

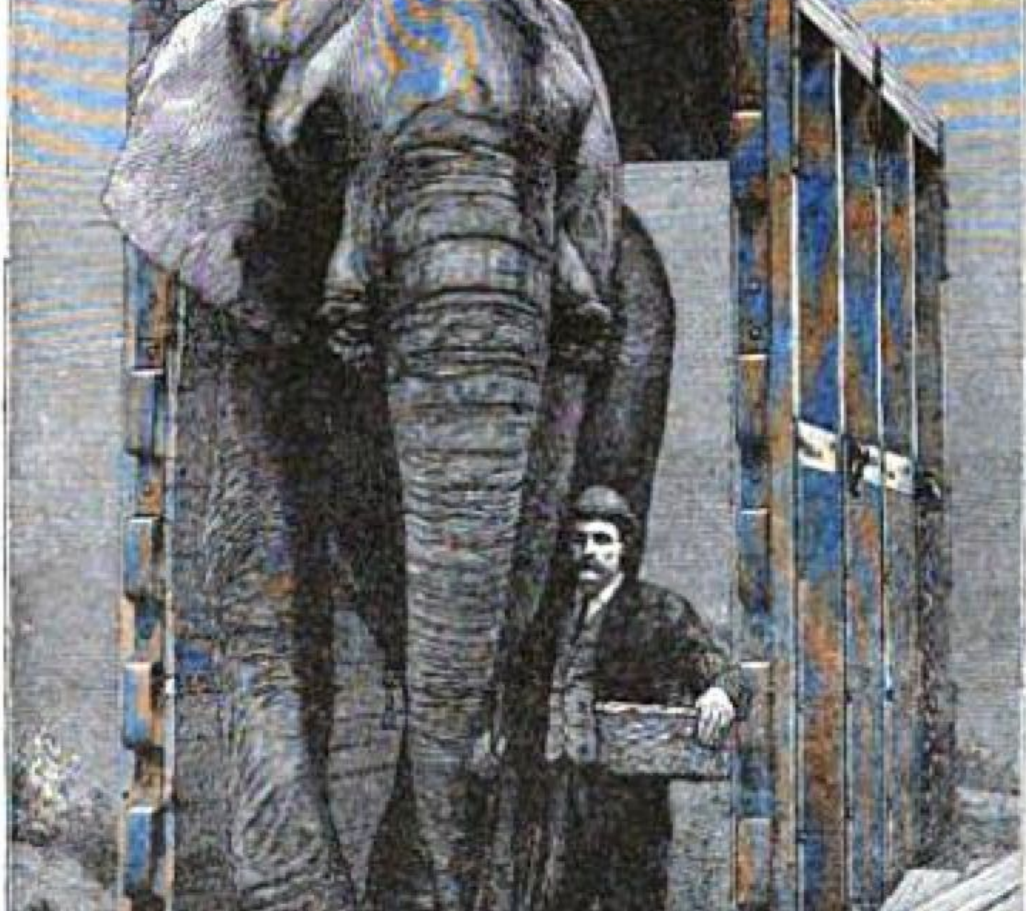
# Scribe

# Toward a General Framework for Community Transcription

Paul Beaudoin | New York Public Library Labs

@nonword | paulbeaudoin@nypl.org

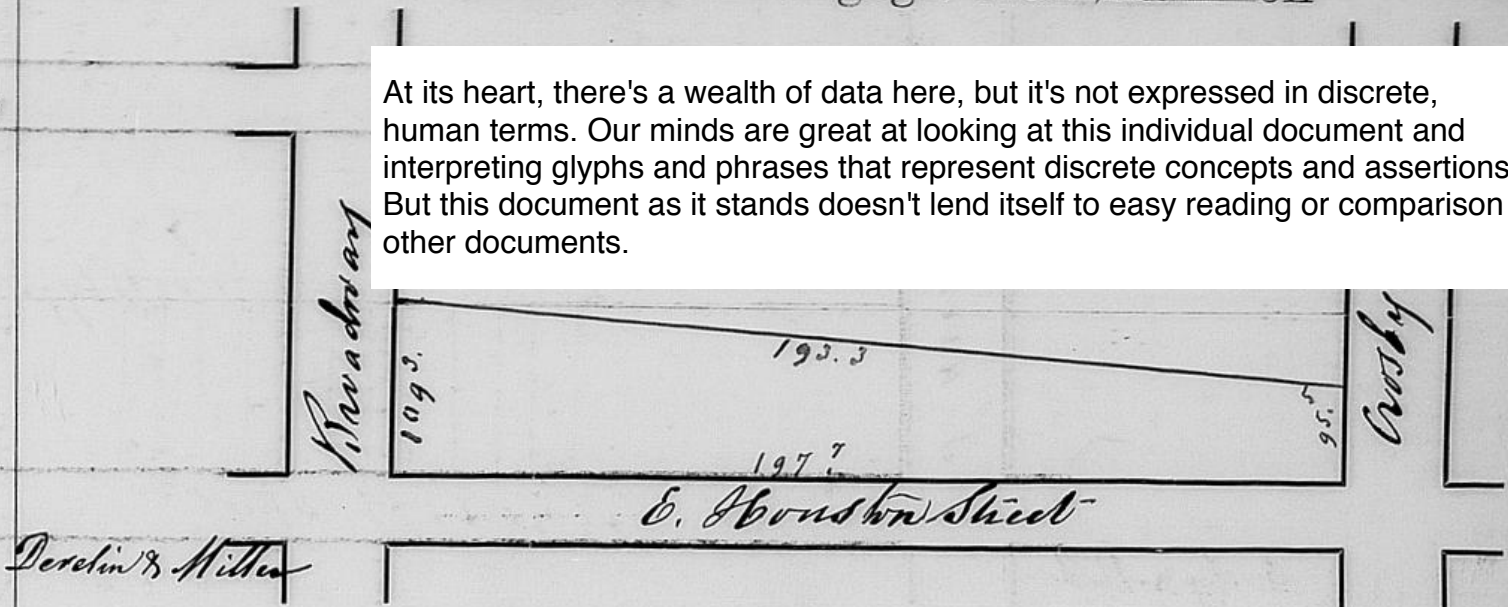
This is a talk about Scribe, a framework for community transcription but I mostly want to encourage people to think about things we can do to aid discovery post-digitization



Cultural institutions of a certain age have this problem of having a lot of data that's stored away in forms that are difficult to analyze.

1880  
December 13

At its heart, there's a wealth of data here, but it's not expressed in discrete, human terms. Our minds are great at looking at this individual document and interpreting glyphs and phrases that represent discrete concepts and assertions. But this document as it stands doesn't lend itself to easy reading or comparison to other documents.



Mortgager, *Phineas J. Barnum*  
No. of Street or Avenue, *Broadway N.E. corner Houston Street*  
Dimensions of Ground, *109.<sup>3</sup> on Broadway, 197.<sup>7</sup> on Houston Street and 95.<sup>5</sup> on Crosby St.*  
Description of Building, *Various stores and dwellings cover the ground*  
Valuation of Land, *\$ 350,000*  
" " Building, *" 50,000 } \$ 400,000.*  
Amount Loaned, *\$ 200,000.*

# Digitization should include data extraction

We should strive to extract semantic content from the array of pixels that digitization produces.

Because frequently it's not about the image. Frequently there's a structure of data inside that document that is more useful for discovery, aggregate analysis.

Maybe we can build a general tool for data extraction that uses humans.

SCRIBE

Document transcription, crowdsourced



NYPL and Zooniverse got together in late 2014 to collaborate on an NEH funded project to solve the problem of data extraction IN GENERAL



SHAKESPEARE'S WORLD



SNAPSHOTS AT SEA



JUNGLE RHYTHMS



CHIMP & SEE

**zooniverse.org**

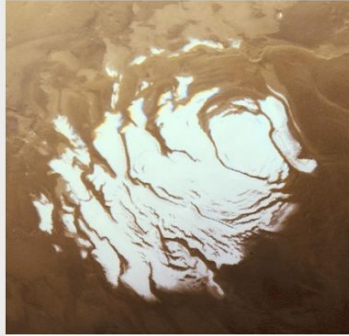
ANNOTATE



SCIENCE GOSSIP



WILDEBEEST WATCH



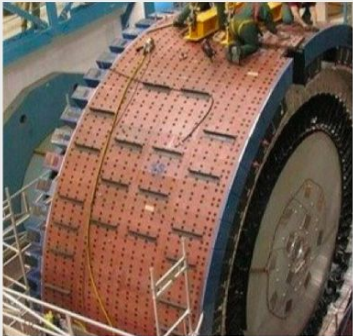
PLANET FOUR: TERRAINS



OLD WEATHER



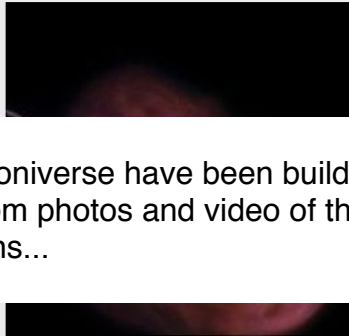
GALAXY ZOO



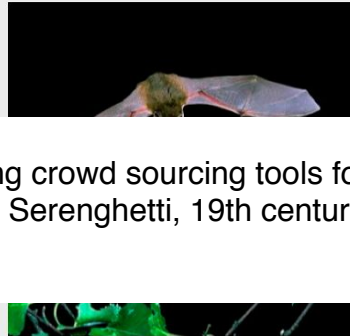
HIGGS HUNTERS



FLOATING FORESTS



RADIO GALAXY ZOO



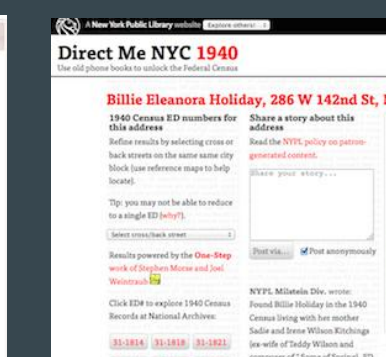
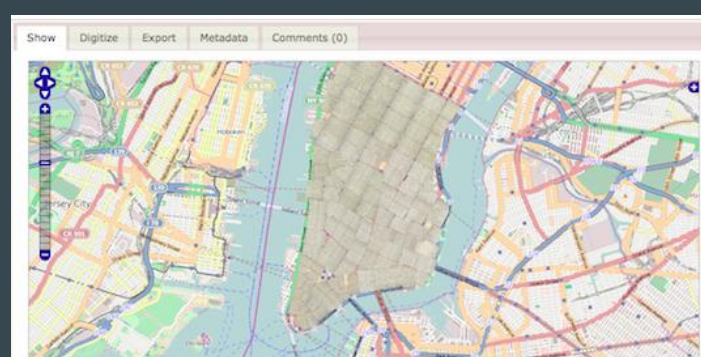
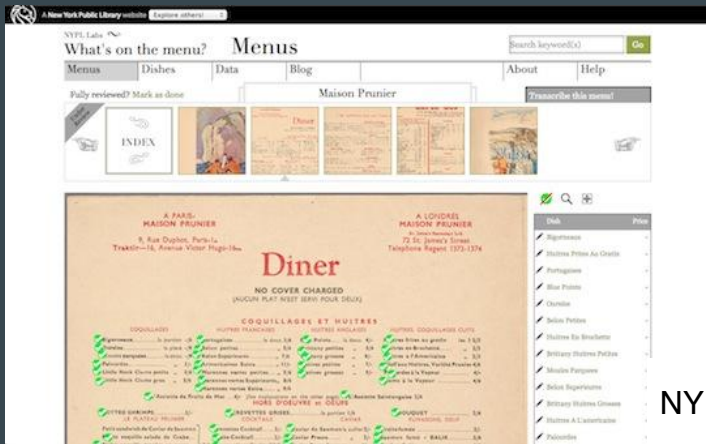
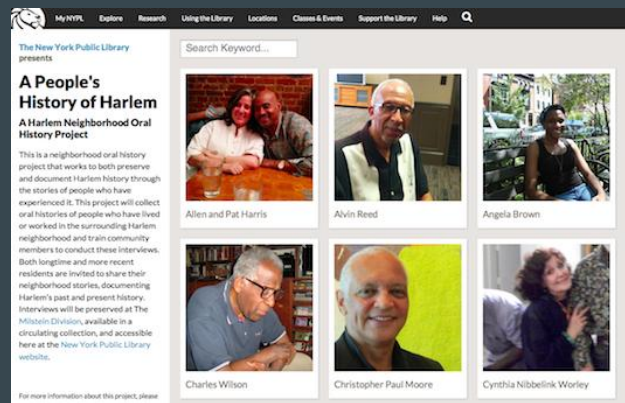
BAT DETECTIVE



CHICAGO WILDLIFE WATCH

The Zooniverse have been building crowd sourcing tools for a while to extract data from photos and video of the Serengeti, 19th century ships logs, arctic penguins...





NYPL Labs has produced a number of projects around data extraction.

Labs & Zooniverse came together to pool experience and resources to build a general purpose community data extraction tool.



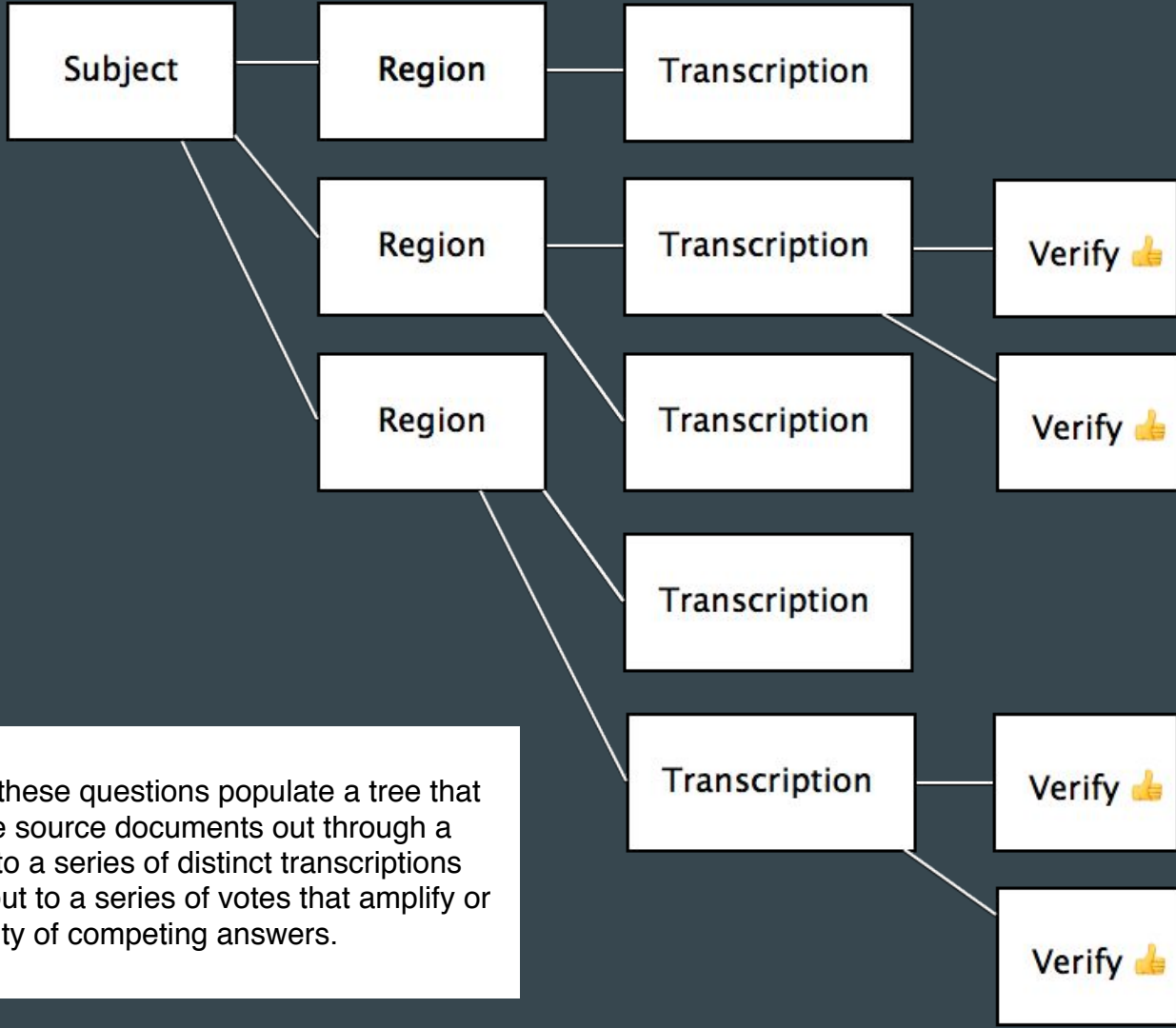
# Scribe: Configured around questions

Questions about the document

Questions about the locations of things

Questions about those things

Questions about the answers to those questions



The answers to these questions populate a tree that extends from the source documents out through a series of marks to a series of distinct transcriptions and potentially out to a series of votes that amplify or reduce the validity of competing answers.

# Scribe: Emigrant City

Running 4 months

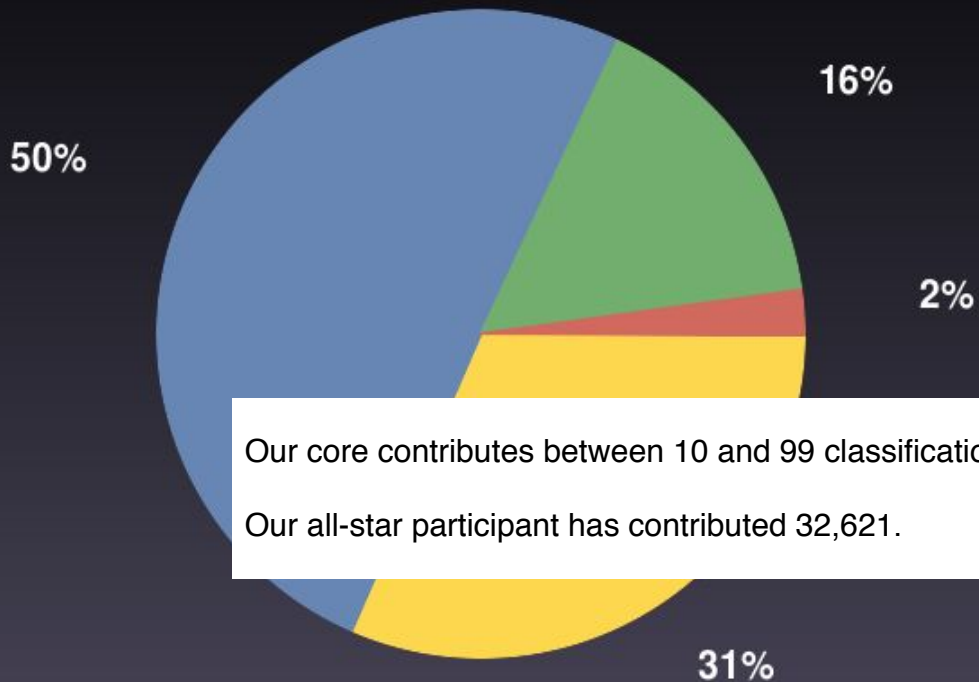
5,000+ unique contributors

500,000+ classifications

Emigrant City is a project around mortgage records from the Emigrant City Savings Bank in NYC late 19th, early 20th centuries.

# Contributor Types

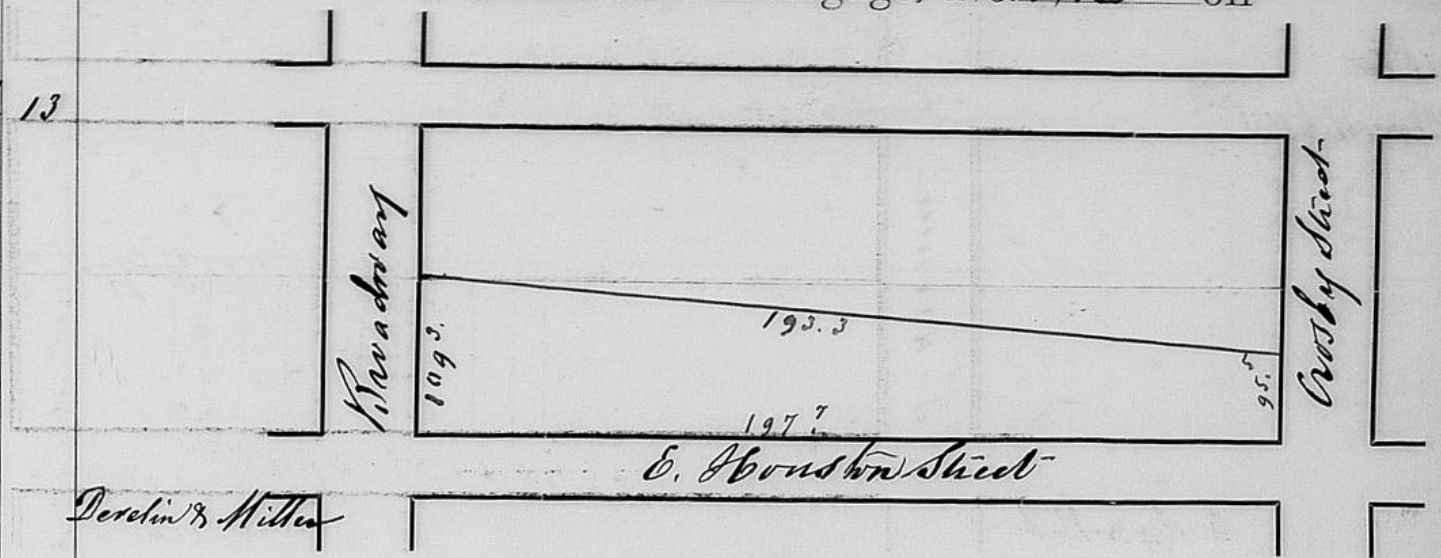
■ 1 - 9   ■ 10 - 99   ■ 100 - 499   ■ 500 +



Our core contributors between 10 and 99 classifications.

Our all-star participant has contributed 32,621.

1880  
December 13



Mortgager, *Phineas J. Barnum*

No. of Street or Avenue, *Broadway N.E. corner Houston Street*

Dimensions of Ground, \_\_\_\_\_

Description of Building, \_\_\_\_\_

Valuation of Land, \_\_\_\_\_

“ “ Building, \_\_\_\_\_

Amount Loaned, \_\_\_\_\_

The Emigrant City flow works like this:

1. One day a user was presented with this and asked to mark the "Amount Loaned"

000

197?  
E. Houston Street  
18 Miller

Phineas J. Barnum  
Broadway N.E. corner Houston Street  
109.3' on Broadway, 197.7' on Houston St and 95.5' on Broadway St.  
Various stores and dwellings cover the ground  
\$ 350,000 }  
" 50,000 } \$ 400,000.

\$ 200,000.

2. A different contributor, encountered that user's mark and typed what they saw

Search forum

Discuss this page.

Within 20 minutes, three distinct users encountered the same mark and transcribed the same text. This immediately verified the transcription as valid, promoting it to consensus data without the need for further verification.

Enter the record date

e.g. "1867 May 11" "1867 July 30" "1875 December 31"

\$200.000

Need some help?

Bad region

Illegible?

Return to Marking

\$200.000.

Desdin & Miller

Phineas J. Ba

197?  
E. Houston Street

Mortgager,

No. of Street or Avenue,

Dimensions of Ground,

Description of Building,

Valuation of Land,

" " Building

Amount Loaned,

Broadway  
109<sup>3</sup> on Broad

Various stones and dwellings cover the ground

\$250,000 & \$400,000.

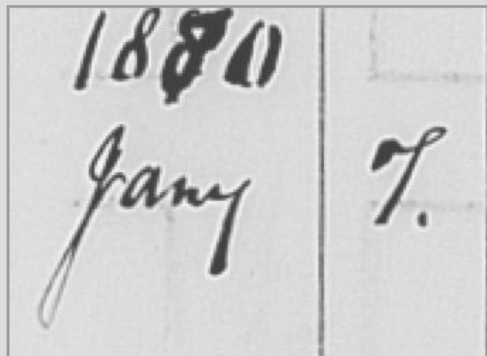
# RECORD DATE

---

1880 Jany 7. ( Interpretted as *1880-01-07T00:00:00Z* )

Confidence: **67%** Status: **awaiting votes** Distinct Transcriptions: 2

Hide region



Recent experimental work (at writing, not yet in master) adds to Scribe a facility for mapping the answers to these questions to a custom schema, with optional/mandatory, repeatable fields with specific types like date, int, dimensions, etc.

We see that applied here to a not perfectly confident consensus transcription with two distinct transcriptions - probably disagreeing on the year, which is either 1870 or 1880.

We've passed this free text transcription through some regex/ date parsing magic to derive a fully qualified DATE value.



About

Browse

Download

Tips & Tricks

# Browse

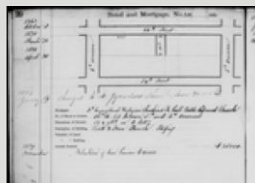
Preview the data by searching by keyword below:

Record Date 

1885 - 1890-03-15

Search

Found 1074 matches



Record Date: 1886  
January 16; 1879  
November; 1863 October  
3; 1870. March 31.; 1871  
April 20



Record Date: 1882  
February 4; April 27/85

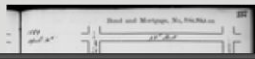


Record Date: 1889  
Novemb. 1

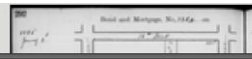


Record Date: 1889  
January 23

This allows us to run smart queries against this field AS A DATE.  
Here I'm filtering on mortgage records between 1885 (jan 1) and March 15, 1890



Record Date: 1889 April  
20th



Record Date: 1885  
January 21

Amount Loaned

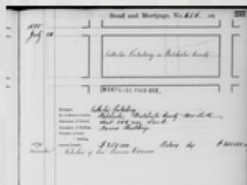
200000-

Search

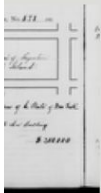
Found 61 matches



Amount  
Loaned: \$400,000



Amount  
Loaned: \$250,000



Amount  
Loaned: \$200,000



Amount  
Loaned: \$200,000



Amount  
Loaned: \$250,000



Amount  
Loaned: \$200,000

Similarly, because we're interpreting the free text transcription of the Amount Loaned field as "monetary", making common substitutions and stripping non-numerics, we can run queries like this to inspect the highest payout mortgages.

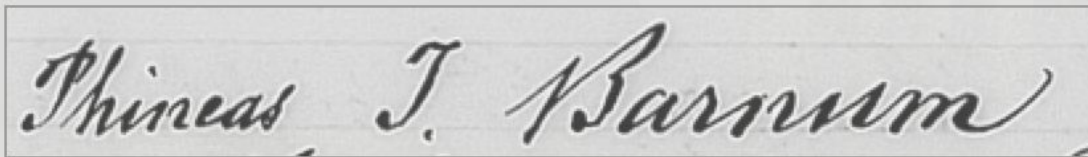
# MORTGAGER

---

**Phineas T. Barnum**

Confidence: **67%** Status: **awaiting votes** Distinct Transcriptions: 2

[Hide region](#)

A rectangular box containing a handwritten signature in cursive script that reads "Phineas T. Barnum". The signature is written in black ink on a light-colored background with faint horizontal lines.

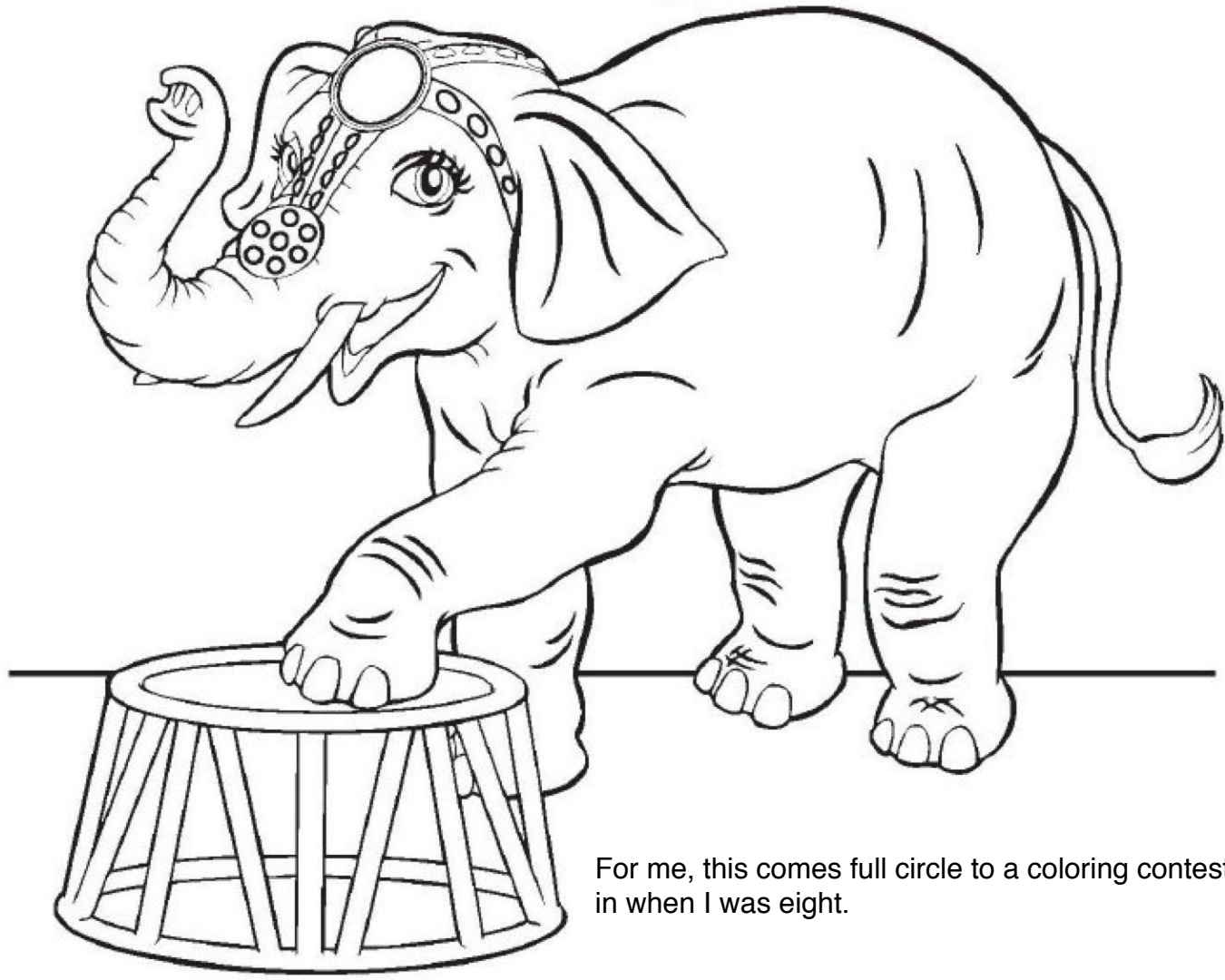
Folks familiar with the big players in late 19th century Manhattan will recognize this name

# STREET ADDRESS

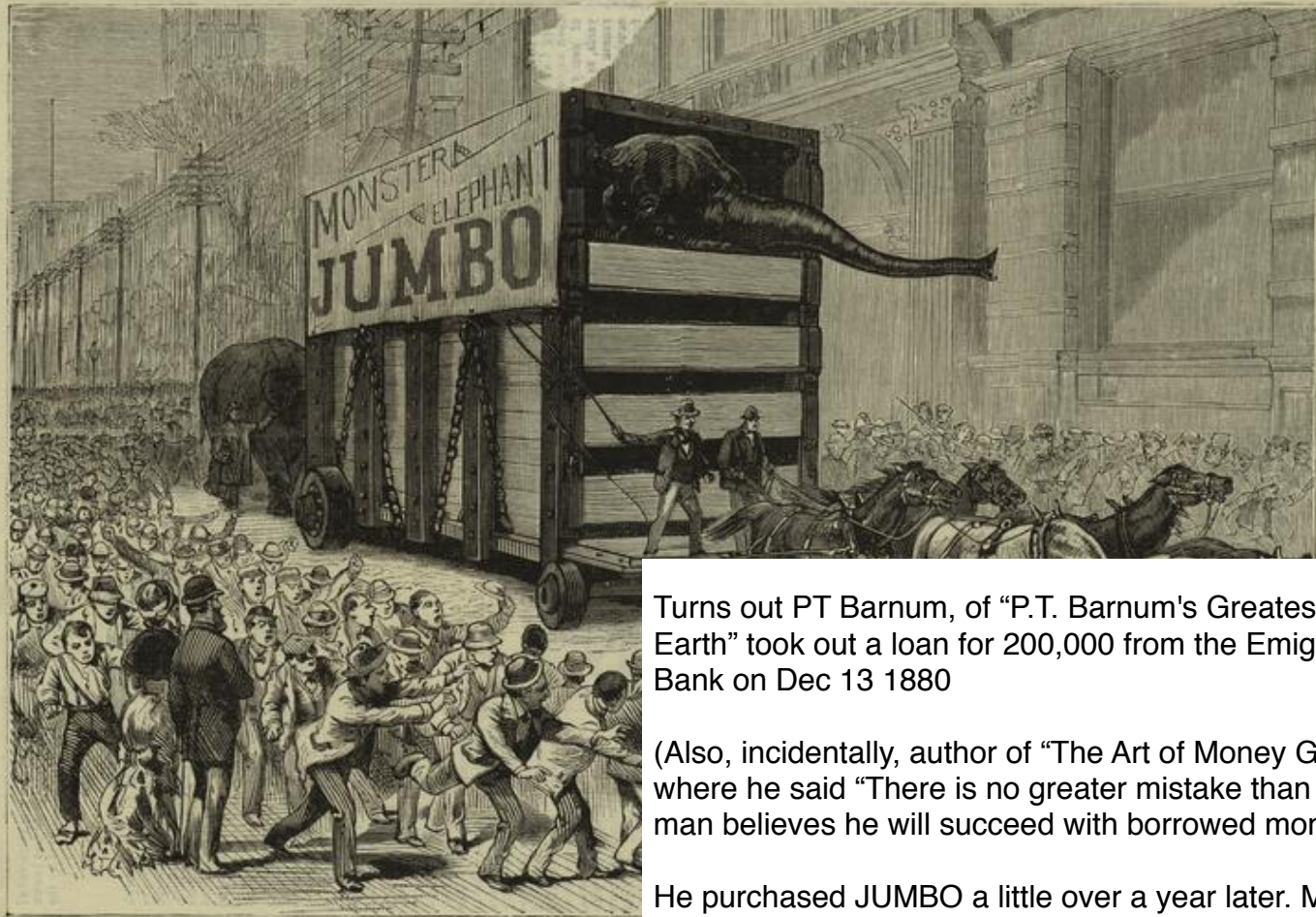
---

**Broadway N.E corner Houston Street**

Confidence: **33%** Status: **awaiting votes** Distinct Transcriptions: 3



For me, this comes full circle to a coloring contest I participated in when I was eight.



Turns out PT Barnum, of “P.T. Barnum's Greatest Show On Earth” took out a loan for 200,000 from the Emigrant Savings Bank on Dec 13 1880

(Also, incidentally, author of “The Art of Money Getting” in 1880, where he said “There is no greater mistake than when a young man believes he will succeed with borrowed money.”)

He purchased JUMBO a little over a year later. Maybe using some borrowed funds.

JUMBO'S ARRIVAL IN

G. New York City - Parades - 1882

# /Scribe Toward a General Framework for Community Transcription

[scribeproject.github.io](https://scribeproject.github.io)  
[emigrantcity.nypl.org](https://emigrantcity.nypl.org)  
[labs.nypl.org](https://labs.nypl.org)

Paul Beaudoin | New York Public Library Labs  
@nonword | [paulbeaudoin@nypl.org](mailto:paulbeaudoin@nypl.org)

We encourage you to check out Scribe & Emigrant City as experiments in this greater effort to build generalized tools to extract data from our digitized assets.