BEDA Accelerator: Washington, D.C.

February 5, 2019
Agenda

- Introductions
- Overview of UBID generation methodology & data requirements
- Example of UBID Generation: UBID Demonstrator & Drawing Tool
- Analysis of DC UBIDs
- Identify viable datasets for integration of DC UBIDs
- Discussion of Implementation Strategy & IT Requirements
Problem Statement

The lack of a standardized way to identify buildings makes it difficult to accurately associate data with a specific facility, creating a barrier to effective asset management, research, and analysis.

Where the current address system breaks down:

- Different address abbreviation, e.g., st or street; ave or avenue; apt or #;
- Simple misspellings or incorrect addresses
- Large buildings with multiple entrances and possibly multiple addresses
Problem Statement

The lack of a standardized way to identify buildings makes it difficult to accurately associate data with a specific facility, creating a barrier to effective asset management, research, and analysis.

How this looks for benchmarking:

<table>
<thead>
<tr>
<th>Assessor Database</th>
<th>CoStar Data</th>
<th>Covered Buildings List</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Building</th>
<th>Assessor</th>
<th>CoStar Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bldg 1</td>
<td>123 Main St</td>
<td>123 Main Street</td>
</tr>
<tr>
<td>Bldg 2</td>
<td>456 5th Street</td>
<td>789 Central Ave.</td>
</tr>
<tr>
<td>Bldg 3</td>
<td>Tax ID 5578</td>
<td>Tax Lot 5577</td>
</tr>
<tr>
<td>Bldg 4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bldg 5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Problem Statement

The lack of a standardized way to identify buildings makes it difficult to accurately associate data with a specific facility, creating a barrier to effective asset management, research, and analysis.

How bad is this problem?

- An analysis of 800k buildings in Houston, TX yielded an 80% match rate based on address from pre-cleansed datasets; an additional 20-30 person hours were required to reach a 95% match rate using fuzzy matching algorithms and hand matching.
- Even small towns like Department of Planning in South Burlington, VT estimates 2 hours/month go into developing data workarounds for bad matches.
- According to Ecotope and SF Department of Environment, average match rates are 50-60%. UBIDs could save days to weeks of manual data matching efforts.

Acknowledgement: UC Berkeley Student Consulting & Research Group
Solution: A Natural Key for Buildings

Flexible fuel vehicles can be identified by the 2nd, 3rd and 8th digits of the VIN

1HGBH41JXMN109186

2nd and 3rd characters: The Manufacturer

1st character: Where the vehicle was built

4th and 8th characters: Portrait of the vehicle-brand, engine size and type

9th character: Security code that identifies the VIN as being authorized by the manufacturer

10th character: Model year of the car

11th character: Indicates which plant assembled the vehicle

Last 6 characters: Serial number of the vehicle
Solution: A Natural Key for Buildings

87C4VXX7+29R-14-37-11-39
UBID Demonstrator

UBID.PNNL.GOV
Datasets

• Open Data Footprints
  ▪ http://opendata.dc.gov/datasets/a657b34942564a8b06f293cb0934cbd_1
  ▪ 163,467 entries
  ▪ No local ID ("GIS_ID" field empty)

• Energy Benchmarking 2016
  ▪ https://doee.dc.gov/publication/2016-building-benchmarking-dataset
  ▪ 1,846 entries
  ▪ pid, dc_real_pid, and pm_pid are local IDs, pid only with no duplicates and value for every entry

• Other datasets used for analysis:
  ▪ Street Centerlines
    ▪ http://opendata.dc.gov/datasets/street-centerlines
  ▪ Address Points
    ▪ http://opendata.dc.gov/datasets/address-points
• 0 invalid geometries from footprints and benchmarking

• 1,608 benchmarking points (UBID₀) matched to footprint UBIDs
  ▪ 238 UBID₀ not matched
  ▪ 191 duplicate UBID₀ created

UBID Matching Process

Benchmarking Point
PID: PM05823132
87C4VXJM+452-0-0-0-0

Footprint (Bounding Box)
87C4VXJM+456-29-5-29-6
UBID Matching Main Issue Overview

- Benchmarking points that represent multiple building footprints and multiple benchmarking points that represent the same building footprint

Some are matched and some are unmatched
Some unmatched points are the “campus” issue on the previous slide.

Others are points that are in the street, and the first step is to differentiate these.

- **Unmatched Points:** UBID$_0$ in Street
  - 5ft buffer: 32 UBID$_0$
  - 10 ft buffer: 66 UBID$_0$
  - 15ft buffer: 71 UBID$_0$

![Map Diagram](image.png)

- Correct Parcel
- Correct Footprint

**2301 11th St**
• Matching to nearest footprint is a quick solution, but there will be false positives (like example in previous slide)

• Best solution is manual review of the ~70 points
  ▪ a little time consuming, but only needs to be done once

• Other solution is matching addresses
  ▪ Not perfect, usually 60-80% success rate, but 60-80% for 5% of database isn’t bad
  ▪ Matching addresses requires some data processing to match the formatting, could be almost as time consuming as manual review

• For future benchmarking, worth making reporter quickly confirm that the geocoded address doesn’t lie in street
One (unmatched) UBID₀ that represents multiple buildings

Example Below:

- Estimated area (with area map tool): 77k * 4 floors = ~308k sqft
- Reported area = 280k sqft
- Conclusion: Benchmarking data represents all buildings in this multi-family housing unit but didn’t match because fell outside bounding box of all footprints

About 140 instances
• Merge all footprints that correspond to the UBID_0 using the “Square” and “Lot” fields
  ▪ Could use either Parcel Lot or Address Point dataset to facilitate the merge
  ▪ Some data processing labor involved
• Worth doing for UBID_0 in street in case they have multiple buildings
One (matched) UBID<sub>0</sub> for multiple buildings

• We assumed there are cases like the unmatched ones, that happened to land within a footprint bounding box – but no way to detect these

• Can be improved in future benchmarking by including critical data to identify these

• Quick estimate (not up to date):
  - 544 UBID<sub>0</sub> that are matched to footprints with multiple addresses in the same lot
  - Even if we can flag the UBID<sub>0</sub> with multiple addresses in same lot, could be difficult to determine if the UBID<sub>0</sub> is only for one in matched to or for all the buildings

Hypothetical example
Multiple UBID\textsubscript{0} for multiple buildings

- Example:
  - 901, 907, and 907 6\textsuperscript{th} St SW
- Benchmark XY all on 907 address
- Area
  - 901 area: 20,450*9 = ~184k
  - 907 area: 17,400*9 = ~157k
  - 3 reported areas (381k, 100k, and 53k) and a tax record of 1M sqft
- The two are very similar architecturally and to the other buildings on the plot
- Conclusion: The taxable area (1M sqft) represents all buildings on the property and the three benchmarking are some combination of sub spaces
Solution

- Detection: Duplicate UBID\(_0\) that also have multiple buildings on parcel
  - Some labor involved in this detection process
- Impossible to know, even with manual inspection, what benchmarking entries represent which spaces
- Question for DC: What would be the appropriate solution for this example?
  - Idea for future: mark as not compliant because impossible to know which spaces are being benchmarked
  - Idea 1: merge footprints and create one UBID for the parcel
    - match all benchmarking entries
    - Aggregate benchmark data and match only one entry
Multiple UBID₀ for single building

- Example: 203 N St SW
- Calculate Area in Google: 21,700*8 = ~173k sqft
- Reported Area:
  - 115,323
  - 23,876
  - 35,992
  - Total: 175k
- Conclusion: multiple spaces benchmarked separately
• Similar to previous case – impossible to know which spaces in the building are being benchmarked

• Question for DC: What would be the appropriate solution for this example?
  ▪ Idea 1: no action (i.e. match all benchmarking UBID\textsubscript{0} to one footprint UBID)
  ▪ Idea 2: Aggregate data and match only one entry
False Matching: Incorrect Location

• Example:
  - 1230 S Capitol SE
  - 1263-1265A 1st St SE

• UBID₀ location and use type match first address, second address is a few blocks away

• Conclusion: Incorrect coordinates entered for second address
Solution

• Garbage in – garbage out?
• Potentially flag (with address?) for revision
  • Just looking through this will not be clean because addresses can vary slightly in zipcode, address number, street format
• For future: when benchmarking ask reporters to confirm geolocation on map (5 seconds)
• If benchmarking geolocation isn’t well aligned with center of the bounding box of the footprint, there is a chance it could be falsely matched to neighboring footprints
We can’t definitively find or fix every false positive, but it’s possible to look at a subset of data to extrapolate our success rate.

For every benchmarking UBID\textsubscript{0} that intersects with multiple bounding boxes, compare the closest and 2\textsuperscript{nd} closest centroids. If the distances are close (say within ~20%) we can flag these for manual review.

Another possibility: Compare distance between UBID\textsubscript{0} and matched centroid to the area of the bounding box or the percent area increase between the footprint and the bounding box.

Another possibility: Look at edge cases with large percent area increase between footprint and bounding box.
Other Possible Cases

• Other cases that may be worth investigating, but would require more time to detect these
  1. Multiple UBID\textsubscript{0} with different location but on same building
  2. Multiple UBID\textsubscript{0} with different exact location on same property with multiple buildings
  3. One UBID\textsubscript{0} represents subsection of building
Thank you
Engagement & Implementation

• What do we need?
  – Technical Leads – who are the folks programming and supporting your database infrastructure?
  – Two+ databases – where do you want to see UBIDs incorporated and matched to each other?

• Process:
  – Mark will Skype/WebEx/etc. in with your technical team to understand your database architecture
  – Using the tooling developed at PNNL, UBIDs can be added into your existing systems. In the process, Mark can develop a replicable process for use by additional stakeholders in your organization.
Next Steps

• Timeline
• Points of Contact
• Relevant Datasets for UBID Integration
• Desired Outcomes & Metrics for Success