Museum Frame Development – A Universe is Comprised of Many Worlds:

Comparing the Efficacy of Web Scraping and Other Approaches to Generating Establishment Lists

Lisa M. Frehill, Jason Enos and Matthew Birnbaum

Office of Digital and Information Strategy

Institute of Museum and Library Services

Effective: 28 February 2018

Outline – Talk & Paper

- Background
    - Frame development – accuracy, uniqueness, and efficiency
    - What is a museum?
    - Children's museums
        - Specific issues
        - Testbed for new methods
- Data and Methods
    - Existing Children's Museum Subset File
    - New entries – two-stage web scraping
- Findings - Accuracy and uniqueness → efficiency indicators
    - Prior methods (IRS 990 mining, lists)
    - New two-stage web scraping (yelp.com and yellowpages.com)
- Conclusions and Next Steps

What is a museum? Common definition

(1) Non-profit (or government)

(2) Organized on a permanent basis for essentially educational or aesthetic purposes

(3) Owns or uses tangible or intangible objects, either animate or inanimate

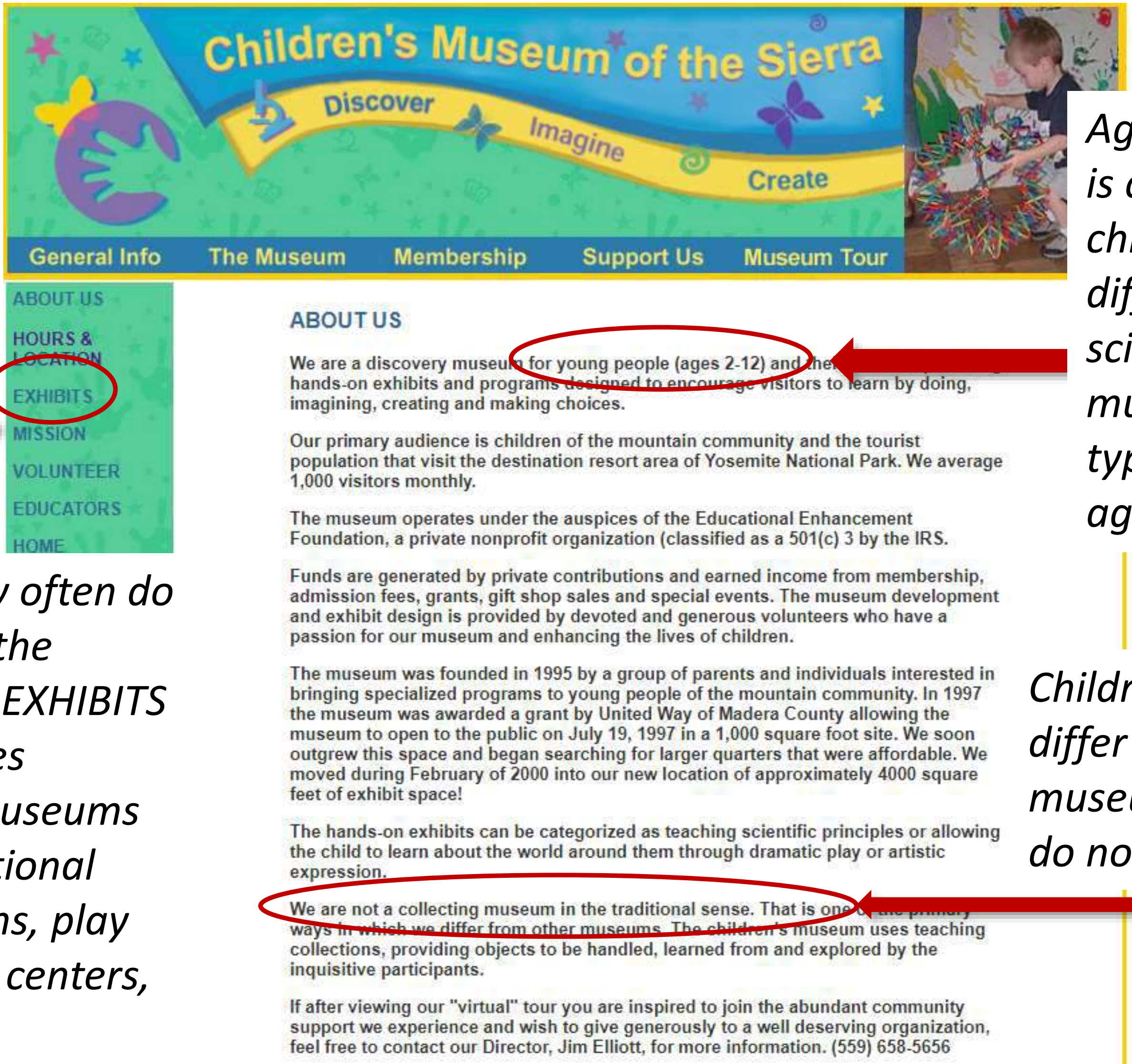
(4) Cares for these objects and

(5) Exhibits these objects to the general public on a regular basis through facilities that it owns or operates

(6) Uses a professional staff (Paid or unpaid)

- *Various ways museums vary greatly in the specific details associated with the definition*
- *Museum Disciplines: characterize the content and audience (e.g., children's museums; science museums; history museums; zoos; arboretums)*
- *Within-discipline homogeneity: provides useful analytical boundaries*

Children’s Museum – Example Description



ABOUT US

We are a discovery museum for young people (ages 2-12) and the [illegible] hands-on exhibits and programs designed to encourage visitors to learn by doing, imagining, creating and making choices.

Our primary audience is children of the mountain community and the tourist population that visit the destination resort area of Yosemite National Park. We average 1,000 visitors monthly.

The museum operates under the auspices of the Educational Enhancement Foundation, a private nonprofit organization (classified as a 501(c) 3 by the IRS.

Funds are generated by private contributions and earned income from membership, admission fees, grants, gift shop sales and special events. The museum development and exhibit design is provided by devoted and generous volunteers who have a passion for our museum and enhancing the lives of children.

The museum was founded in 1995 by a group of parents and individuals interested in bringing specialized programs to young people of the mountain community. In 1997 the museum was awarded a grant by United Way of Madera County allowing the museum to open to the public on July 19, 1997 in a 1,000 square foot site. We soon outgrew this space and began searching for larger quarters that were affordable. We moved during February of 2000 into our new location of approximately 4000 square feet of exhibit space!

The hands-on exhibits can be categorized as teaching scientific principles or allowing the child to learn about the world around them through dramatic play or artistic expression.

We are not a collecting museum in the traditional sense. That is one of the primary ways in which we differ from other museums. The children's museum uses teaching collections, providing objects to be handled, learned from and explored by the inquisitive participants.

If after viewing our "virtual" tour you are inspired to join the abundant community support we experience and wish to give generously to a well deserving organization, feel free to contact our Director, Jim Elliott, for more information. (559) 658-5656

Age-group specification is common with children’s museums – differentiates them from science-oriented museums, which typically have a broader age-range.

Children’s museums often differ from other museums – many of them do not “collect”

Though they often do not collect, the presence of EXHIBITS differentiates children’s museums from educational organizations, play spaces, arts centers, and retail establishments.

Challenges: The same museum might have different names at different times …

We should support the Children’s Museum of Judea.

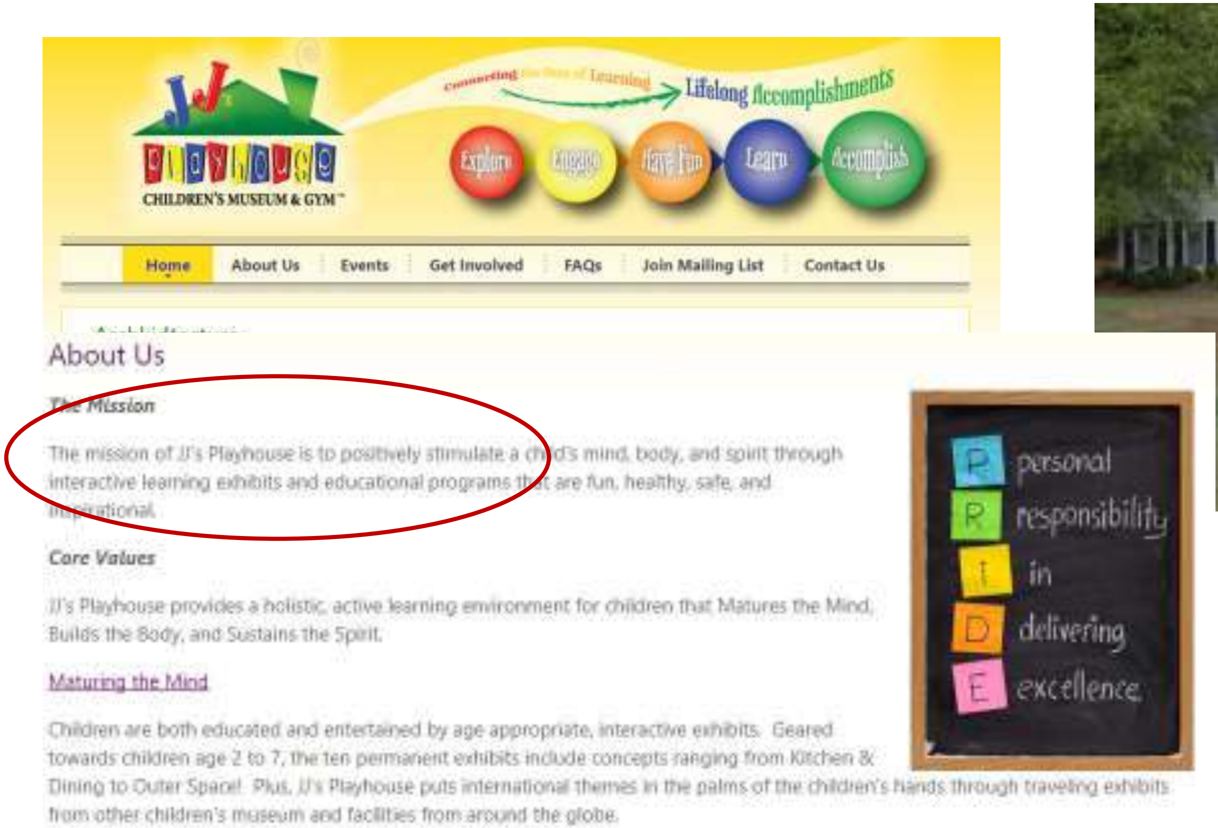
No, no, no, we should support the Judean Children’s Museum.

I am supporting the Carnegie Children’s Museum of Judea.

C’mon, let’s go to the CMOJ!

Challenges: Similar Names – One's a Museum, the other, not quite

JJ'S Playhouse



About Us

The Mission

The mission of JJ's Playhouse is to positively stimulate a child's mind, body, and spirit through interactive learning exhibits and educational programs that are fun, healthy, safe, and inspirational.

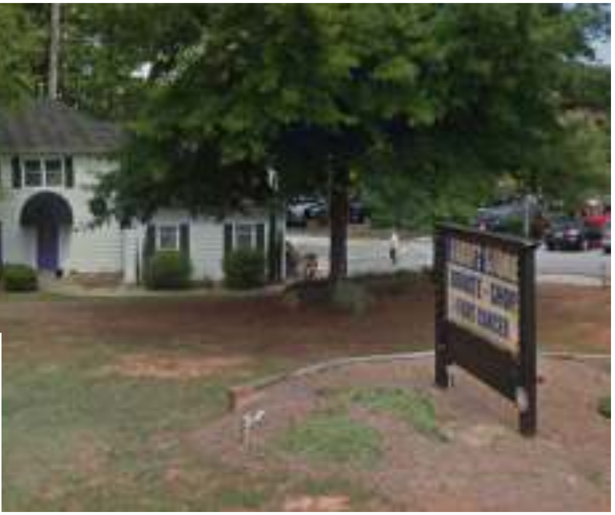
Core Values

JJ's Playhouse provides a holistic, active learning environment for children that Matures the Mind, Builds the Body, and Sustains the Spirit.

Maturing the Mind

Children are both educated and entertained by age appropriate, interactive exhibits. Geared towards children age 2 to 7, the ten permanent exhibits include concepts ranging from Kitchen & Dining to Outer Space! Plus, JJ's Playhouse puts international themes in the palms of the children's hands through traveling exhibits from other children's museum and facilties from around the globe.

Linda's Playhouse



Assessment of Existing Children’s Museum Subset File

- **Dataset 1:**
  - Initial data compiled in 2014 for the museum universe data file
  - 873 file entries with museum discipline = children’s museum (CMU)
  - With addition of new variables (below), duplicate entries removed, final n = 591
- Key variables - Existing
  - Names and addresses of entities
  - Geocode data
  - NAICS and NTEEC codes
  - Source flags:
    - IRS 990 BMF
    - Factual
    - Association lists*
    - Agency records
    - Private Foundation
- New variables
  - Type of entry (Museum or not)
  - Level of duplication
    - Dummy variable – duplicate vs. unique
    - Number of file entries for the establishment

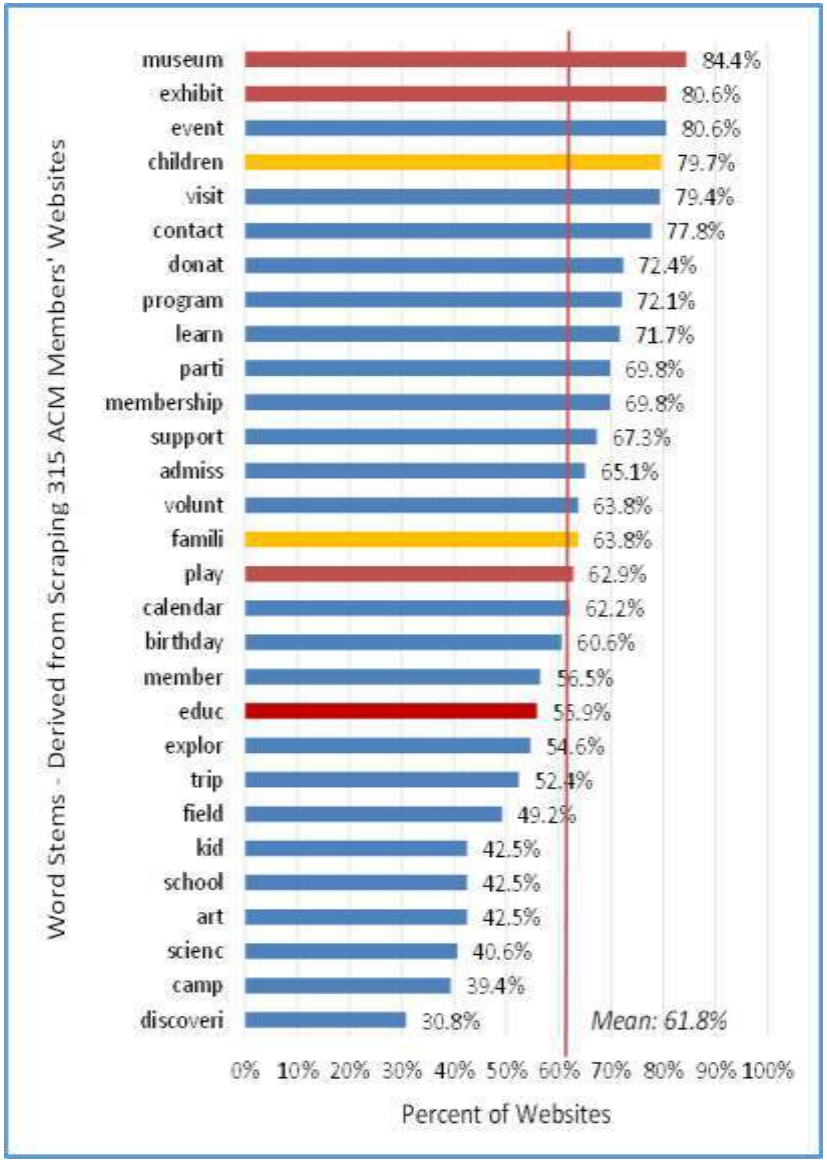
Research questions we can answer:

- What was the uniqueness and validity of entries supplied by the original sources of data?
- *How reliable are the NTEEC and NAICS codes in identifying museums versus other types of organizations? → See paper!*

**Association of Children’s Museums (ACM) most important for this paper*

Dataset 2: Two-Stage Web Scraping - Stage 1

1. Access ACM's online listing of members – URLs available for 315 U.S. museums (valid children's museums)
2. BeautifulSoup module used to scrape front pages
3. Single UTF-8 encoded text strings, punctuation stripped, text strings tokenized using the Natural Language Toolkit for Python with the Porter stemming dataset (e.g., "child" can be used in place of "children", "child", "childhood", and other variants)
4. 500 stems so identified, sorted in descending order of frequency
5. Removed highly common terms that would not differentiate children's museums
6. Retained 28 most common terms



Dataset 2 - Web Scraping Stage 2

1. Used APIs provided by Yelp.com and Yellowpages.com,
   - 1st stage, 28 common terms,
   - U.S. Census Bureau's 2016 Incorporated Places Dataset for places of >10,000
2. Both services assigned a unique identifier to each business → facilitated automated deduplication due to overlap of geographic areas
3. Python script to web scrape the presumed unique URLs for each entry identified in the "children's museum" category
4. Worksheet with all front page information assembled:
   - Tokenized the strings (NLTK)
   - Stem presence identified (yes/no) → *Another paper*
5. Manual review to code additional variables:
   - **Accuracy** – two variables (museum or not) AND (children's museum vs. other type of museum)
   - **Duplication** (old = already in children's museum file (i.e., Dataset 1, or new)
   - Noted **reasons** for inaccuracy

Research questions we can answer:

- What was the uniqueness and validity of establishments pulled from yellowpages.com and yelp.com "children's museums" categories?
- What are the strengths and weaknesses of each source for children's museum universe file updating?

Metrics / Analysis

Summarized for:

- Original children’s museums subset (Dataset 1)
  - IRS 990
  - Factual
- New web scrape results (Dataset 2*)
  - Yelp.com
  - Yellowpages.com

		Accuracy		
		No	Yes	
Unique	No	Not accurate / not unique	Accurate / not unique	Total Not Unique
	Yes	Not accurate & unique	Accurate & unique	Total Unique
		Total Not Accurate	Total Accurate	Total Entries with Source

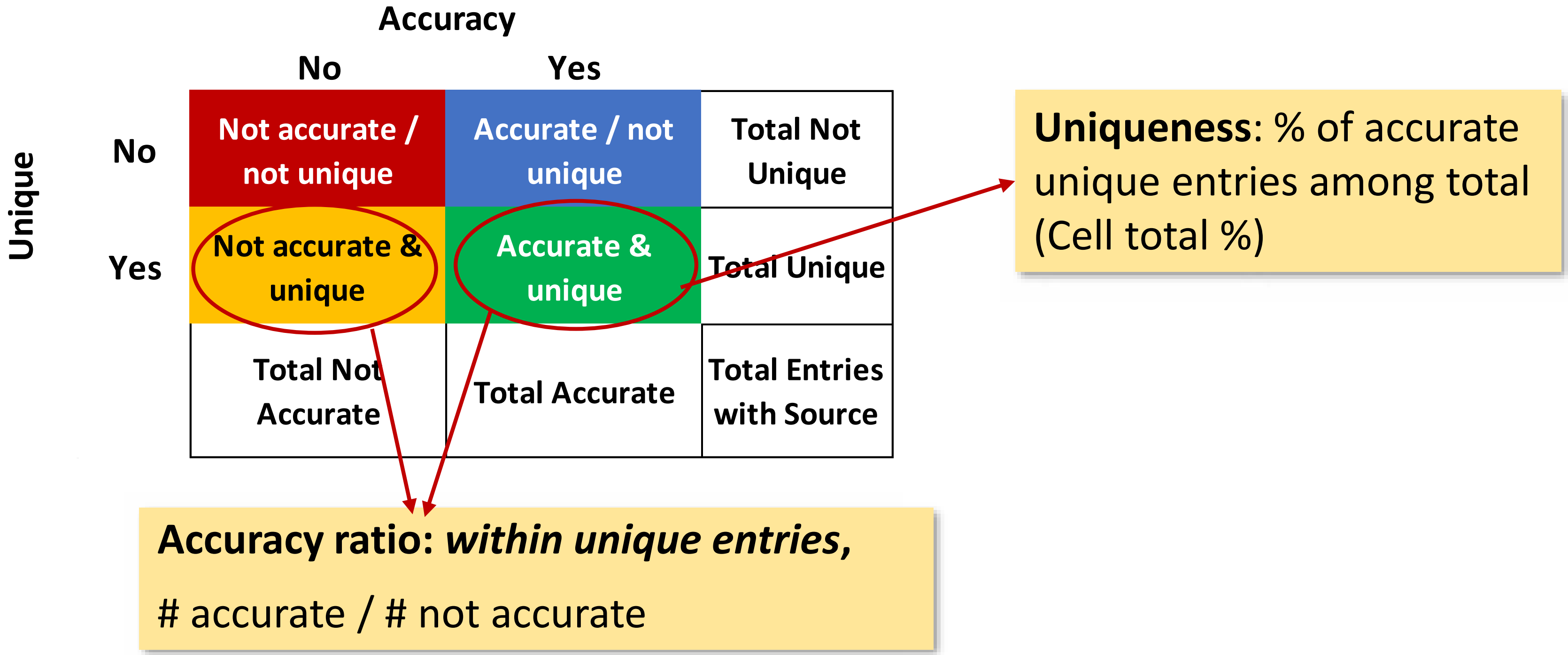
False positives: % of unique entries that are not accurate (Cell row %)

Overall accuracy: % of all entries from the source that conform to the museum definition (Column marginal)

**Unique: refers to a comparison of the web scrape results to Dataset 1 (No = Old; Yes = New)*

Additional Analysis – Efficiency Indicators for Web Scrape Results (Data Set 2)

Overlap detection efficiency: 1 – (# Missed Overlaps / # Entries)



		Accuracy		
		No	Yes	
Unique	No	Not accurate / not unique	Accurate / not unique	Total Not Unique
	Yes	Not accurate & unique	Accurate & unique	Total Unique
		Total Not Accurate	Total Accurate	Total Entries with Source

Uniqueness: % of accurate unique entries among total (Cell total %)

Accuracy ratio: *within unique entries,*

# accurate / # not accurate

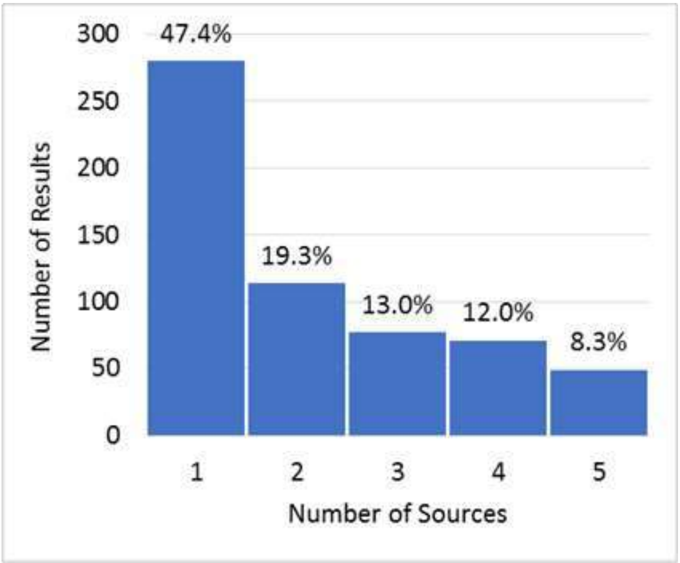
(1) Overlaps are not taken as "duplication" – when duplicate cases were identified during review, these were "Missed Overlaps"

(2) Unique: refers to a comparison of the web scrape results to Dataset 1 (No = Old; Yes = New)

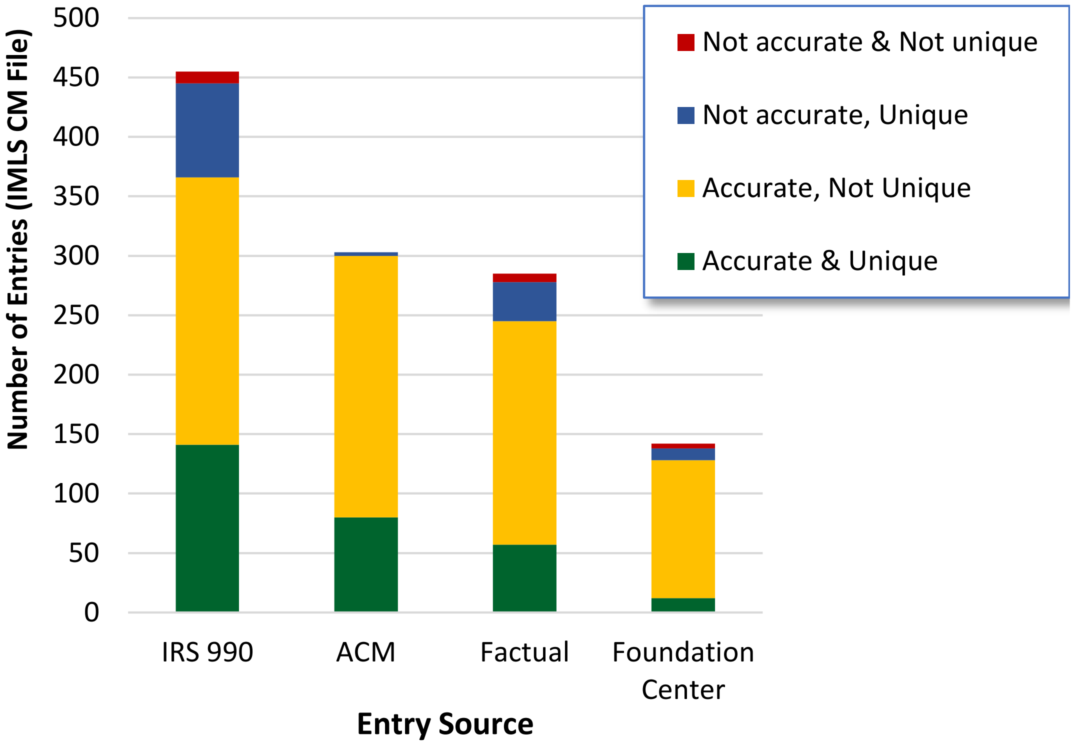
Original Children’s Museum Entries in Museum File

(n = 873 → 591 after deduplication)

Number of Sources for List Results

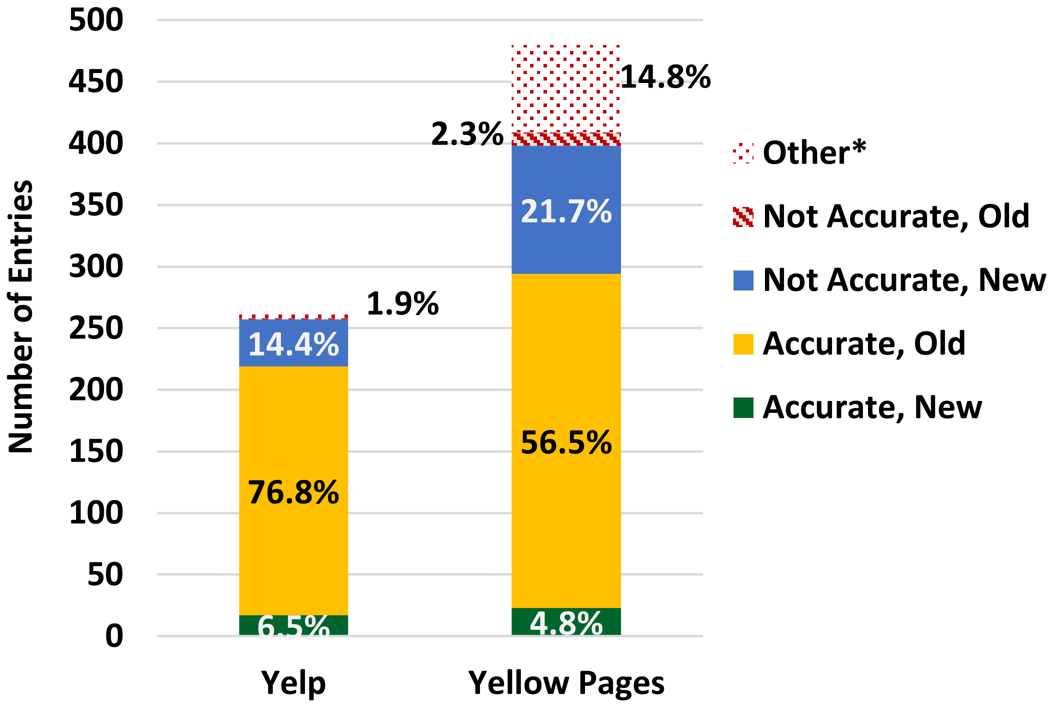


Accuracy and Uniqueness by Source of Entry



Web Scraping Results - Tale of the Tape

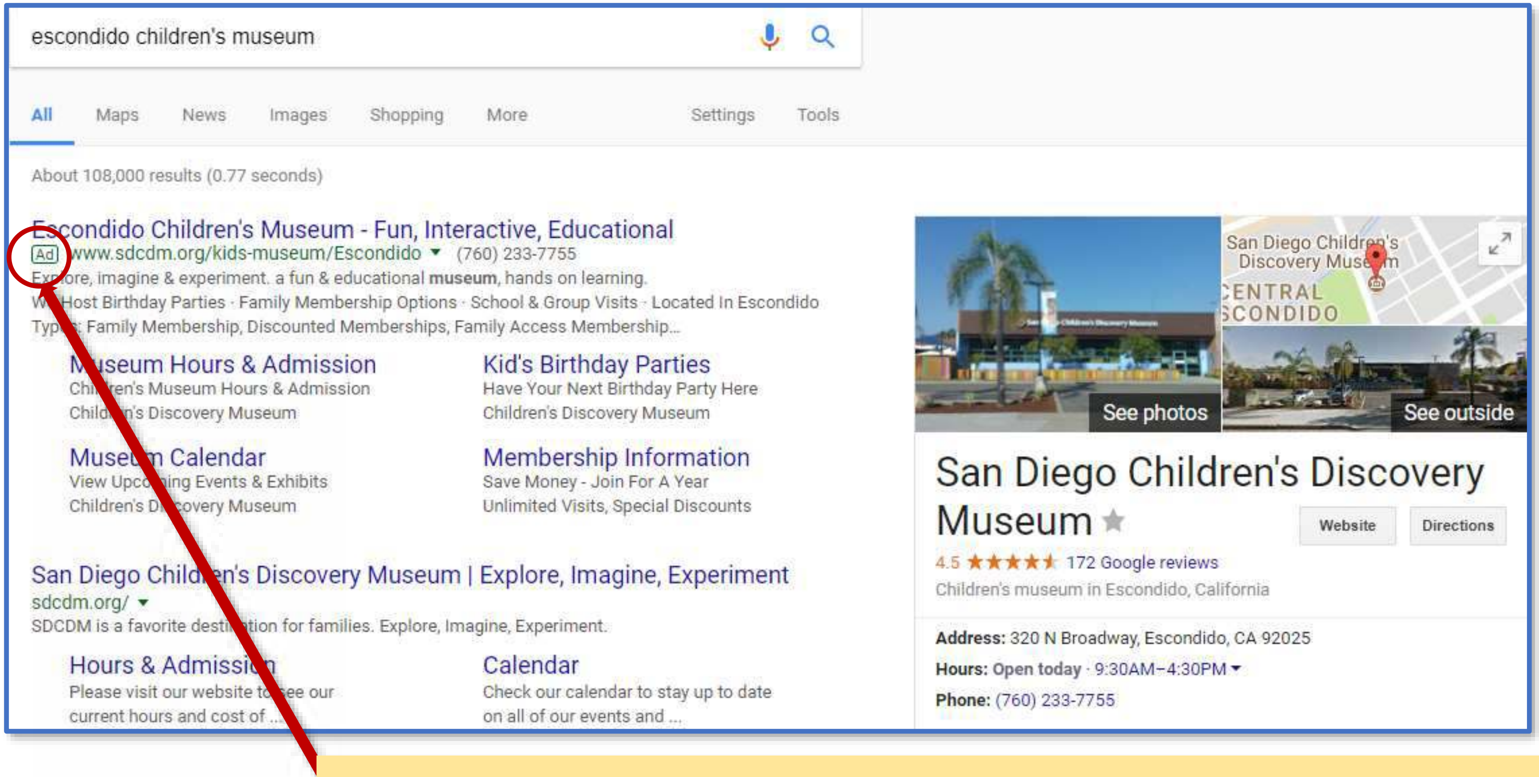
	Yelp	Yellow Pages
Total entries	7,200	19,246
De-overlapped entries	263	480



**Other: Yelp – 2 Non-U.S. entries and 3 missed overlaps; Yellowpages – 71 missed overlaps.*

Challenges: Source of duplication - Museums use different names to make sure that people can find them (Cross-referencing)

With a location in ESCONDIDO, close to San Diego, the San Diego Children's Discovery Museum will appear when a user searches for the Escondido Children's Museum.



escondido children's museum

About 108,000 results (0.77 seconds)

Escondido Children's Museum - Fun, Interactive, Educational
Ad www.sdcdm.org/kids-museum/Escondido ▾ (760) 233-7755
Explore, imagine & experiment. a fun & educational **museum**, hands on learning.
We Host Birthday Parties · Family Membership Options · School & Group Visits · Located In Escondido
Types: Family Membership, Discounted Memberships, Family Access Membership...

Museum Hours & Admission
Children's Museum Hours & Admission
Children's Discovery Museum

Kid's Birthday Parties
Have Your Next Birthday Party Here
Children's Discovery Museum

Museum Calendar
View Upcoming Events & Exhibits
Children's Discovery Museum

Membership Information
Save Money - Join For A Year
Unlimited Visits, Special Discounts

San Diego Children's Discovery Museum | Explore, Imagine, Experiment
sdcdm.org/ ▾
SDCDM is a favorite destination for families. Explore, Imagine, Experiment.

Hours & Admission
Please visit our website to see our current hours and cost of ...

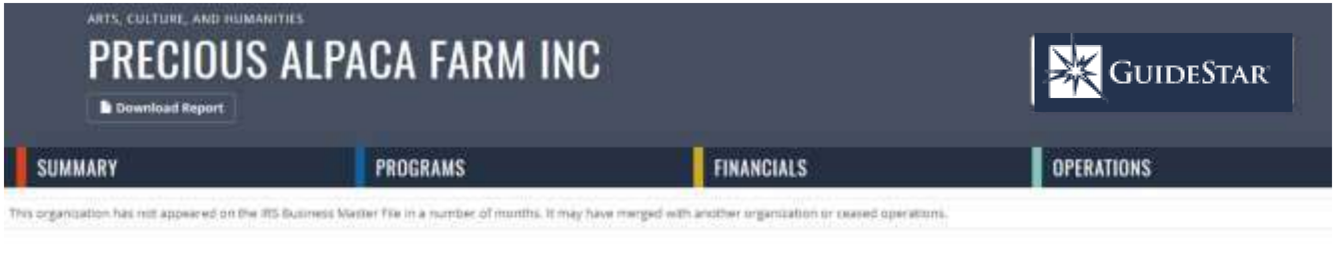
Calendar
Check our calendar to stay up to date on all of our events and ...

San Diego Children's Discovery Museum ☆
4.5 ★★★★★ 172 Google reviews
Children's museum in Escondido, California

Address: 320 N Broadway, Escondido, CA 92025
Hours: Open today · 9:30AM–4:30PM ▾
Phone: (760) 233-7755

Google Ad words – makes it easy for an institution to increase its hits / easy for people to find

Challenges: Durability of web content & BOTS



ARTS, CULTURE, AND HUMANITIES

PRECIOUS ALPACA FARM INC

Download Report

GUIDESTAR

SUMMARY | PROGRAMS | FINANCIALS | OPERATIONS

This organization has not appeared on the IRS Business Master File in a number of months. It may have merged with another organization or ceased operations.

with a mission

Ruling Year
2006

EIN
75-3184790

IRS Filing Requirement
This organization is required to file an IRS Form 990-N.

Principal Officer
Brenda Brown

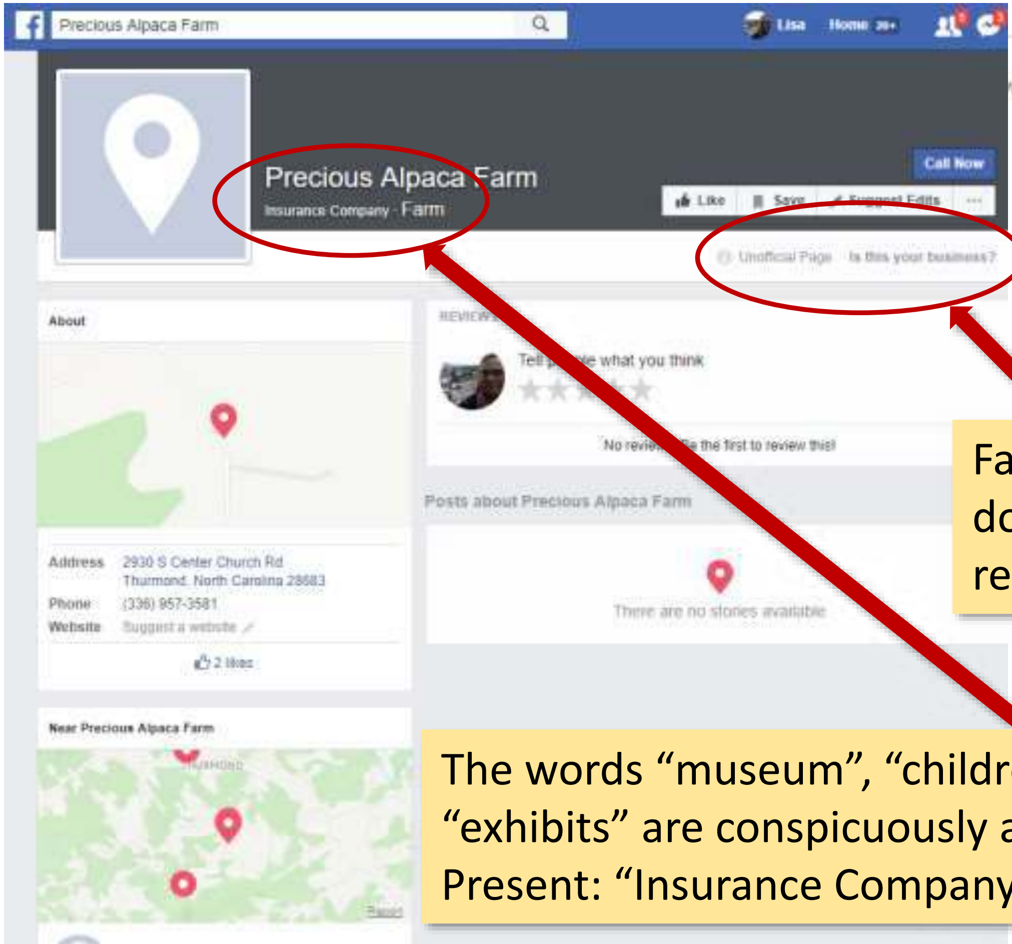
[illegible] Number
1515674101

Download Tax Forms
Show Forms 990

Main Address
PO BOX 37
THURMOND , NC 28683

Cause Area (NTEE Code)
Natural History, Natural Science Museums (A56)

Show More Contacts



Precious Alpaca Farm

Lisa Home 20+

Precious Alpaca Farm
Insurance Company · Farm

Call Now

Like | Save | Suggest Edits | ···

Unofficial Page · Is this your business?

About

Address 2930 S Center Church Rd
Thurmond, North Carolina 28683
Phone (336) 957-3581
Website Suggest a website

2 likes

REVIEWS

Tell p[illegible]e what you think

No revie[illegible] Be the first to review this!

Posts about Precious Alpaca Farm

There are no stories available

Near Precious Alpaca Farm

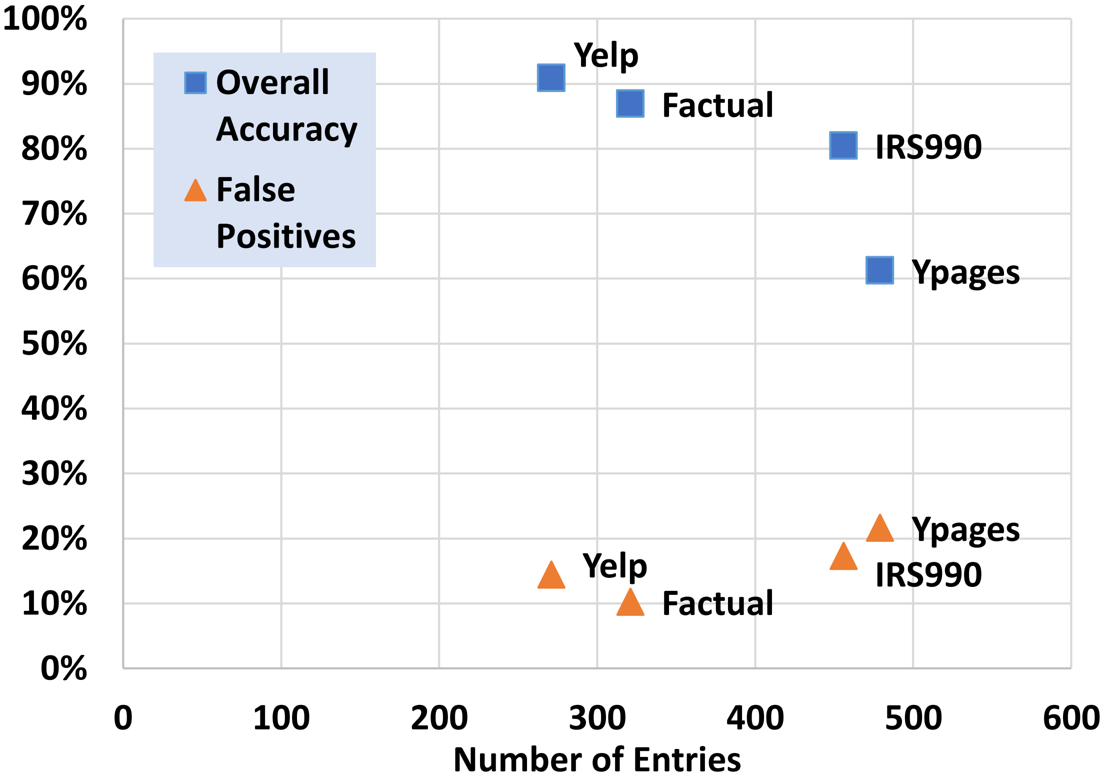
FaceBook doesn't make it real.

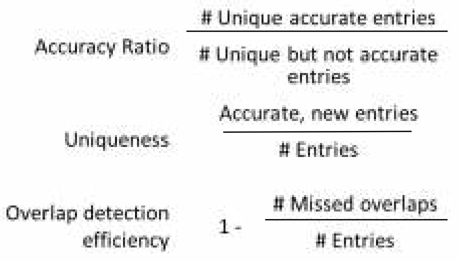
GuideStar's entry suggests it may have ceased operations AND the Alpaca Farm has indicated it is a *Natural History or Natural Science Museum with NTEE Code A56*.

The words "museum", "children", and "exhibits" are conspicuously absent. Present: "Insurance Company" and "Farm"

Efficiency Indicators

	Yelp	Yellow Pages
Overlap Detection Efficiency	98.86%	85.21%
Accuracy Ratio	2.24	4.52
Uniqueness	4.56%	3.75%





Accuracy Ratio $\frac{\#\text{ Unique accurate entries}}{\#\text{ Unique but not accurate entries}}$

Uniqueness $\frac{\text{Accurate, new entries}}{\#\text{ Entries}}$

Overlap detection efficiency $1 - \frac{\#\text{ Missed overlaps}}{\#\text{ Entries}}$

Note: as a point of comparison, a pull of IRS 990 data on 2/27/2018 yielded an overlap detection efficiency of 98.61% for 359 NTEEC A52, A52I, A52O, and A52Z entries

Conclusions

- Yelp & Factual:
  - Advantage: dynamic, implicitly crowd-sourced data – highly accurate results
  - Yelp → "Permanently closed" field – useful to deal with durability of web content issue
  - Disadvantage: Far fewer frame entries identified
- IRS 990:
  - Advantage: large number of identified entries with slightly lower accuracy than Yelp and Factual
  - Disadvantages: misses government / municipal-operated museums & limited coverage college/university museums
- Two-stage web scraping and other frame entry validation shortcuts
  - Stage 1: validated lists
  - Stage 2: broader web scrape
- Developed efficiency metrics – tradeoffs / lead to additional questions:
  - Oversample in establishment surveys to account for expected level of false positives in frame *vs.* expending additional up-front effort to use existing sources to validate frame entries?
  - Should we continue to use sources that fail to meet a standard efficiency level? What is that level?

Moving Forward / Next Steps

- Developing algorithms to assign a unique identifier to museums that will work with multiple sources of frame entries – building on another recent project that used FuzzyWuzzy for name matching
- Working to identify an effective set of terms (and NOT terms) to build an algorithm to validate frame entries (working with additional data scraped for this project)
- Children's museums were a relatively homogeneous testbed - adjust approach for more heterogeneous museum establishments
- Can we web scrape relevant information from museum webpages to build dataset with elements that permit validation?

THANK YOU

Lisa M. Frehill, Senior Statistician
lfrehill@imls.gov

Jason Enos, Data Analyst
jenos@imls.gov

Matt Birnbaum, Supervisory Social Scientist
mbirnbaum@imls.gov

INSTITUTE of Museum and Library SERVICES