	0.80
I/C	0.92	1.00
C	0.64	1.00
C/R	0.79	0.99
R	0.99	0.91
Overall	0.85	0.94

- Very accurate features
- Very detailed type information

Future Work

- Automatically merge feature type hierarchies
 - Our types are very detailed and would not translate directly/automatically into existing hierarchies
 - How can we automatically augment existing FTT to accommodate highly detailed feature types?
- Sub-Parcel Footprints
- More feature types (non-manmade)
 - New Sources
- Automatically learn and extract the contents of online raster maps

The End

- Thanks to the gang!
 - John P. Wilson, advisor – USC GIS Lab
 - Craig A. Knoblock, advisor – USC/ISI
 - Snehal Thakkar, mediator – USC/ISI
 - Yao-Yi Chiang, vector extraction – GeoSemble
- Questions?
- Comments?