

# Museum Data Challenge

Northern Virginia



June 2, 2013

Dylan Barlett  
[@dylanbarlett](https://twitter.com/dylanbarlett)  
[dylanbarlett.com](http://dylanbarlett.com)

# Background

- Institute of Museum and Library Services
  - Independent federal agency
  - Main source of federal support for libraries and museums
- Museums Count
  - “First national, Web-based census of the museum field”
  - Goal: comprehensive, reliable data about the size, distribution, and scope of the museum sector in the United States

# Museum Counts Dataset

- 35K records, varying completeness
- Fields
  - Institution Name (all-caps)
  - Address
  - Location (geocoded)
  - DUNS
  - EIN
  - Misc FIPS & Census

# IMLS Challenges

1. Develop tools to verify and clean existing name and location data.
2. Help build better museum records by filling-in missing data fields.
3. Augment the museum records with institutional URLs and other details that would be helpful for public way finding.

# Traditional Options

**amazon**mechanical turk  
beta Artificial Artificial Intelligence



# Automated Reconciliation Sources



Instant  
API



Facebook  
Places



Exempt  
Organizations  
& Form 990

1.8M records

# Stack

fuzzywuzzy

facebook-sdk

duckduckgo2



~350 SLOC



Declarative ORM



1.8M rows

~25 SLOC Bash script to download and unzip CSV data

# Results

| imls_name                              | legal_name                                 | fb_name                               | fb_id           | url   |
|--|--|---------------------------------------|-----------------|---|
| HAGGIN MUSEUM                          |  | The Haggin Museum                     | 124208817621223 |   |
| NORTH LAKE TAHOE HISTORICAL SOCIETY    |  | North Lake Tahoe Historical Society   | 123837933270    |   |
| TULARE HISTORICAL MUSEUM               |  | Tulare Historical Museum              | 130492753687040 |   |
| HAYS ANTIQUE TRUCK MUSEUM              |  | Hay's Antique Truck Museum            | 113961838634163 |   |
| ALASKA AVIATION HERITAGE MUSEUM        | Alaska Aviation Heritage Museum            | Alaska Aviation Heritage Museum       | 117214631631030 | <a href="http://www.alaskaairmuseum.org">http://www.alaskaairmuseum.org</a>   |
| ALASKA BOTANICAL GARDEN                | Alaska Botanical Garden Inc.               | Alaska Botanical Garden               | 147172085306877 | <a href="http://www.alaskabg.org/">http://www.alaskabg.org/</a>               |
| ALASKA DEPARTMENT OF NATURAL RESOURCES |  |                                       |                 | <a href="http://www.dnr.state.ak.us/">http://www.dnr.state.ak.us/</a>         |
| ALASKA GEOGRAPHIC ASSOCIATION          | Alaska Geographic Association              | Alaska Geographic Association         | 162936800398947 |   |
| ALASKA MINING HALL FAME FOUNDATION     | Alaska Mining Hall of Fame Foundation      | Alaska Mining Hall of Fame Foundation | 100864353292128 |   |
| ALASKA NATIVE HERITAGE CENTER          |  |                                       |                 | <a href="http://www.alaskanative.net/">http://www.alaskanative.net/</a>       |
| ALASKA RAPTOR CENTER                   |  | Alaska Raptor Center                  | 125452720857140 | <a href="http://www.alaskaraptor.org/">http://www.alaskaraptor.org/</a>       |
| ALASKA SEALIFE CENTER                  |  | Alaska SeaLife Center                 | 98827926470     | <a href="http://www.alaskasealife.org">http://www.alaskasealife.org</a>       |
| ALASKA STATE MUSEUM                    |  | Alaska State Museum                   | 115538301809692 |   |
| ALASKA ZOO                             | Alaska Zoo                                 | The Alaska Zoo                        | 136848033588    | <a href="http://www.alaskazoo.org">http://www.alaskazoo.org</a>               |
| ALFRED STARR NENANA CULTURAL CENTER    |  | Alfred Starr Cultural Center          | 439468142759377 |   |
| AMERICAN BALD EAGLE FOUNDATION         | American Bald Eagle Foundation Haines C... | American Bald Eagle Foundation        | 162882710425424 |   |
| ANCHORAGE MUSEUM                       |  |                                       |                 | <a href="http://www.anchoragemuseum.org">http://www.anchoragemuseum.org</a>   |
| CAMPBELL CREEK SCIENCE CENTER          |  | Campbell Creek science center trails  | 225774204113265 |   |
| CENTER FOR ALASKAN COASTAL STUDIES     |  | Center for Alaskan Coastal Studies    | 104160828208    |   |
| CHUGACH ALASKA CORPORATION             |  | Chugach Alaska Corporation            | 106030972794814 | <a href="http://www.chugach-ak.com/">http://www.chugach-ak.com/</a>           |
| CITY MUSEUM                            |  |                                       |                 | <a href="http://www.citymuseum.org/site/">http://www.citymuseum.org/site/</a> |



# Future

- Replace SQLite with MySQL or PostgreSQL
- Tune radius and fuzzy matching thresholds
- REST API for OpenRefine (Werkzeug?)
- More data sources
  - Facebook Graph
  - Google Places
  - [OpenCorporates](#)
  - [EasyPost](#) (address standardization)