

# Approaches to Digital Cleaning:

*Blending traditional methods with crowdsourcing and AI to clean a public sector establishment survey population frame*

Jake Soffronoff

*Author — Survey Methodologist  
Institute of Museum and Library Services (IMLS)*

Ai Rene Ong

*Collaborator — Researcher  
American Institutes for Research (AIR)*

Matthew M. Sweeney

*Collaborator — Senior Researcher  
American Institutes for Research (AIR)*

FCSM 2024



*Disclaimer: The author of this presentation is solely responsible for its content. The opinions expressed are the author's own and do not represent the views of IMLS, AIR, or any other organization.*

# The Institute of Museum and Library Services (IMLS)

- One of America's cultural agencies, along with the National Endowment for the Arts (NEA) & the National Endowment for the Humanities (NEH)
- The mission of IMLS is to advance, support, and empower America's museums, libraries, and related organizations through grantmaking, research, and policy development.



# Background:

## The National Museum Survey (NMS)

- More than a decade of effort
  - *The museum field needs nationally valid data*
- Frame discussed today developed for pilot NMS, which was successfully completed in Summer 2023
  - <https://www.ims.gov/webinars/ims-national-museum-survey-pilot-summary-findings-webinar>
- Full report on IMLS' development of the NMS population frame coming later this year

# Lessons Learned from Past Efforts

No “off-the-shelf” population frames available

- *IMLS defines “museum” broadly (includes zoos, botanical gardens, etc.)*
- *Association lists are opt-in, institutional parent-child relationships are challenging for public resources (IRS-990s), as is contact info, etc.*

Previous efforts: Data aggregation + web scraping approach

- *Museum Universe Data File (MUDF – 2013) | Museum Data File (MDF – 2017, 2018)*

Issues:

- **GIGO**: Inconsistent source quality → inconsistent frame quality
  - *Inconsistent “museum” definition, inclusion of for-profit museums, parent-child institutional relationships unclear, etc.*
- **Duplication**: Multiple entries for single institutions → challenge to clean fully
- **Onlinedness**: Tying museums to web presence → missed units

# New Approach

- “Phone Book” approach
- Based on physical location:
  - Capture museums without an online presence
  - Include museums “administratively hidden” within parent orgs
    - *E.g., academic museums “hidden” within universities*
  - Ensure accurate address data
- Independently, actively curated
- Final records include contact information & discipline - something like\*:

ID	Museum name	Address	Email address	Phone number	Discipline
1	Jake’s Good Museum	100, S. Example St. Example, NJ 12345	jake@jakesgoodmuseum.org	(732)-123-4567	Specialized Museum

\*In the final frame the address is split into street address, city, state, zip code, etc., and multiple contacts may be included where available to give the field team the best chance of reaching respondents. URL is also included wherever available to allow for manual review, web scraping, etc.

# Starting Point: Yelp + Official Museum Directory

- Yelp: Our “Phone Book”
  - Usable T&C’s
  - Up-to-date, curated information
  - Rigorous content control:
    - Constant crowdsourced updating
      - + Yelp-employed moderators who verify user-submitted data
      - + Establishment representatives “own” pages & verify information
- Official Museum Directory (OMD)
  - Email addresses for limited subset of institutions

# Needed Data Cleaning

## Excess rows

- 107,610 from Yelp → 21,465 pilot population frame → 7,050 pilot sample

The data IMLS received needed to be augmented and updated:

ID	Museum name	Address	Email address	Phone number	Museum discipline
1	Jake's Jazz Museum	100, S. Example St., NY	N/A	(212)-123-4567	N/A
2	Jazz Museum of Jake	105, S. Example St., NY	N/A	N/A	N/A
3	Jake's Aquarium Restaurant	1, N. Example St., NY	N/A	(212)-000-0101	N/A

**ID 2 is a duplicate of ID 1 — contact information fields are different**

**This is not a museum**

**Missing email address**

**Some records are missing phone no.**

**Missing museum discipline**

\*Note: Not real data; this illustrative example was created to describe challenges faced by the population frame team.

# Needed Cleaning

- Identify and remove non-museums (excess rows)
- Determine NMS-based museum disciplines for all rows (e.g., “art” or “history” museums)
- Acquire and append missing or invalid information on the frame
- Remove duplicate entries





# Strategies to clean and update the data

IMLS' population frame team implemented multiple strategies to clean and augment the Yelp data:



## Identify museums and non-museums

- Matching other data sources
- Manual coding
- Keyword search



## Identify museum disciplines

- Manual coding



## Update contact information

- Matching other data sources
- Manual coding
- ChatGPT, Web scraping



## Append additional museums

- Museum associations, etc.



# Let's skip these and get to the interesting stuff...



- Data modeling – no-can-do!
  - Insufficient ancillary variables, class imbalance
- Keyword searching
  - Maintain rows including “museum,” “arboretum,” etc.
- Data matching
  - Fuzzy matching for deduping
  - Align triangulated/augmenting resources
- Manual coding
  - Too slow with available resources:
    - *Estimated 454 days to complete*
- Web scraping
  - Python scrapes museums’ URLs for email addresses

# Amazon Mechanical Turk (MTurk)

Platform where discrete tasks are crowdsourced to a distributed human workforce who are paid upon the completion of each discrete task (aka “Human Intelligence Tasks” or “HITs”).

Why MTurk?	
1	Data modeling had failed and manual review was far too slow
2	MTurk has been employed to label data training data in many other studies. <sup>1,2</sup> Many MTurk workers are already accustomed to labelling data
3	Expected to speed up manual review processes
4	MTurk is cost effective*

1. Guo, L., Mays, K., Lai, S., Jalal, M., Ishwar, P., & Betke, M. (2020). Accurate, Fast, But Not Always Cheap: Evaluating “Crowdcoding” as an Alternative Approach to Analyze Social Media Data. *Journalism & Mass Communication Quarterly*, 97(3), 811–834. doi: 10.1177/1077699019891437

2. Kasthuriarachchy, B., Chetty, M., Shatte, A. & Walls, D. (2021). Cost Effective Annotation Framework Using Zero-Shot Text Classification. *International Joint Conference on Neural Networks (IJCNN)*, Shenzhen, China, pp. 1-8, doi: 10.1109/IJCNN52387.2021.9534335.

\*Though lower payment options were available, to ensure that the work was ethically compensated IMLS worked with its subcontractor to ensure that MTurk workers would be paid the equivalent of at least \$15/hour for the tasks they completed on the agency's behalf.

# MTurk Testing: Is [ROW] a Museum?

<b>MTurkers:</b> Is this a museum? <i>(SUCCEFULLY SCREENED MTurkers vs. AIR Live Reviewers)</i>		
N=529	COMBINED: Not a museum/ Non-Museum Art Gallery	Yes this is a museum
COMBINED: Not a museum/ Non-Museum Art Gallery	224 (42%)	37 (7%)
Yes this is a museum	67 (13%)	186 (35%)
Unsure	15 (3%)	

<b>Outside contractor:</b> Is this a museum? <i>(SUCCEFULLY SCREENED Live reviewers vs. AIR Live Reviewers)</i>		
N=536	COMBINED: Not a museum/ Non-Museum Art Gallery	Yes this is a museum
COMBINED: Not a museum/ Non-Museum Art Gallery	223 (42%)	23 (4%)
Yes this is a museum	52 (10%)	175 (33%)
Unsure	63 (12%)	

# First MTurk Iteration

Ensure MTurk workers carefully read the instructions

1<sup>ST</sup> Iteration

- Added 3 attention screener questions to ensure MTurk workers were reading instructions

- In-line question (“Answer X below” embedded in the instruction text)
- Two content questions around nuances of NMS museum definition

**Due to irregular responses throughout this project, the research team is increasingly more diligent regarding answers for these tasks. Therefore, you may experience delays in hit approval (but it will remain within 72 hours). Additionally, irregular or improper responses will result in non-payment, and repeated or blatant offenses may result in the research team restricting accounts from completing tasks in the future.**

What should you answer below to get credit?

Please Select an Option Below

Which of the below is NOT considered a museum according to the definition above?

Please Select an Option Below

Why is a non-museum art gallery NOT considered to be a museum for the purposes of this task?

Please Select an Option Below

Link: `#{url}` Identification Number: `#{ID}`

Is the place in the link above a museum as defined for this task?

Please Select an Option Below

# Our Experience: Quality Issues. Bots?

- Compared to live coders, too many rows were being classified as museums
- Too many rows were being evaluated by too few MTurk workers in too little time
  - E.g., 4s is implausibly quick to read through the instructions, research the entity and make an informed decision
  - The faster the MTurk worker, the more rows marked as “museums”
- Some MTurk workers clicked through without selecting an option, defaulting to identifying all entities as the first available option (“Museum”)

# Lessons from Combatting Non-Probability Survey Research Bots Inform MTurk Workarounds

The population frame team tested at the 2<sup>nd</sup> iteration to ensure that these changes do result in reduced errors.

MTurk workers were coding too many entities as museums

2<sup>nd</sup> Iteration

- “Museum” no longer default
- Added “bannable” response option (i.e., “If you select this, you will be blocked.”)
- Added 2 ban criteria based on no. of HITS taken and % museum

## Test:

- Rerun a sample of 500 records previously coded as “museums”
- **Just 40% were re-identified as “museums” by other MTurk workers**

## Response:

- “This is a museum” no longer default answer choice
- Added “if you select this, you will be blocked” and banned anyone who chose that option
- Added banning criteria to catch script-running bots: Ban those who completed >52 hits at a rate of >46% museums, and those who completed >104 hits at a rate of >23% museums\*

\*Thresholds for banning were determined through an examination of the data to find levels where MTurk workers’ museum screen-in rates became systematically suspiciously high

# Identifying museums and non-museums: MTurk Challenges

At the 3<sup>rd</sup> iteration, the population frame team compared the proportion identified as museums between those who answered the math screener correctly and those who did not.

MTurk workers were using scripts to complete the HITs

3<sup>rd</sup> Iteration

- Added a math screener question

## Test:

- Proportion “museum” for those who answered math screener correctly were lower than those who did not

As those who answered wrongly were banned from taking more HITs, the number of correctly math screener responses decreased over batches.

## Final Test:

- Rerunning 25,548 records previously identified as museums led to just 7,225 (28%) being again identified as museums.

***Presumption: The instruction modifications helped reduce task completion error.***

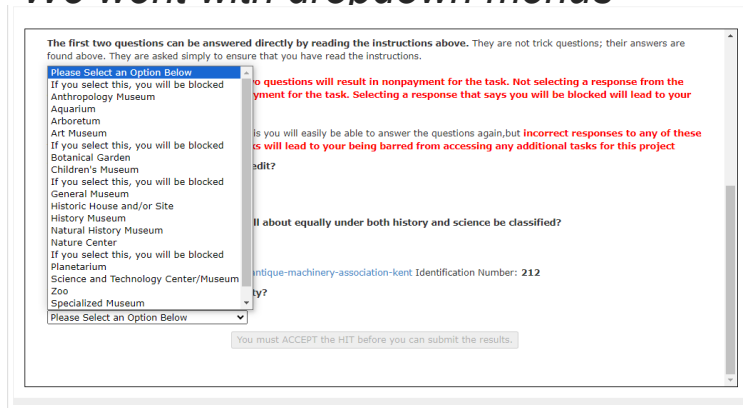


# Additional MTurk Work

Assigning museum disciplines to population frame units:

1. Layout test: Radio buttons and dropdown menus yielded very similar error rates

- *We went with dropdown menus*



The first two questions can be answered directly by reading the instructions above. They are not trick questions; their answers are found above. They are asked simply to ensure that you have read the instructions.

Please Select an Option Below

- If you select this, you will be blocked
- Anthropology Museum
- Aquarium
- Arboretum
- Art Museum
- If you select this, you will be blocked
- Botanical Garden
- Children's Museum
- If you select this, you will be blocked
- General Museum
- Historic House and/or Site
- History Museum
- Natural History Museum
- Nature Center
- If you select this, you will be blocked
- Planetarium
- Science and Technology Center/Museum
- Zoo
- Specialized Museum
- Please Select an Option Below

to questions will result in nonpayment for the task. Not selecting a response from the ment for the task. Selecting a response that says you will be blocked will lead to your

is you will easily be able to answer the questions again, but incorrect responses to any of these is will lead to your being barred from accessing any additional tasks for this project

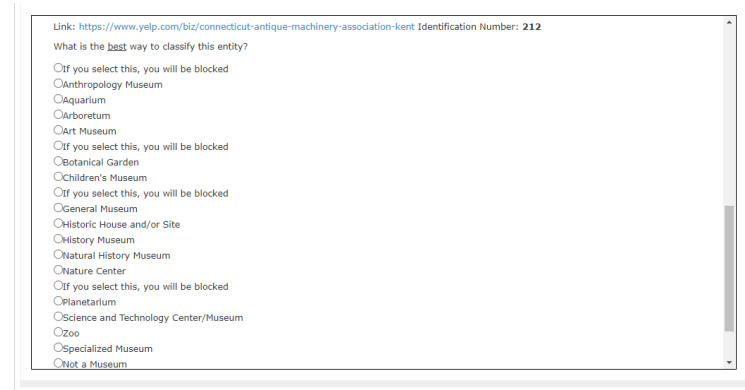
edit?

If about equally under both history and science be classified?

ntique-machinery-association-kent Identification Number: 212

ty?

You must ACCEPT the HIT before you can submit the results.



Link: <https://www.yelp.com/biz/connecticut-antique-machinery-association-kent> Identification Number: 212

What is the best way to classify this entity?

- If you select this, you will be blocked
- Anthropology Museum
- Aquarium
- Arboretum
- Art Museum
- If you select this, you will be blocked
- Botanical Garden
- Children's Museum
- If you select this, you will be blocked
- General Museum
- Historic House and/or Site
- History Museum
- Natural History Museum
- Nature Center
- If you select this, you will be blocked
- Planetarium
- Science and Technology Center/Museum
- Zoo
- Specialized Museum
- Not a Museum

2. Grouped vs. ungrouped disciplines: Grouped disciplines yielded more reliable results

- *Ungrouped list: 15 disciplines | Bucketed list: 9 discipline groups*

# Testing ChatGPT

**ChatGPT (Chat Generative Pre-trained Transformer)**: A large language model chatbot developed by OpenAI that was trained on a corpus of publicly available text data that can generate human-like text to respond to queries (“prompts”). Previous research used ChatGPT for annotating data<sup>1,2</sup>; we tried:

- **Categorizing entries as museums or non-museums: Unsuccessful**
  - ChatGPT’s preconceived notions about what “museums” are were difficult to overcome
- **Identifying museum disciplines : Unsuccessful**
  - Nuanced units incorrectly classified
  - Similar entity names confused the platform
- **Obtaining missing data for population frame units: Mixed Results**
  - **Effective** finding URLs, **ineffective** finding email addresses.

1. Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*.

2. Kuzman, T., Ljubešić, N., & Mozetič, I. (2023). ChatGPT: beginning of an end of manual annotation? Use case of automatic genre identification. *arXiv preprint arXiv:2303.03953*.

# ChatGPT Takeaways

- The “black box” nature of ChatGPT makes it hard to rely on
  - *Take nothing for granted — it’s a whole new world of ways to fail*
- Specify terms as much as possible
- Thoroughly check results against known benchmarks
- Iteratively “cognitively test” how ChatGPT interprets terms by asking “why did you answer that way?” to find and patch disconnects

**Overall:** Worth seeing what sticks, but contrary to the hype  
***do not expect untold worlds of success***

# Putting it Together: Tools in the Toolbox

Treating methods as tools, rather than answers, each with its own function. Mixing methods can yield good results.

For example, our approach for obtaining contact information implemented three methods, each providing value:

1. **MTurk:** Used to classify rows as museums or non-museums
2. **ChatGPT:** Used to obtain museum URLs for all identified museums
3. **Web scraping:** Scrape museum URLs for email addresses.
4. **More MTurk:** Even though it was not as reliable as web scraping for email addresses, it could provide some lift at minimal cost for museums whose contact information could not be obtained using (2) and (3) above.

# Thank you!

Contact:

Jake Soffronoff

Survey Methodologist

Office of Research & Evaluation

Institute of Museum and Library Services (IMLS)

[jsoffronoff@imls.gov](mailto:jsoffronoff@imls.gov)

## Presentation Appendix

~AKA~

Details there weren't enough time for.

*For more, please keep an eye out for IMLS' NMS population frame report to be published in 2024*

# Yelp vs. NMS population frame

## Yelp Data

**107,610** rows of museums and non-museums

No email addresses

No museum disciplines



## Final Frame

**21,464** rows of museums

**16,489** rows have email addresses (77% of the frame)

Museum discipline identified for all rows

# Current NMS Population Frame

On launch, there were very few email bounce backs from the pilot's sample. Though 650 emails bounced back initially, this was reduced to about 55 bounce backs when the second available email for these museums was used for subsequent outreach.

21,464 rows of museums

16,489 rows have email addresses

Museum discipline identified for all rows

**Sampled museums:** 7,050 for the NMS pilot (6,600 with email addresses, 450 without email pushed to web via phone and snail mail)

**# initial email bounce backs:** 650

**# of email bounce backs after using secondary email address:** ~55 bounce backs (Note: the team is also currently testing ChatGPT to acquire additional emails in-field)

**7 field methods experiments:** Phone and mailed prenotification and reminder experiments



# Known Frame Weaknesses



Despite efforts to deduplicate and clean the frame, there may still be duplicates where the same institution may have multiple names and/or changed names



The frame may still be missing smaller museums: these museums are more likely to have been omitted from Yelp as they receive less of the foot traffic that would lead to their inclusion



The frame likely includes some records that are not museums by IMLS definition (e.g., for-profit museums that would be screened out while taking the NMS), and/or non-museums that were not caught during manual data cleaning



Contact information is still missing for many records and needs to be updated wherever possible

# Path Forward

1

## Direct outreach

- Prior to the pilot, IMLS engaged in extensive communications outreach to raise awareness of the NMS
- IMLS plans to continue engaging in outreach to museums in between surveys to update its contact records

2

## Less-direct outreach

- Validate against MDF
- Potentially set up interagency agreements with outside public sector experts to verify the frame

3

## Partner with outside agencies

# NMS Museum Definition and Eligibility Criteria

IMLS' museum definition is broad and includes disciplines that are frequently not categorized as museums by other organizations (e.g., zoos, botanical gardens)

## NMS Eligibility Criteria (2023):

A unit of Federal, State, local, or tribal government, or a not-for-profit institution that:

- Serves the public in a physical location it owns or operates
- Provides exhibitions & programs
- Has as its primary function to house, display, and care for animate or inanimate objects that form the core of its exhibitions, programs, and research
- Under normal circumstances, is open to the public 120 days or more per year, either through specific hours of operation or by appointment
- Has at least one staff member, or the full-time equivalent, whether paid or unpaid

## Eligible Disciplines (2023):

- Anthropology museums
- Aquariums
- Arboretums
- Art museums
- Botanical gardens
- Children's museums
- General museums
- Historic houses and/or sites
- History museums
- Natural history museums
- Nature centers
- Planetariums
- Science and technology centers/museums
- Specialized museums
- Zoos

# Updating Contact Info: MTurk

The population frame team tested the reliability of using MTurk to obtain email addresses and phone numbers for museums:

## Test 1: Updating Contact Info with MTurk

**Method:** The team randomly sampled 100 museums that had pre-existing email addresses and phone numbers from matching with OMD or as coded by internal manual coders. These were presented as a task on MTurk, where MTurk workers were asked to visit the provided websites and fill in that site's contact information (email address and phone number). The contact information received was then compared against the contact information on the frame.

	Email Address	Phone Number
% matched OMD	46%	82%

**Results:** Due to time constraints, this test attempting to collect email and phone numbers was halted when 71 HITs were completed; the preliminary results uncovered reliability issues and as such the tests were abandoned.

# ChatGPT Prompt (1) – Email Address

"content" = paste("Based on the available information, what are the best contact Email Address(es) for", df\$Name[i], "in the state of", df\$State[i], "? Please include any available email address for the location and/or domain name, separated by a comma, including those to specific people within the organization. Once you identify an email address, please find any other available email addresses that use the same domain name and include them as well. If you have the email address only respond with the email address. If you do not have any available email addresses, respond with N. Do not include any other information or text output beyond the email address(es) or N.")

# ChatGPT Prompt (2) – Business URL

```
"content" = paste("Based on the available information, what are the best web address hyperlink for", df$Name[i], "in the state of", df$State[i], "? If the location does not have a website, enter in any social media profile that you can find. Otherwise, respond with N. Do not include any other information beyond a hyperlink to a website, a hyperlink to a social media page, or N.")
```

# ChatGPT Prompt (3) – Phone Number

```
"content" = paste("Based on the available information, what are the best phone number", df$Name[i], "in the state of", df$State[i], "? If the location does not have a phone number, respond with N. Do not include any other information beyond a phone number or N. Please format the phone number as XXX-XXX-XXXX.")
```

# ChatGPT Prompt (4) – Mailing Address

```
"content" = paste("Based on the available information, what is the best mailing address for", df$Name[i], "in the state of", df$State[i], "? If the location does not have a mailing address, respond with the physical address. Otherwise, respond with N. Do not include any other information beyond an address or N.")
```



# ChatGPT Prompt (5) – Contact Person

"content" = paste("Based on the available information, who is the director, leader, manager, board chairperson, or president of", df\$Name[i], "in the state of", df\$State[i], "? If the location does identify an organizational leader, respond with N. Do not include any other information beyond the name or N.")

# ChatGPT Prompt (6) – Identifying Museums

"content" = paste("a museum is an entity: That has a physical location Whose primary function is housing, displaying, caring for living or inanimate objects/exhibits. Note: this primary function is NOT selling items. For example, although a museum may have a museum store, the store is not the point of a museum. A non-museum art gallery is also not a museum: it could be private or public, but either way it primarily exists to sell art, not to display it for the public, and so it is not a museum. Museums includes places like Zoos, Aquariums, Botanical Gardens, and Arboretums; Nature Centers; Science Centers; History Museums and Historic Sites; Art Museums; Children's Museums; Natural History Museums; and Specialized Museums. Note: for the purposes of this task, museums must have staff or volunteers that interact with the public to show their objects and exhibits. So, places like recreational parks should not be marked as museums. Based on this definition, is", df\$Name[i], "in the state of", df\$State[i], "a museum? Please only respond with, Yes, No, or Unknown. Only respond with Unknown if there is not enough information to confirm or deny the result. Do not respond with anything else other than Yes, No, or Unknown.")

# ChatGPT Prompt (7) – Classifying Museum Type

"content" = paste("Based on your knowledge, what type of museum is", df\$Name[i], "in the state of", df\$State[i], "? Please choose from the following list: aquariums, arboretums, art museums, botanical gardens, children's/youth museums, general museums (those having two or more significant disciplines), historic houses/sites, history museums, natural history/anthropology museums, nature centers, planetariums, science/technology centers, specialized museums (limited to a single distinct subject), and zoological parks. Only respond with one of these categories, otherwise respond with Unknown.")

# MTurk Museum Identification Task Instructions (Final)

## Data Collection Instructions/Codebook

**NOTE: The instructions are VERY important to getting paid for this task**

- For this task, a **museum**:
  1. **Has a physical location**
  2. Primary function is **housing, displaying, caring for living or inanimate objects/exhibits.**
  3. Open to the public through **specific hours of operation** or by appointment
  4. Has **at least one staff member** whether paid or unpaid

**Museums includes places like Zoos, Aquariums, Botanical Gardens, and Arboretums; Nature Centers; Science Centers; Answer Orange below to get credit; History Museums and Historic Sites; Art Museums; Children's Museums; Natural History Museums; and Specialized Museums.**

## An entity is NOT a museum if:

1. Its primary function is **to sell items**.
  - Although a museum may have a museum store, the store is not the main focus of the museum.
  - A non-museum art gallery is **not** a museum because it primarily exists to sell art, not to display and preserve it.
2. Its staff or volunteers **do not interact with the public**.
  - Recreational parks' staff do not interact with the public to show objects or exhibits, so they are **not** museums. Places like recreational parks should not be marked as museums.
3. It **does not have a curated exhibit or collection**.
  - Historic landmarks with **no curated exhibits or objects, tours, etc.** will not be considered a museum.
  - A historic landmark such as the First Ladies National Historic Site is considered a museum because the home of First Lady Ida Saxton-McKinley is curated for the public and is open for visits.
  - A community garden that does not have staff to take care of and display a curated collection of plants **is not** a museum, but a botanical garden whose purpose is showing a curated collection of plants **is** a museum.
  - National, state, or local parks are museums only if they have a curated exhibit that is open to the public. For example, Gettysburg National Military Park offers the public access to the Gettysburg National Military Park Museum & Visitor Center.

### **Instructions:**

1. Open and read the website carefully. some entities may require further investigation
2. Decide if the entity is a museum.

We appreciate your effort!

**The first three questions are simple and ensure that you read the instructions.**

**Incorrect responses to any of these three questions will result in nonpayment for the task. Not selecting a response from the dropdown menu will also result in nonpayment for the task. Selecting a response that says you will be blocked will lead to your being blocked.**

If you choose to perform multiple tasks for this you will easily be able to answer the questions again, but **incorrect responses to any of these three questions across more than two tasks will lead to your being barred from accessing any additional tasks for this project**

# MTurk Screening Questions

***The instructions were updated to include additional screening elements:***

Which of the below is NOT considered a museum according to the definition above?

Please Select an Option Below

- Please Select an Option Below
- Nature Centers
- Zoos
- If you select this, you will be blocked
- Aquariums
- Recreational Parks
- Botanical Gardens and Arboretums
- Parks that have a curated exhibit open to the public

to be a museum for the purposes c

d for this task?

Why is a non-museum art gallery NOT considered to be a museum for the purposes of this task?

Please Select an Option Below

- Please Select an Option Below
- It is a public art gallery
- It is a private art gallery
- It is a gallery primarily focused on selling art
- If you select this, you will be blocked
- It is a gallery primarily focused on displaying art
- Silver

ined for this task?

together and place the answer in the box 99

What should you answer below to get credit?

Please Select an Option Below

- Please Select an Option Below
- Red
- Blue
- If you select this, you will be blocked
- Green
- Orange
- Yellow

d a museu

OT consider

Is the place in the link above a museum as defined for this task?

Please Select an Option Below

- Please Select an Option Below
- If you select this, you will be blocked
- No, this is not a Museum
- Yes, this is a Museum
- If you select this, you will be blocked
- Unsure
- This is a non-museum art gallery

mbertwo} together and plac

S

# MTurk Museum Discipline Classification Task Instructions (Final)

**NOTE: The instructions are VERY important to getting paid for this task**

Each of the entities that you will review is believed to be some kind of museum, but we want to know what kind it is.

For this task, **Museums includes places like Zoos, Aquariums, Botanical Gardens, and Arboretums; Nature Centers; Science Centers; Answer Purple below to get credit; History Museums and Historic Sites; Art Museums; Children's Museums; Natural History Museums; and Specialized Museums.**

## **Instructions:**

1. Open and review the entities website(s) carefully.
  2. Provide the categorization you feel best fits the entity.
- If the entity falls under more than one category, but falls more closely under one category than another, choose the categorization that best fits the entity.
  - If the entity seems to fall under more than one category about equally - for example, if its exhibits fall about equally under both history and science - classify as "general museum."
  - If you are not sure if it is a museum at all, classify as "not a museum."

We appreciate your effort!



**The first two questions can be answered directly by reading the instructions above.** They are not trick questions; their answers are found above. They are asked simply to ensure that you have read the instructions.

**Incorrect responses to either of these two questions will result in nonpayment for the task. Not selecting a response from the dropdown menu will also result in nonpayment for the task. Selecting a response that says you will be blocked will lead to your being blocked.**

If you choose to perform multiple tasks for this you will easily be able to answer the questions again, but **incorrect responses to any of these two questions across more than two tasks will lead to your being barred from accessing any additional tasks for this project**

[Top of Form](#)

---

What should you answer below to get credit?

How should a museum whose exhibits fall about equally under both history and science be classified?

# MTurk Instructions: Grouped vs. Ungrouped Museum Disciplines

## Layouts for museum discipline grouping tests:

### Grouped

We appreciate your effort!

The first two questions can be answered directly by reading the instructions above. They are not trick questions; their answers are found above. They are asked simply to ensure that you have read the instructions.

**Incorrect responses to either of these two questions will result in nonpayment for the task:**

If you choose to perform multiple tasks for this you will easily be able to answer the questions again, but **incorrect responses to any of these two questions across more than two tasks will lead to your being barred from accessing any additional tasks for this project**

What should you answer below to get credit?

Red

How should a museum whose exhibits fall about equally under both history and science be classified?

Not a Museum

Link: <https://www.yelp.com/biz/pfeiffer-big-sur-state-park-big-sur> Identification Number: 160

What is the **best** way to classify this entity?

Art Museum  
Art Museum  
Botanical Garden/ Arboretum/ Nature Center  
Children's Museum  
General Museum  
Other/ Specialized museum  
History Museum/ Historic Site/ Historic House  
Natural History/ Anthropology Museum  
Science/ Technology Center/ Museum, Planetariums  
Zoo/ Aquarium  
Not a museum

ACCEPT the HIT before you can submit the results.

### Ungrouped

We appreciate your effort!

The first two questions can be answered directly by reading the instructions above. They are not trick questions; their answers are found above. They are asked simply to ensure that you have read the instructions.

**Incorrect responses to either of these two questions will result in nonpayment for the task:**

If you choose to perform multiple tasks for this you will easily be able to answer the questions again, but **incorrect responses to any of these two questions across more than two tasks will lead to your being barred from accessing any additional tasks for this project**

What should you answer below to get credit?

Red

How should a museum whose exhibits fall about equally under both history and science be classified?

Not a Museum

Link: <https://www.yelp.com/biz/pfeiffer-big-sur-state-park-big-sur> Identification Number: 160

What is the **best** way to classify this entity?

Anthropology Museum  
Anthropology Museum  
Aquarium  
Arboretum  
Art Museum  
Botanical Garden  
Children's Museum  
General Museum  
Historic House and/or Site  
History Museum  
Natural History Museum  
Nature Center  
Planetarium  
Science and Technology Center/Museum  
Specialized Museum  
Zoo

ACCEPT the HIT before you can submit the results.

FEATU

Fully managed data labeling service. Learn more »

Link: [\\${url}](#) Identification Number: **#{ID}**

What is the best way to classify this entity?



- Select a response below
- If you select this, you will be blocked
  - Anthropology Museum
  - Aquarium
  - Arboretum
  - Art Museum
  - If you select this, you will be blocked
  - Botanical Garden
  - Children's Museum
  - If you select this, you will be blocked
  - General Museum
  - Historic house and/or site
  - History Museum
  - Natural history museum
  - Nature Center
  - If you select this, you will be blocked
  - Planetarium
  - Science and Technology Center/ Museum
  - Specialized Museum
  - Zoo
  - Not a museum

# MTurk Contact Information Task Instructions

## Data Collection Instructions/Codebook

**NOTE: The instructions are VERY important to getting paid for this task**

### Instructions:

1. Open the Yelp link. Next, open the entity's website.
2. Locate and enter the best phone number and email address you can identify for someone wanting to reach out to the location. Type green to get credit. Some entities may require further investigation through a web search beyond the listed website in order to find good contact information.

We appreciate your effort!

**Missing responses to any of these three questions will result in nonpayment for the task. Failing to type the correct answer to the "credit" question will result in nonpayment for the task. Failing to correctly follow these instructions more than twice will result in being blocked:**

[Top of Form](#)

Link: [\\${url}](#) Identification Number: **}\${ID}**

What should you type to get credit:

Please find and enter an email address for the location:

Please find and enter a phone number for the location:

Submit

\*Note: The second iteration of this task also provided the business URL of the museums (from Yelp, other matched sources and scraped using ChatGPT)