White Paper Report

Report ID: 104596

Application Number: HD5148011

Project Director: Lacy Schutz (lschutz@mcny.org)

Institution: Museum of the City of New York

Reporting Period: 9/1/2011-1/31/2013

Report Due: 4/30/2013

Date Submitted: 4/30/2013

Final Report
NEH Grant No. HD-51480-11

Improving Digital Record Annotation Capabilities with Open-sourced Ontologies and Crowd-sourced Workers

Project Director: Lacy Schutz
Museum of the City of New York

April 30, 2013

Project Activities

The Museum of the City of New York received funding from the National Endowment for the Humanities (NEH) to collaborate with Tagasauris, a New York-based startup, on an innovative project to increase the accessibility and discoverability of its collections to researchers and the public. This project stemmed from a key challenge the Museum faces: its two imaging staff could scan and process up to 700 digital images per day, a much faster rate than its two catalogers, who could annotate about 120 records per day. There was a backlog of approximately 24,000 object records that had digitized images but nonexistent or inadequate metadata. This annotation problem could not be remedied simply by hiring additional catalogers; the Museum lacked both the physical space and the budgetary resources to add the two or three full-time catalogers who would be required to keep pace with its photographers, let alone address the backlog.

Tagasauris proposed a solution for this problem by developing an online platform that divides the work of annotating digital object records into micro-tasks that can be completed by huge labor pools available through crowd-sourced marketplaces. Crowd-sourced marketplaces have the potential to provide millions of annotations, allowing museum or library staff to shift their focus from metadata creation to higher-value tasks that utilized their expert knowledge, including overseeing and reviewing the work of crowd-sourced workers.

NEH funding supported a Digital Humanities Start-Up project to develop a micro-task structure and test the Tagasauris online platform with crowd-sourced workers. To ensure that metadata terms were accessible to a general audience, Tagasauris first reconciled the Museum's standard cataloging lexicon with sources of semantic, linked open data (a method for sharing and connecting pieces of data, information, and knowledge on the Web, such as Wikipedia). Using this new ontology, Tagasauris syndicated a set of micro-tasks to Amazon’s Mechanical Turk, where 8,639 crowd-sourced workers enriched digital records for 51,671 objects by adding 472,628 tags with descriptive metadata using the Tagasauris platform. The project was carried out in three stages, detailed below.

Stage 1: Creation, development, and use of an open-source ontology

The first task in Stage 1 was to publish the traditional lexicon currently in use at the Museum as a linked data resource on the semantic Web, using Freebase (an open, Creative Commons-licensed repository of structured data) as the semantic Web platform. Tagasauris completed this task in December 2011, publishing 37,083 rows of data. Tagasauris then reconciled the entities in the Museum’s lexicon with the existing entities in Freebase by linking terms wherever there was overlap. 22,000 terms had clear equivalent entities in Freebase and were reconciled with those entities. This left 15,083 entries from the Museum’s lexicon with no good match in Freebase. 12,856 of these entries were low-confidence terms that could correspond with multiple possible entities in Freebase, all of which might be incorrect. For example, an object in the Museum’s collection tagged with the term “Majestic Theatre” could correspond with any of three different theaters by that name that have operated in the New York City area since the 19th century. Additionally, there were 2,227 entries with no reconciliation data at all—terms for which no corresponding entity existed in Freebase or Wikipedia. Tagasauris contributed each of the 12,856 low-confidence and the 2,227 irreconcilable subject terms to Freebase. This prepared the lexicon portion of the project for the launch of the crowd-sourcing metadata annotation platform in Stage 3.

Prior to launch, the Museum and Tagasauris undertook two activities to increase the number of reconcilable Museum subject terms in Freebase, in order to improve the quality of annotation and subsequent search results. First, the Museum exported additional contextual metadata from its collections management system to help disambiguate between potential multiple matches in Freebase, and to populate potential new Freebase entries for those terms with no existing matches. This additional metadata helped enrich the quality of the entities created in Freebase.

Second, Tagasauris also implemented a new quality-control micro-task to capture input on two of the most common problems experienced during lexicon reconciliation: choosing the Freebase entity that best reconciles to an MCNY subject heading, referred to as “reconciliation”; and evaluating whether two different entities should be considered the same, referred to as “merge or match.” These are situations where human judgment is unsurpassed in quality and is essential to ensuring reliable data.

Stage 2: Design of the micro-task architecture
With input from the Museum’s collections staff, Tagasauris designed a task architecture consisting of interconnected micro-tasks and associated actions, as described below, to deliver the exact metadata the Museum required:

- **Descriptive Tags**: Identification of main subject(s) and action(s) in the image
- **Caption Tags**: Relevant tags extracted from the caption (if available)
- **Tag Day/Night**
- **Tag Indoors/Outdoors**
- **Tag Photographic Styles**: Aerial, bird's eye, streetscape, background, cross-processed, full frame, full length, head and shoulders, landscape, lens flare, looking at camera, low angle view, multiple exposure, panoramic, portrait, scenic, still life, studio shot, urban scene, wide angle, X-ray image
- **Tag People Count**: None, 1, 2, 3, 4, group
- **Tag Gender**: Male, female
- **Tag Age**: Infant (birth to age 2), child (3 to 12), teen (13 to 19), young adult (20 to 35), mid adult (36 to 65), senior adult (66 and older)
- **Tag Geolocation**: Currently location is mapped to Google Maps using Named Entity Recognition or human judgment. At the suggestion of Museum staff, Tagasauris is looking into adding the capability to mark cross-streets.
- **Tag Emotion** *(as needed)*: Joy, trust, fear, surprise, sadness, disgust, anger, anticipation, optimism, love, submission, awe, disapproval, remorse, contempt, aggressiveness, etc. This micro-task was also suggested by Museum staff for application to special collections where emotional content is likely to be depicted (e.g. Currier & Ives prints).
- **Quality control**: Crowd-sourced workers voted on the appropriateness of fellow workers’ tags.

- **Semantic reasoning module**: Presents the possibility to reconcile similar terms. For example, when tagging geolocation, it would reconcile terms such as “14th Street and Park Avenue South” in New York City to adjacent locations like Union Square.

For each of these micro-tasks, Tagasauris identified a subset of the ontology to be used. For example, if an annotation micro-task required the worker to identify a location depicted in an image, the subset of the ontology used for this micro-task included only location info. To ensure quality results, Tagasauris created thorough video instructions for each micro-task, hosted at http://instructions.tagasauris.com/.

Each task was performed by at least three workers. In terms of workflow, Tagasauris experimented with requiring crowd-sourced laborers to work through all of these micro-tasks for a single image, and with letting workers self-select the micro-tasks they prefer. The latter option produced better results, as demonstrated through higher-quality tagging, higher levels of worker engagement, and faster throughput.

Tagasauris investigated and identified appropriate labor channels from crowd-sourced marketplaces to determine which ones provided the highest quality of labor and final product by thoroughly investigating available crowd-sourced marketplaces, including Amazon’s Mechanical Turk, oDesk, and Museum followers on social networks like Twitter and Facebook.

As a result of this analysis, all micro-tasks were designed to run using the interface for Amazon’s Mechanical Turk, a piecework marketplace in which workers are paid by the number of tasks they complete. Mechanical Turk was selected because it has by far the largest degree of engagement from workers than any other competitor. It has a mature, proven application programming interface (API) and a micro-labor market that includes more than 500,000 workers. This API-driven approach presents an optional way for Tagasauris to engage micro-labor. It transforms fixed human costs into variable expense and does not require long-term contracts or minimum commitments. Additionally, the Mechanical Turk system allowed Tagasauris to publish tasks of its own design into the marketplace.

Tagasauris also experimented with oDesk, a marketplace for hourly rather than piecework laborers, by hiring a group of oDesk workers to log into the Mechanical Turk interface and perform tasks under the direction of Tagasauris staff. Via Skype, chat, and instant messaging, Tagasauris directed the laborers to focus on a specialized task for a subset of the collection, allowing them to accomplish that task more quickly or guiding them to produce better-quality results. This guided interaction mimicked the sort of oversight a Museum cataloger might provide to crowd-sourced workers, if this project were to be implemented on a larger scale. It has produced such good results in testing that Tagasauris is now building chat and instant messaging capabilities into all of its tasks and workflows.

Stage 3: Annotation of the Museum’s digital assets using the open-source ontology and the Tagasauris crowd-sourcing platform

Tagasauris created digital workflows encapsulating the interconnected micro-tasks, actions, ontologies, and labor channel preferences. Please refer to *Image 1* below for a screenshot of the Museum's workflow as seen on the Tagasauris interface. In preparation for testing the annotation system, the Project Director/Director of Collections, Lacy Schutz, selected objects from the Museum's collection that had already been tagged with metadata by professional staff catalogers, pulling together a sample that represented the current range of existing metadata for objects within the collection, from minimal to extensive. In April 2012, Tagasauris began syndicating jobs to the target labor channels for all micro-tasks. Over the next nine months, April 2012 to January 2013, 8,639 unique crowd workers tagged 51,671 objects with descriptive metadata using the Tagasauris platform and micro-tasks, completing a total of 172,193 tasks that resulted in 472,628 metadata tags. The workers were geographically restricted to U.S. IP addresses.

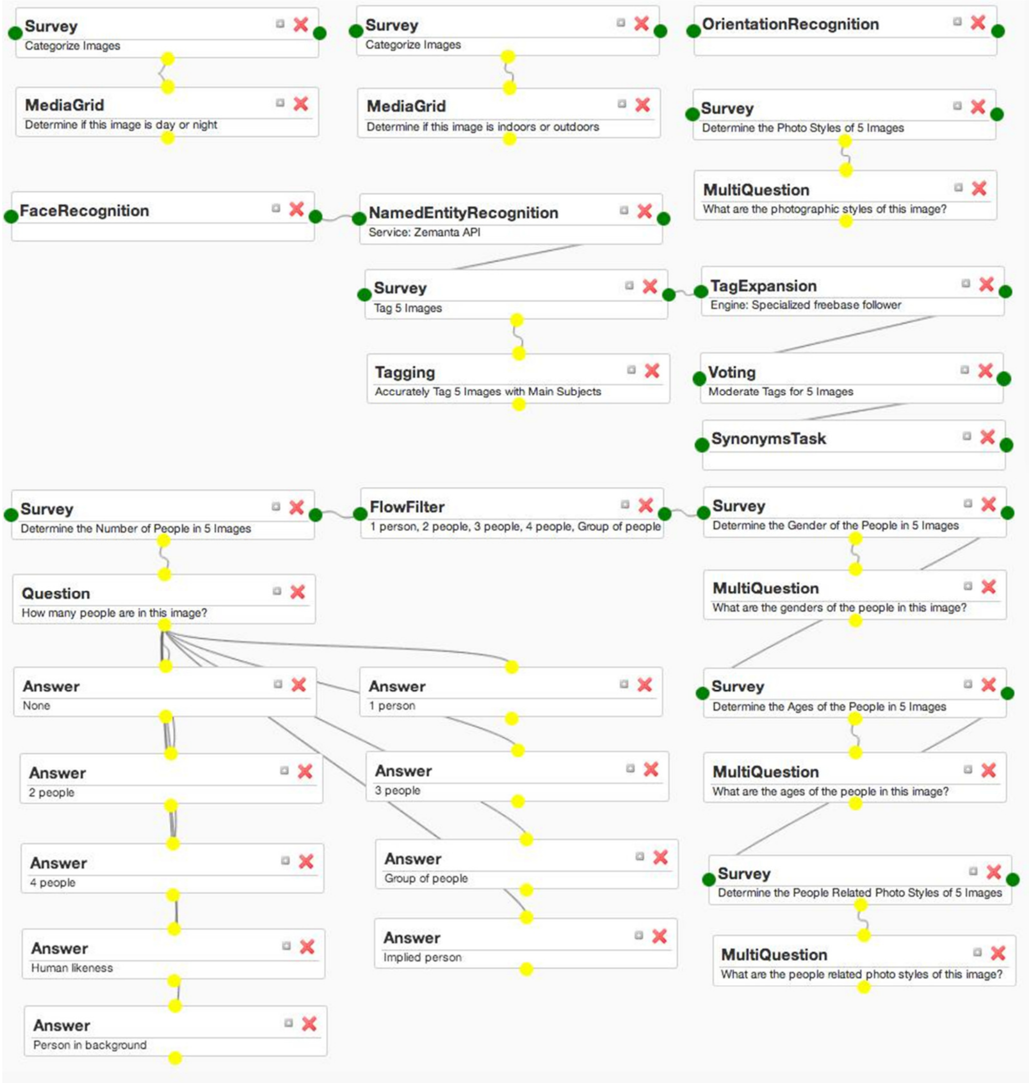

Survey
Categorize Images
MediaGrid
Determine if this image is day or night
Survey
Categorize Images
MediaGrid
Determine if this image is indoors or outdoors
OrientationRecognition
Survey
Determine the Photo Styles of 5 Images
MultiQuestion
What are the photographic styles of this image?
FaceRecognition
NamedEntityRecognition
Service: Zemanta API
Survey
Tag 5 Images
Tagging
Accurately Tag 5 Images with Main Subjects
TagExpansion
Engine: Specialized freebase follower
Voting
Moderate Tags for 5 Images
SynonymsTask
Survey
Determine the Number of People in 5 Images
Question
How many people are in this image?
FlowFilter
1 person, 2 people, 3 people, 4 people, Group of people
Survey
Determine the Gender of the People in 5 Images
MultiQuestion
What are the genders of the people in this image?
Answer
None
Answer
1 person
Answer
2 people
Answer
3 people
Answer
4 people
Answer
Group of people
Answer
Human likeness
Answer
Implied person
Answer
Person in background
Survey
Determine the Ages of the People in 5 Images
MultiQuestion
What are the ages of the people in this image?
Survey
Determine the People Related Photo Styles of 5 Images
MultiQuestion
What are the people related photo styles of this image?

Image 1: Screenshot of the Museum's custom workflow from the Tagasauris interface

While micro-task jobs were in progress, Tagasauris collected annotations and continually monitored workers, product quality, and the progress made. Once crowd-tagging was completed in January 2013, Tagasauris then began to compare the results of the crowd-sourced tagging to the Museum-generated metadata it already had for each object. The results of this comparison are discussed below under Evaluation.

While the final results of the project will not be publicized widely until the Museum distributes its NEH white paper following the submission of this report, the Museum and Tagasauris publicized their ongoing activities and preliminary results at professional conferences and meet-ups throughout the project. These activities are described further under Audience and Dissemination.

Accomplishments

The project team successfully accomplished all activities outlined in our three-stage work plan and completed an extensive evaluation of the results, which confirmed that we had met and exceeded the project goals.

In Stage 1, we delivered the Museum's controlled vocabulary as a linked data resource using Freebase as the semantic Web platform. The new linked lexicon includes entities from Wikipedia, Open Libraries Project, the National Register of Historic Places, and WorkNet.

In Stage 2, we delivered 15 micro-tasks, including the configured actions, ontologies, instructions and labor channel preferences. For each micro-task, we identified subsets of the Museum's lexicon to be used. We selected Amazon's Mechanical Turk as the default labor channel for these tasks.

In Stage 3, we configured a digital workflow that encapsulated and interconnected the 15 micro-tasks. We created jobs that syndicated images to be annotated to the target labor channels for all micro-tasks. We collected annotations and continually monitored workers, quality, and progress made.

We were successful in coordinating the technical and consulting services of Tagasauris and Orange Logic to virtually integrate the Tagasauris annotation platform with a test/staging platform based on the Museum's Orange Logic collections management platform. This allowed for the automated round-trip flow of collections metadata from the Museum, to the Tagasauris platform, Amazon Mechanical Turk, and back.

Subsequent to the completion of Stage 3, we have created and launched three additional micro-tasks to assist with the statistical evaluation of the results we had acquired.

Audience and Dissemination

The potential audience for the work tested in and resulting from this Digital Humanities Start-up project is vast, encompassing collections and digital information staff at museums, libraries, and humanities organizations across the United States. Because testing was just completed and the work has not yet been implemented on a large scale, the actual audience so far been limited to attendees at conferences and meet-ups where Museum or Tagasauris staff conducted presentations about the ongoing activities or preliminary results of the project. This audience has consisted primarily of professionals, scholars, and students working in the fields of digital humanities and information management.

The multidisciplinary nature and approach of the project appeals to constituents across a broad range of technology and information subfields, including computer science, computer vision and signal processing, cognitive science, web science, semantic web, and machine learning communities. All of these audiences are invested in addressing the grand challenge posed by the "Semantic Gap"--the large disparity between descriptions of multimedia content that can be computed automatically, and the richness and subjectivity of semantics in user queries and human interpretations of audio-visual media--that constitutes the core of this project.

The project was highlighted in one major press article during the grant period. On September 20, 2012, Tiffany Crawford, blogger for the Huffington Post, published a profile of Todd Carter, CEO of Tagasauris, that featured his work with the Museum.

Beyond this piece of press coverage, Museum and Tagasauris staff concentrated on promoting the project at events engaging audiences of humanities scholars and students; professionals and students of archival management, library sciences, and informatics; and professional special interest groups, such as the Linked Open Data in Libraries, Archive, and Museums (LODLAM) community and the Galleries, Libraries, Archives, and Museums (GLAM) communities. Outreach to these constituents was conducted at the following events:

On February 23, 2012, Todd Carter, participated in a regional Linked Open Data for Libraries, Archives, and Museums event for the Cultural Heritage Sector in the New York Metropolitan Area. Co-organized by the Metropolitan New York Library Council, The New York Public Library's NYPL Labs, and New York University, the event focused on cutting-edge semantic web technologies and furthered the goals defined in the World Wide Web Consortiums Library Linked Data Incubator Report and the various outputs of the Stanford Linked Data Workshop. During the morning plenary session, which was intended for a general audience of 175 attendees, Carter participated in a panel discussion and Q&A about relevant technologies, best practices and current trends with Corey Harper (NYU Libraries), Evan Sandhaus (NY Times Research & Development) Seth van Hooland (Université Libre de Bruxelles), Erik Mannes (Ghent University - IBBT), and Rebecca Guenther (Library Consultant, formerly Library of Congress). During the afternoon session, which was designed for metadata experts and application developers, Carter led a breakout session exploring Linked Data in practice, in which 40 participants were split among three breakout groups.

On March 7, 2012, Carter participated in a panel discussion hosted by the Metropolitan New York Library Council Digitization Special Interest Group, entitled “Semantic Technologies & Linked Data for Digitized Collections.” An audience of 175 attendees listened as Carter, Evan Sandhaus (NY Times), Ben Vershbow (Manager, NYPL Labs), and Doug Reside (Digital Curator of Performing Arts, NYPL) discussed the challenges and benefits of launching linked data and digital crowdsourcing projects. The panel addressed ways that sharing this type of knowledge work with the general public serves their institutions and change the ways that staff interact with their public.

On August 7, 2012, the Museum's Manuscripts and Ephemera Archivist, Lindsay Turley, and Tagasauris's Director of Product Development, Arpi Mardirossian, led a presentation at the Society of American Archivists annual meeting in San Diego, California. Addressing an audience of 60 information and library science professionals, they gave an illustrated talk on the preliminary results of the tagging project, discussing how the Tagasauris platform and micro-task structure has provided granular, consistent metadata with a dramatic increase in the rate of production.

On October 2, 2012, the Museum and Tagasauris jointly hosted an event called "Humanities and Technology Unite," which centered around the digital record annotation project. Held on site at the Museum of the City of New York, this event attracted more than 200 attendees and was produced by Sharon Middendorf, Director of Operations for Tagasauris . (Images are available online, and a video of the panel is available upon request.) Guest speakers discussed their views and ideas around topics such as crowdsourcing, human computation, and semantic web technologies, and how these new developments are changing the landscape for humanities organizations. In addition to Lacy Schutz and Todd Carter, the panel included five experts addressing interrelated topics:

- Perry Collins, Humanities Administrator, NEH - The Challenges faced by Humanities Organizations
- David Lipsey, Chairman, Digital Asset Management & Principal, Optimity Advisors - Re-contextualizing the DAM problem
- Sharon Chiarella, Vice President, Amazon Mechanical Turk - Crowd-sourcing and the elastic labor economy
- David Alan Grier, Professor of Science & Technology, George Washington University, Author - Human computers and the factorization of work
- Panos Ipeirotis, Professor at Stern School of Business NYU, Chief Scientist, Tagasauris - Scalable quality systems for distributed work.

Schutz will also talk about this start-up grant project at this year's American Association of Museums annual meeting, to be held May 19-22, 2013, as part a NEH-sponsored panel on "Tech Transforming Museums and Historic Sites: The Digital Humanities Perspective."

In addition to these multidisciplinary events directed at the intersection of the humanities and technology communities, Tagasauris has actively promoted the innovations associated with the project at technology industry-specific conference and events. These events drew audiences interested in crowd-sourcing, machine learning, and semantics, as well as the connection between these areas and educational and financial topics:

- Presentation at Pratt University with Lacy Schutz and Todd Carter, October 14, 2011 (75 participants)
- Lecture by Todd Carter at Columbia University's Digital Library Series, November 11, 2011 (100 participants)

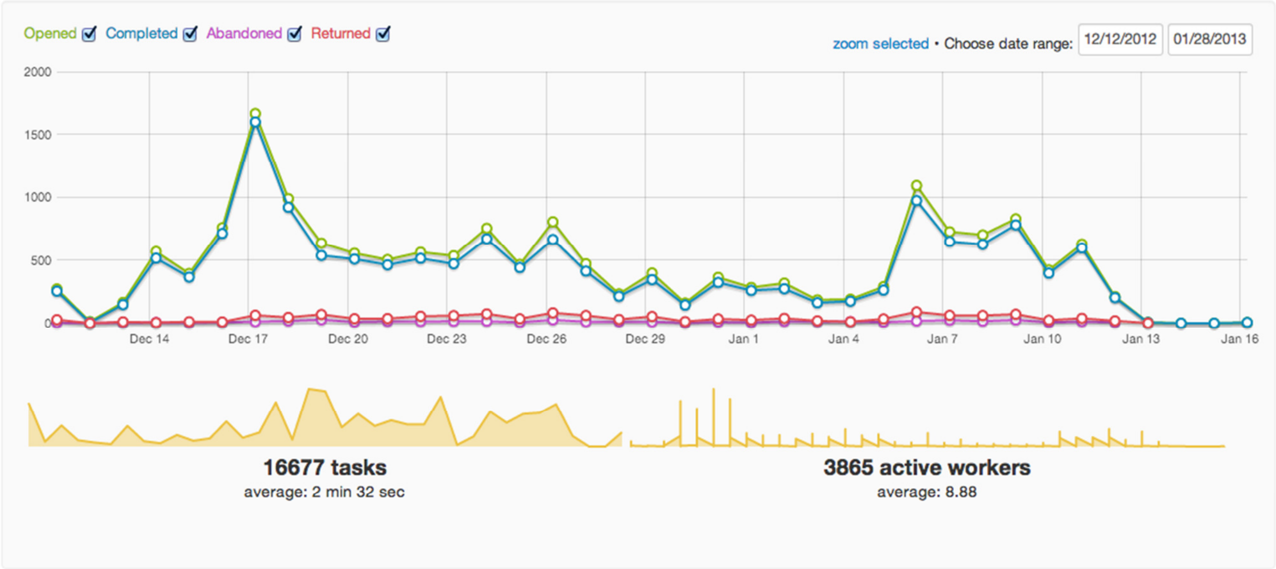
- Presentation by Todd Carter at the Henry Steward Digital Asset Management Conference in Los Angeles, California, November 14, 2011 (150 participants)
- School of Visual Arts presentation by Todd Carter, March 15, 2012 (75 participants)
- Keynote address by Todd Carter for the Visual Resources Association in Albuquerque, New Mexico, April 8, 2012 (450 participants)
- Dean's Lecture by Todd Carter at University of California, Berkeley, April 25, 2012 (75 participants)
- Presentation by Todd Carter for the Society of California Archivists in Ventura, California, April 27, 2012 (150 participants)
- New York Semantic Web Meet-up, October 18, 2012 (45 participants)
- Presentation for the Picture Archive Council of America by Todd Carter, October 21, 2012 (150 participants)
- Presentation by Todd Carter at the Henry Steward Digital Asset Management Conference in Los Angeles, California, November 1, 2012 (450 participants)

Evaluation

Tagasauris syndicated micro-tasks to Mechanical Turk, based on the defined workflow, and collected the resulting metadata from August of 2012 until January of 2013. During this time, 8,639 crowd-sourced workers completed 172,193 tasks, which resulted in a total of 472,628 semantic tags for 51,671 images. This equates to an average of 9.15 tags per image, demonstrating that the Museum could enhance the scalability of its digital asset annotation efforts through this method of crowd-sourced tagging.

The quality of output generated by the crowd-sourced workers was evaluated on a continuous basis through a task infrastructure that Tagasauris constructed. It evaluated worker quality at three levels, generating: 1) a score for a worker for each tag he or she creates, 2) an aggregated score for a worker for one specific micro-task, and 3) a worker's score across all tasks. Tagasauris built a dashboard to display these scores to workers as they completed their tasks, allowing them to compare their performance to that of their cohorts. The goal was to provide workers with instrumentation that would encourage more thoughtful work and a higher-quality product.

Image 2 below shows a screenshot of the Tagasauris interface of a dashboard of worker activity that was continuously monitored. While *Image 2* shows worker engagement, the dashboard also displays worker information, including historical metrics on the work completed and metrics on the quality of that work, an activity feed of work that is being submitted. The dashboard further includes functionality that allows workers to interact with their colleagues in real time, such as by communicating with specific workers, or by rewarding good work and rejecting bad work via the quality-control micro-task, in which they may vote on the appropriateness of fellow workers' tags.



Opened ☑ Completed ☑ Abandoned ☑ Returned ☑

zoom selected • Choose date range: 12/12/2012 01/28/2013

Image 2: Screenshot of Museum's dashboard from the Tagasauris interface showing worker engagement for the period 12/12/2012 - 01/28/2013

Our first evaluation procedure involved sampling 10 images for further analysis. All these images had existing metadata defined by the Museum's catalogers. We compared the existing metadata to the metadata generated in this project. As shown in *Image 3*, the catalogers defined an average of 6.1 tags while crowd-sourced workers defined an average of 15.3 tags. Additionally, machine processes and reasoning identified an average of 2 additional tags. We conclude that crowd-sourced workers provide a higher quantity of tags than the catalogers.

Next, we used comparative Panofsky-Shatform matrices to evaluate the quality of the tags we collected against the gold standard of the Museum's catalogers. In the mid-twentieth century, art historian Erwin Panofsky published his theories about the iconography of art[1] by suggesting that there are three levels at which an artwork can be described: the pre-iconographical (generic things in the image), iconographical (specific things), and iconological (symbolic things). The table below shows an example of this approach. In the 1980s, information scientist Sara Shatford applied Panofsky's model to indexing images[2]. She relabeled Panofsky's terms as Generic (pre-iconographic), Specific (iconographic) and Abstract (iconological), and extended the model further by breaking each of these three levels into four facets: Who, What, Where and When.

[1] Panofsky, E. (1962). *Studies in iconology*. Harper & Row, New York.

[2] Shatford, S. (1986). Analyzing the subject of a picture: a theoretical approach. Cataloging & Classification Quarterly, 6(3):39-62.

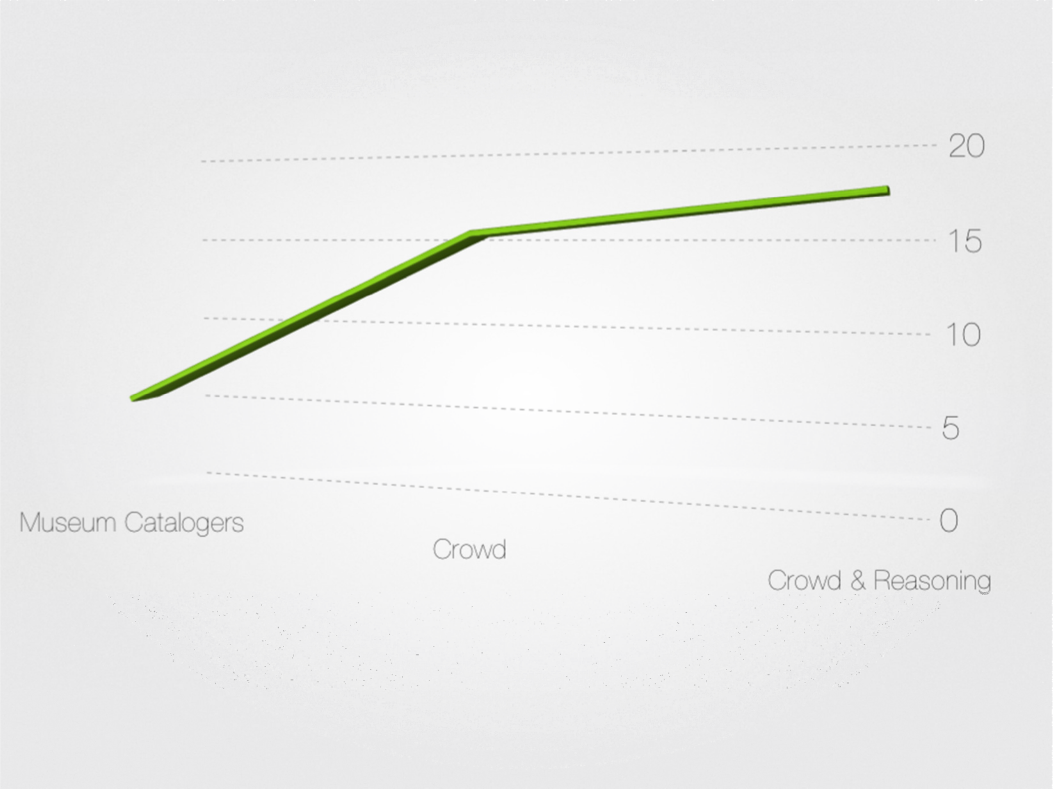
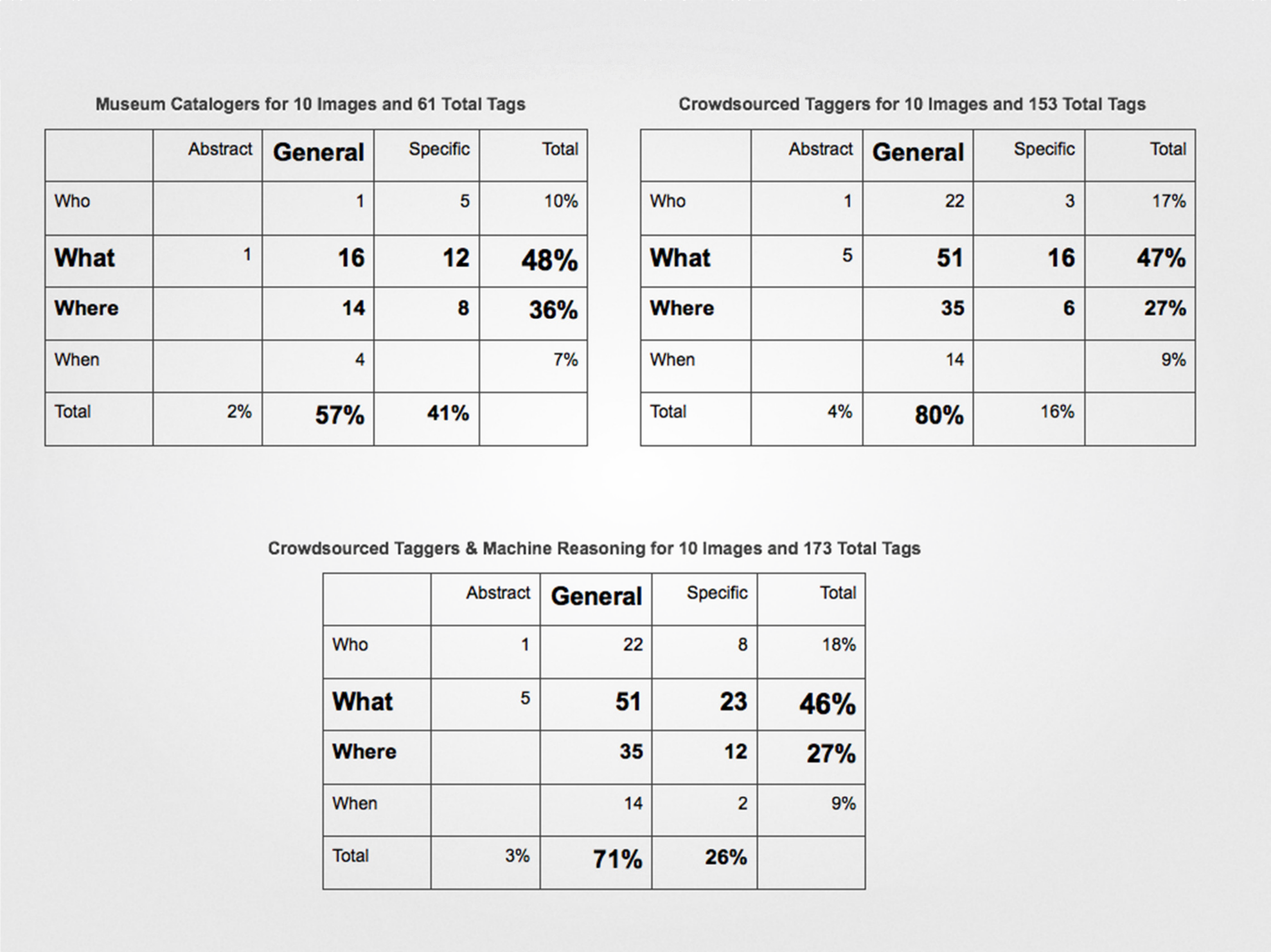


Image 3: Average number of tags per image by Museum catalogers, crowd-sourced workers, and machine reasoning

Refer to *Image 4* for the results of the Panofsky-Shatform analysis on the sample of 10 images. The catalogers identified a total of 61 tags for the 10 images in the sample. The tags break out into the following categories: 0.00% in Abstract-Who, 1.64% in Abstract-What, 0.00% in Abstract-Where, 0.00% in Abstract-When, 1.64% in Generic-Who, 26.23% in Generic-What, 22.95% in Generic-Where, 6.56% in Generic-When, 8.19% in Specific-Who, 19.67% in Specific-What, 13.11% in Specific-Where, and 0.00% in Specific-When.

Crowd-sourced workers and machine processes identified a total of 173 tags for the same 10 images. Note that for this sample set, crowd-sourced workers and machine processes identified 183.60% more relevant tags than the Museum's catalogers. The tags break out into the following categories: 0.58% in Abstract-Who, 2.89% in Abstract-What, 0.00% in Abstract-Where, 0.00% in Abstract-When, 12.72% in Generic-Who, 29.48% in Generic-What, 20.23% in Generic-Where, 8.09% in Generic-When, 4.62% in Specific-Who, 13.29% in Specific-What, 6.94% in Specific-Where, and 1.16% in Specific-When.



Museum Catalogers for 10 Images and 61 Total Tags

	Abstract	General	Specific	Total
Who		1	5	10%
What	1	**16**	**12**	**48%**
Where		**14**	**8**	**36%**
When		4		7%
Total	2%	**57%**	**41%**	

Crowdsourced Taggers for 10 Images and 153 Total Tags

	Abstract	General	Specific	Total
Who	1	22	3	17%
What	5	**51**	**16**	**47%**
Where		**35**	**6**	**27%**
When		14		9%
Total	4%	**80%**	16%	

Crowdsourced Taggers & Machine Reasoning for 10 Images and 173 Total Tags

	Abstract	General	Specific	Total
Who	1	22	8	18%
What	5	**51**	**23**	**46%**
Where		**35**	**12**	**27%**
When		14	2	9%
Total	3%	**71%**	**26%**	

Image 4: Comparative Panofsky-Shatford matrices

Refer to *Image 5* for a visualization of the comparison of the types of tags identified by Museum's catalogers verses crowd-sourced workers and machines. Our first observation is that crowd-sourced workers' tags are distributed comparably to those of the Museum's catalogers. In both the Abstract and Generic categories, the crowd-sourced workers' tags slightly outperform those of the Museum's catalogers. Yet, the Museum's catalogers outperform the crowd-sourced workers in the Specific categories. This can be explained by the fact that crowd-sourced workers lack the contextual knowledge or specific reference materials (for example, access to photographer and publisher notation on the verso side of the print) available to the Museum's catalogers. As such, they are often more limited in their ability to provide tags of a highly specific nature.

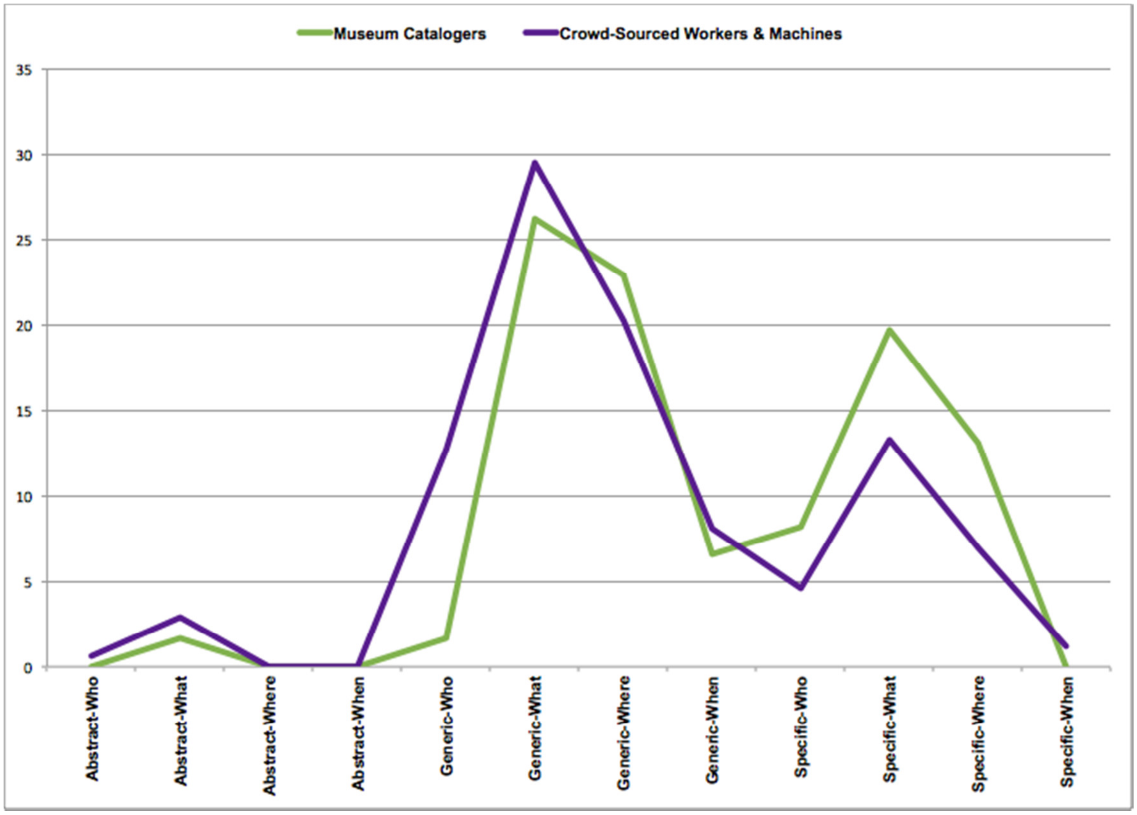
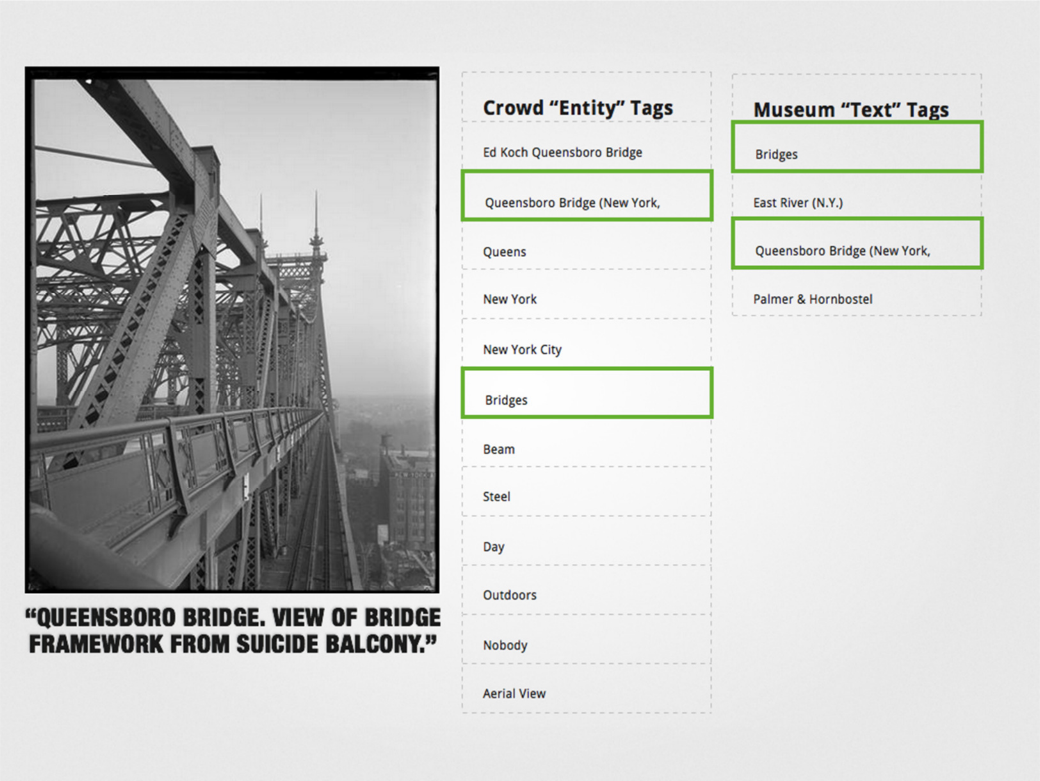


Image 5: Comparison of Panofsky-Shatford matrices for Museum catalogers and crowd-sourced workers and machines

In addition to the above analysis on the 10 sample images, we also present the details of the tags identified by the Museum's catalogers and the crowd-sourced workers for three of the sample images in *Image 6*, *Image 7*, and *Image 8*.

Often, with traditional cataloging methods, due to limited resources, catalogers have to be efficient and judicious about what tags to include. This results in a small but descriptive set of tags for each image. On the other hand, with crowd-sourced workers, there is no such need to limit the number of tags. Therefore, the crowd-sourced tags are much larger sets that consist of not only content-specific tags but also consistent features such as people count, day/night, indoors/outdoors, and photographic styles, etc.

Notice in *Image 6* that there are two tags that both catalogers and crowd-sourced workers identified for this image: "Queensboro Bridge" and "Bridges." In *Image 7* there are no tags in common between catalogers and crowd-sourced workers. In *Image 8* there are five tags in common between catalogers and crowd-sourced workers: "Equitable Life Assurance Co.," "Office," "Computers," "Interiors," and "Men."



“QUEENSBORO BRIDGE. VIEW OF BRIDGE FRAMEWORK FROM SUICIDE BALCONY.”

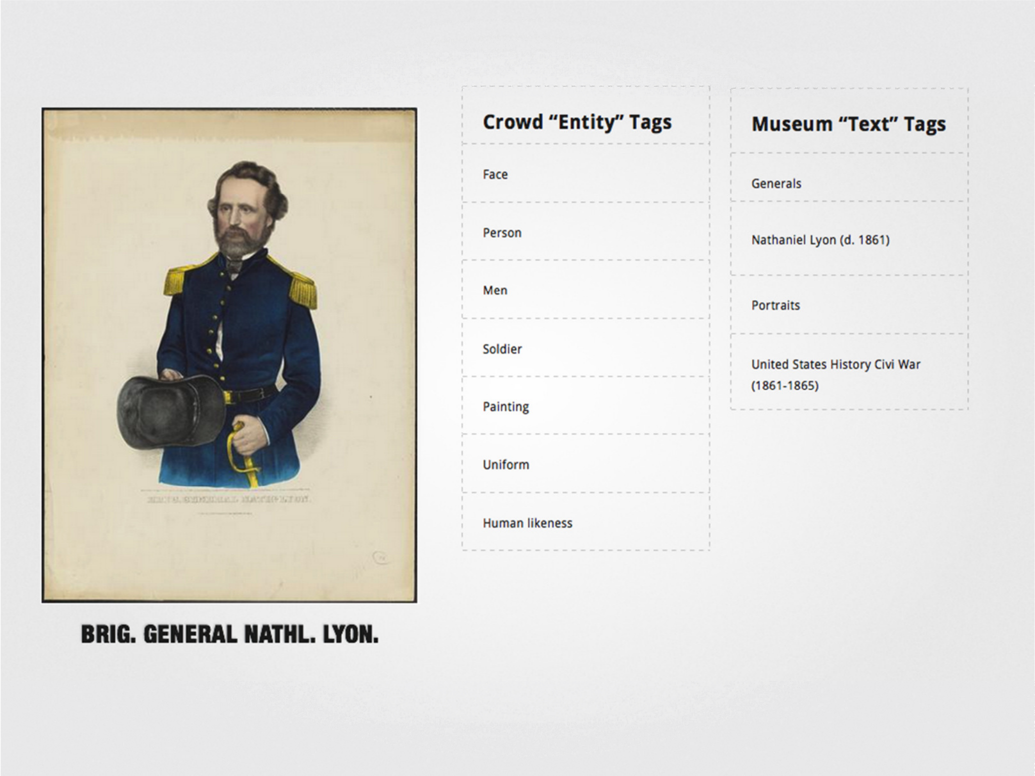
Crowd “Entity” Tags

- Ed Koch Queensboro Bridge
- Queensboro Bridge (New York,
- Queens
- New York
- New York City
- Bridges
- Beam
- Steel
- Day
- Outdoors
- Nobody
- Aerial View

Museum “Text” Tags

- Bridges
- East River (N.Y.)
- Queensboro Bridge (New York,
- Palmer & Hornbostel

Image 6: Comparative results from crowd-sourced workers and Museum catalogers



BRIG. GENERAL NATHL. LYON.

Crowd “Entity” Tags

- Face
- Person
- Men
- Soldier
- Painting
- Uniform
- Human likeness

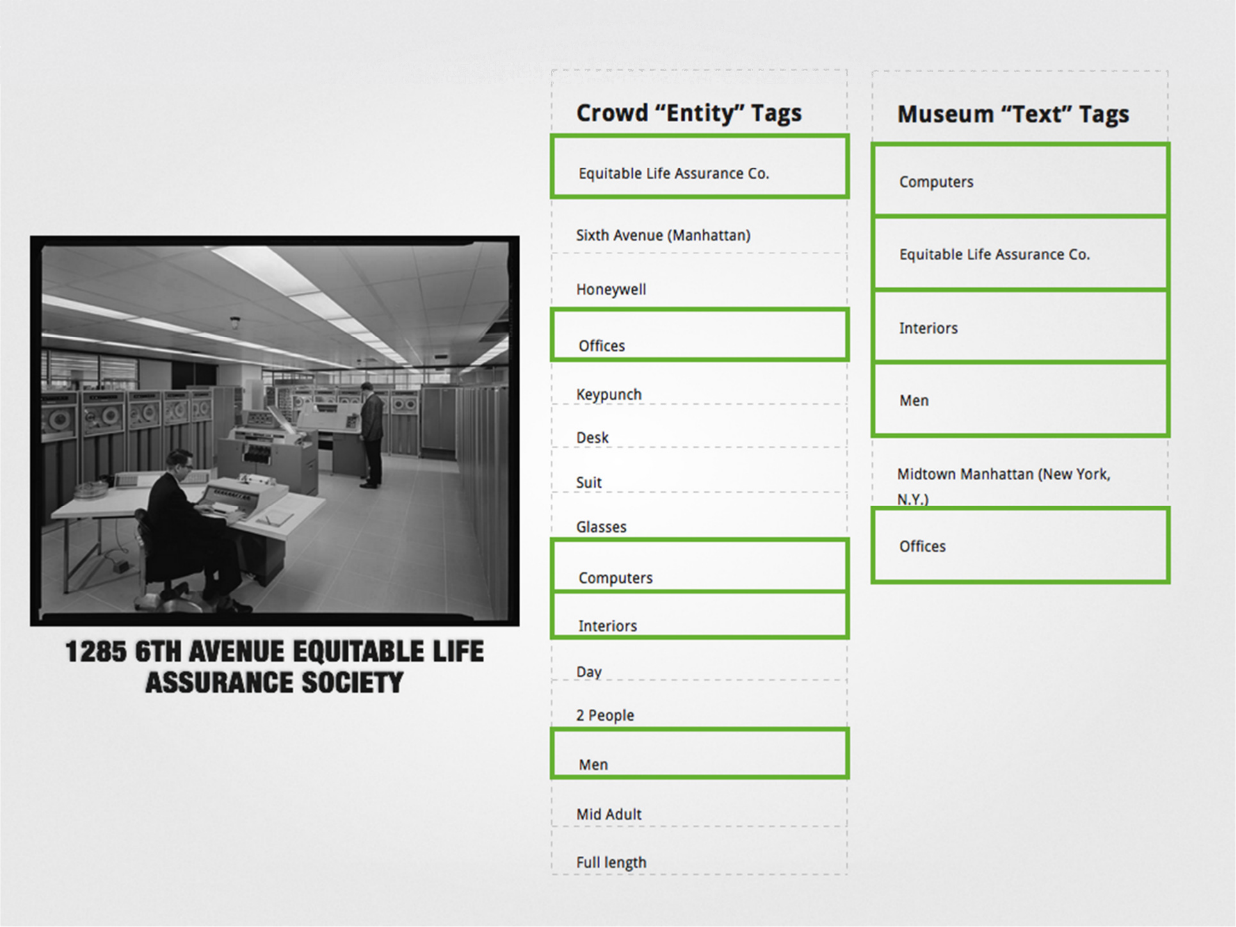
Museum “Text” Tags

- Generals
- Nathaniel Lyon (d. 1861)
- Portraits
- United States History Civi War (1861-1865)

Image 7: Comparative results from crowd-sourced workers and Museum catalogers

It is important to note that there is a clear and significant difference between the tags entered by the catalogers and crowd-sourced workers. The cataloger tags are static, text-based tags that are drawn from the limited lexicon of the Museum. Conversely, the crowd-sourced worker tags are semantic tags drawn from the linked open data sources with the reconciled terms from the Museum's lexicon, as completed in Stage 1 of the project. Refer to *Image 9* for an illustration of how each of these tags packs an entire networks of related information.

By using semantic linked open data sources for the terms that crowd-sourced workers draw from for annotation, we improved the digital record annotation capabilities. We improve on both quantity and quality. Improvements to quantity are clear and are discussed above, while improvements to quality are made as a result of the linked nature of the data which automatically provides access to related facts about each term. Refer to *Image 10* which illustrates all the related facts that are automatically available about "Queensboro Bridge."



1285 6TH AVENUE EQUITABLE LIFE ASSURANCE SOCIETY

Crowd "Entity" Tags

- Equitable Life Assurance Co.
- Sixth Avenue (Manhattan)
- Honeywell
- Offices
- Keypunch
- Desk
- Suit
- Glasses
- Computers
- Interiors
- Day
- 2 People
- Men
- Mid Adult
- Full length

Museum "Text" Tags

- Computers
- Equitable Life Assurance Co.
- Interiors
- Men
- Midtown Manhattan (New York, N.Y.)
- Offices

Image 8: Comparative results from crowd-sourced workers and Museum catalogers

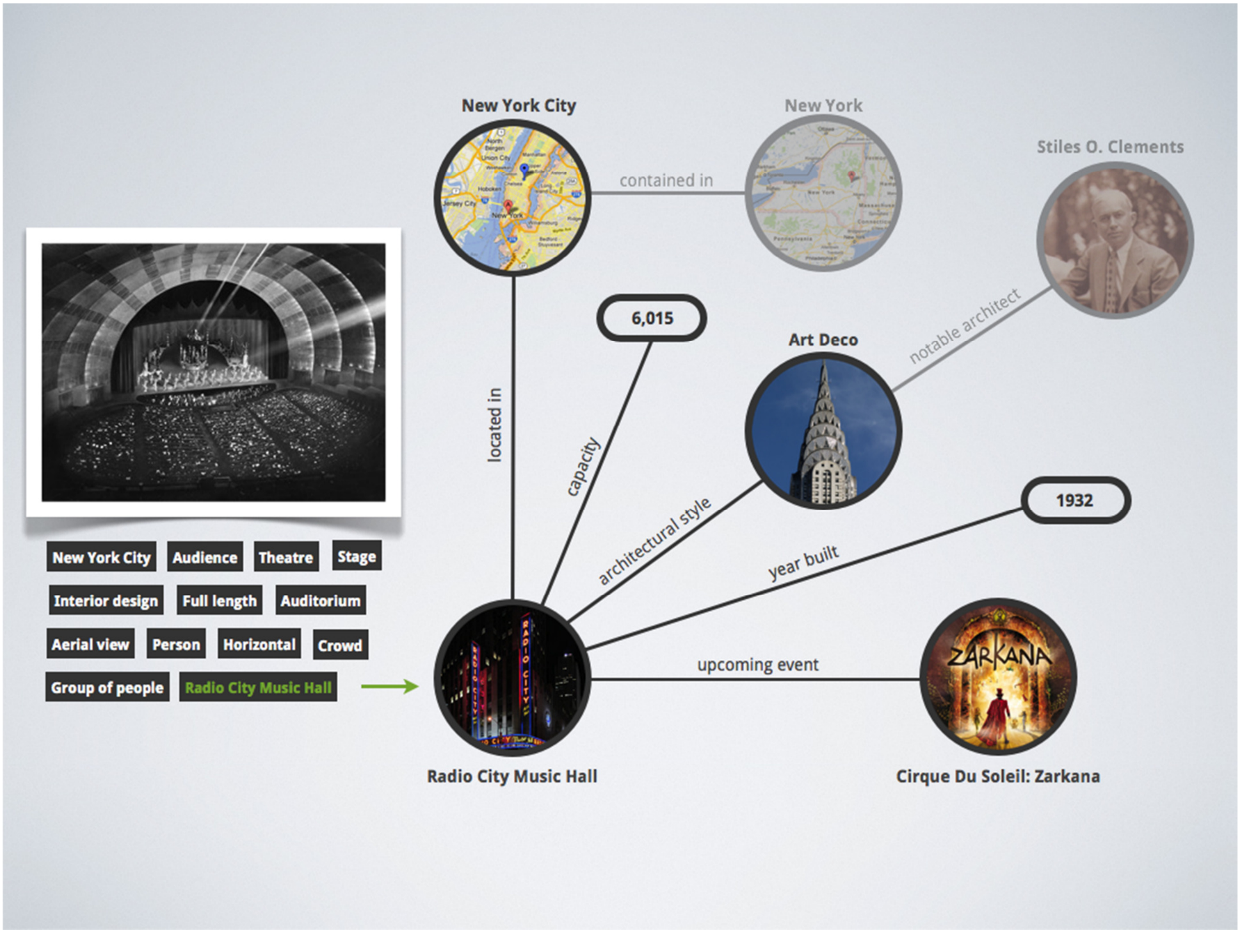
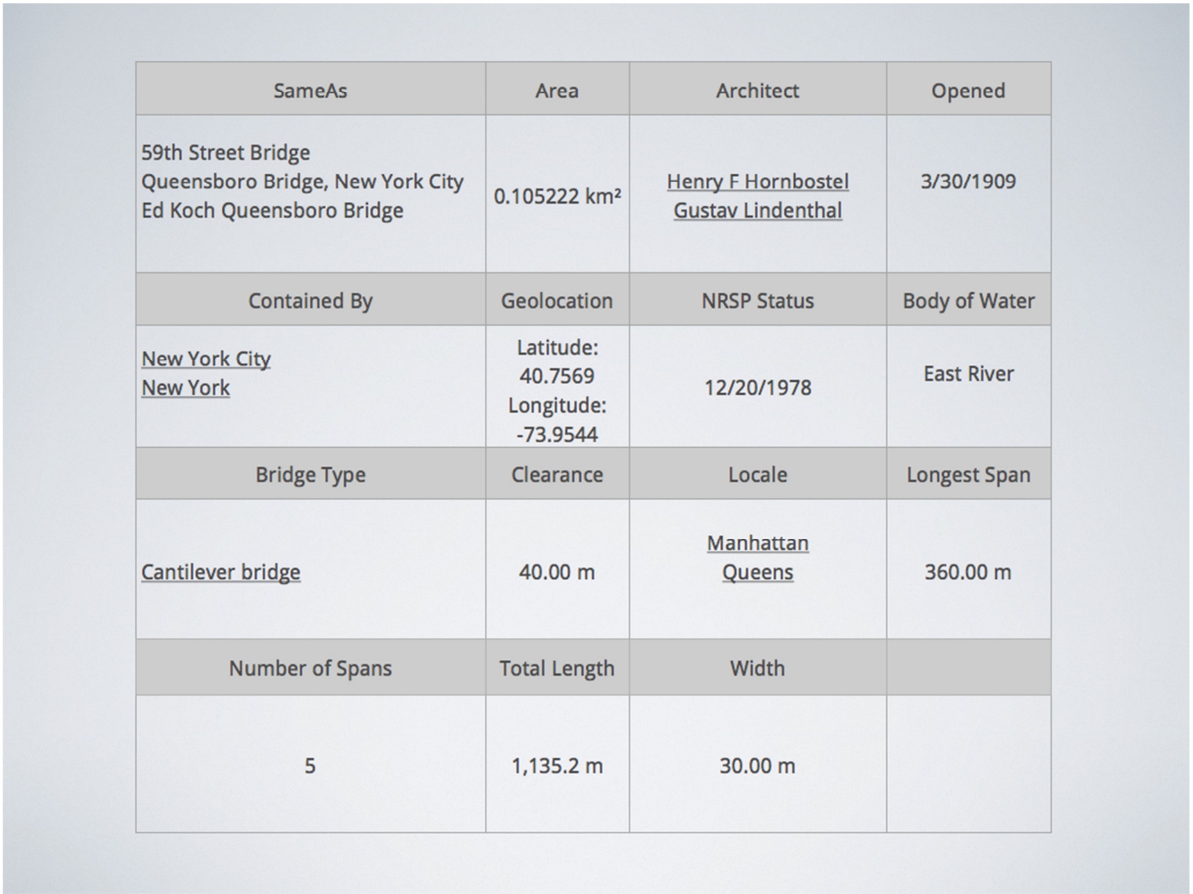


Image 9: All objects related to a particular concept are connected by a single Entity Tag.

Using a base of semantic linked open data as the source of terms for annotation solves two problems. The first problem is uniformity. Individual entities or concepts like New York City are often represented differently by different institutions with multiple keywords (like “nyc,” “new_york_city,” and “new york”), making it difficult to organize related content. This ultimately results in the problem of segmenting content. In such a system, a search for “New York City” would return a different set of results than “NYC.”

The second problem is one of ambiguity. It is not always clear what a particular keyword represents. For example, does “Wall Street” represent an eight-block-long street running from Broadway to South Street on the East River in Lower Manhattan, the financial district of New York City, or a 1987 American drama film directed by Oliver Stone?



SameAs	Area	Architect	Opened
59th Street Bridge Queensboro Bridge, New York City Ed Koch Queensboro Bridge	0.105222 km²	Henry F Hornbostel Gustav Lindenthal	3/30/1909
Contained By	Geolocation	NRSP Status	Body of Water
New York City New York	Latitude: 40.7569 Longitude: -73.9544	12/20/1978	East River
Bridge Type	Clearance	Locale	Longest Span
Cantilever bridge	40.00 m	Manhattan Queens	360.00 m
Number of Spans	Total Length	Width	
5	1,135.2 m	30.00 m	

Image 10: Illustration of Freebase concepts linked to the Entity Tag "Queensboro Bridge"

Controlled vocabularies like those used at the Museum help solve some of these problems by limiting the range of available keywords that can be used to describe an object to the shortlist of approved authority terms included in the vocabulary. However, other problems persist for potential visitors to the Collections Portal who are unfamiliar with the approved subject terms used in the vocabulary. They will also affect any future federated database of New York City archival materials (referenced below under Long Term Impact) because different vocabularies have been used to described similar objects in different collections. Here, controlled vocabularies often lead to unsatisfactory recall, in that they fail to retrieve some documents that are actually relevant to the search in question.

Continuation of the Project

Although the project is now complete, further work is needed to realize the full potential of the work performed and the learning acquired under this NEH Digital Humanities Start-up grant. In order to further maximize the accessibility, discoverability and data interoperability of the Museum's collections to both human and machine users, the Museum and Tagasauris propose a second three-stage work plan as defined below:

Stage 2.1: Reconcile all existing Museums "Keywords" to Freebase Entity Tags

Currently, the Tagasauris annotation platform solution has resulted in two sets of tags for each of the 51,671 objects in the original test sample set: the existing text-based tags that the Museum's catalogers entered (referred to as "existing tags"), and the Freebase Entity Tags that were added during this project by crowd-sourced workers and machines (referred to as "new tags"). As discussed in the Evaluation section of this report, the new tags reap the benefits of semantic linked open data, while the existing tags continue to suffer the problems of static text tags.

We propose to solve this problem by reconciling all the Museum-generated existing tags to Freebase Entity Tags. We would begin by repurposing the reconciliation workflow developed under Stage 1 of the project to the task of reconciling all existing tags for Freebase Entity Tags. Powered by an optimal distribution of crowds and machines, a plurality of judgments would be collected to help choose the Freebase Entity Tag that best reconciles to an existing tag. Then a "merge or match" task would evaluate whether two different tag entities should be considered the same. Lastly, we would merge all pairs of Freebase Entity Tags for each duplication between the existing tags and new tags.

There are compelling and powerful benefits to tagging the Museum's collections with only Freebase Entity Tags. These benefits result from the fact that Entity Tags, unlike text keywords, are references to unique, well-defined concepts complete with metadata and their own URLs. This would also resolve two critical problems: synonyms and homonyms.

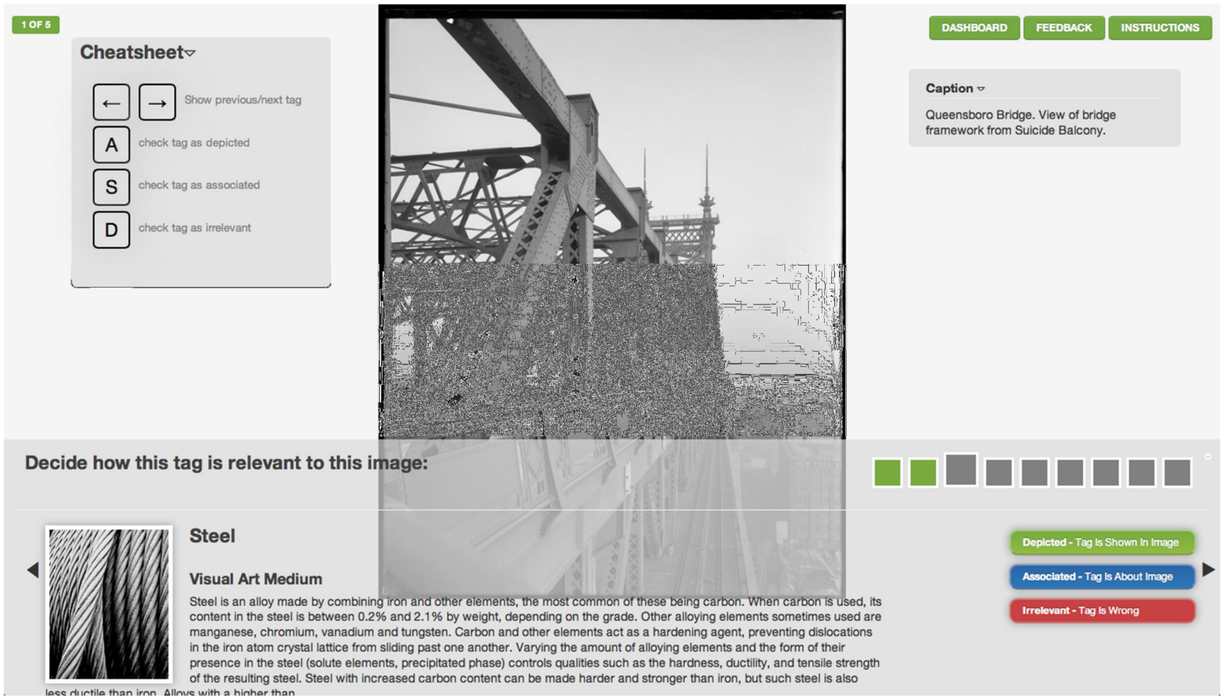
The first problem observed with the existing tags is that synonyms were often used for tagging, which yields incomplete results. Searching the Collections Portal for "car" returns 2,667 results, while a search for "automobile" returns 2,212 different results. Similarly, "streetcar" returns 56 results, while the alternative term "street railroads" returns 2,550 results and "tram" returns just 1 result.

The second problem observed with the existing tags is that of homonyms. Even in the best-case scenario, controlled language is often not as specific as the search terms input by the users themselves. For example, searching the Collections Portal using the vocabulary term "Central Park (New York, N.Y.)" produces 2,455 results. But an evaluation of failed searches in the Collections Portal search logs reveals the search term "Greensward Park." In contrast, both Wikipedia and Freebase recognize Greensward Park as a synonym of Central Park and redirect it accordingly. Wikipedia currently maintains nine English-language redirects (terms that are mostly synonyms) that redirect users to Central Park.

Additionally, by adopting Entity Tags, the Museum can increase the discoverability of objects in its collection. Search engines like Google and Bing have begun reading RDFa--the markup standard used to represent Entity Tags--to acquire richer information about the content labeled with it. These search engines are using this information to improve the presentation and relevance of collection search results in organic Web search, helping drive more traffic to the Museum's Collections Portal.

Stage 2.2: Continuous Improvement

Once all tags have been standardized to Freebase Entity Tags, the next step involves collecting data to measure the relevance of the tags associated with each image. A series of specialized micro-tasks would be syndicated in a repetitive and iterative fashion to crowd-sourced workers. An initial task would collect metrics from a plurality of workers on the degree of relevance of each tag, classifying them as: relevant and depicted (all elements that are observed in the image), relevant and associated (all elements that are about the image but not seen, such as the name of the photographer who took the picture), or not relevant. Refer to *Image 11* for a snapshot of such a task.


1 OF 5
Cheatsheet▿
← → Show previous/next tag
A check tag as depicted
S check tag as associated
D check tag as irrelevant
DASHBOARD
FEEDBACK
INSTRUCTIONS
Caption ▿
Depicted - Tag Is Shown In Image
Associated - Tag Is About Image
Irrelevant - Tag Is Wrong

Queensboro Bridge. View of bridge framework from Suicide Balcony.

Decide how this tag is relevant to this image:

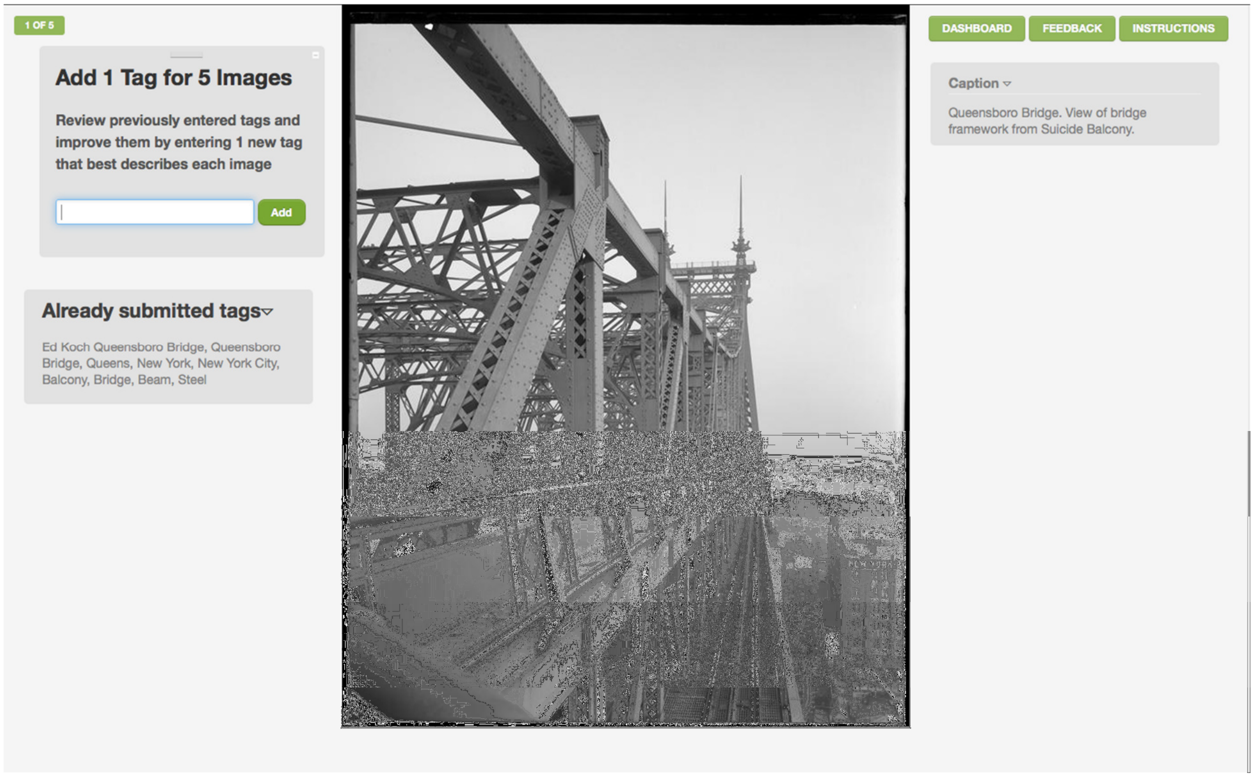
Steel

Visual Art Medium

Steel is an alloy made by combining iron and other elements, the most common of these being carbon. When carbon is used, its content in the steel is between 0.2% and 2.1% by weight, depending on the grade. Other alloying elements sometimes used are manganese, chromium, vanadium and tungsten. Carbon and other elements act as a hardening agent, preventing dislocations in the iron atom crystal lattice from sliding past one another. Varying the amount of alloying elements and the form of their presence in the steel (solute elements, precipitated phase) controls qualities such as the hardness, ductility, and tensile strength of the resulting steel. Steel with increased carbon content can be made harder and stronger than iron, but such steel is also less ductile than iron. Alloys with a higher than

Image 11: Snapshot of a Depicted and Associated Tag Relevance task running against images and tags collected during the third stage of the project.

These metrics, along with Panofsky-Shatford matrices, would provide a framework for collections staff at the Museum to visualize cataloging deficiencies, targeting collections for new tasks to further improve metadata expressiveness. Once deficiencies are identified, further micro-tasks can be syndicated, again in a repetitive and iterative fashion, that require workers to improve on the work of the previous workers. Refer to *Image 12* for a snapshot of just one example of such a task. Additionally, the metrics collected for tags that are marked "not relevant" could be used by collections staff at the Museum to find and eliminate erroneous tags.



Add 1 Tag for 5 Images

Review previously entered tags and improve them by entering 1 new tag that best describes each image

Already submitted tags

Ed Koch Queensboro Bridge, Queensboro Bridge, Queens, New York, New York City, Balcony, Bridge, Beam, Steel

Caption

Queensboro Bridge. View of bridge framework from Suicide Balcony.

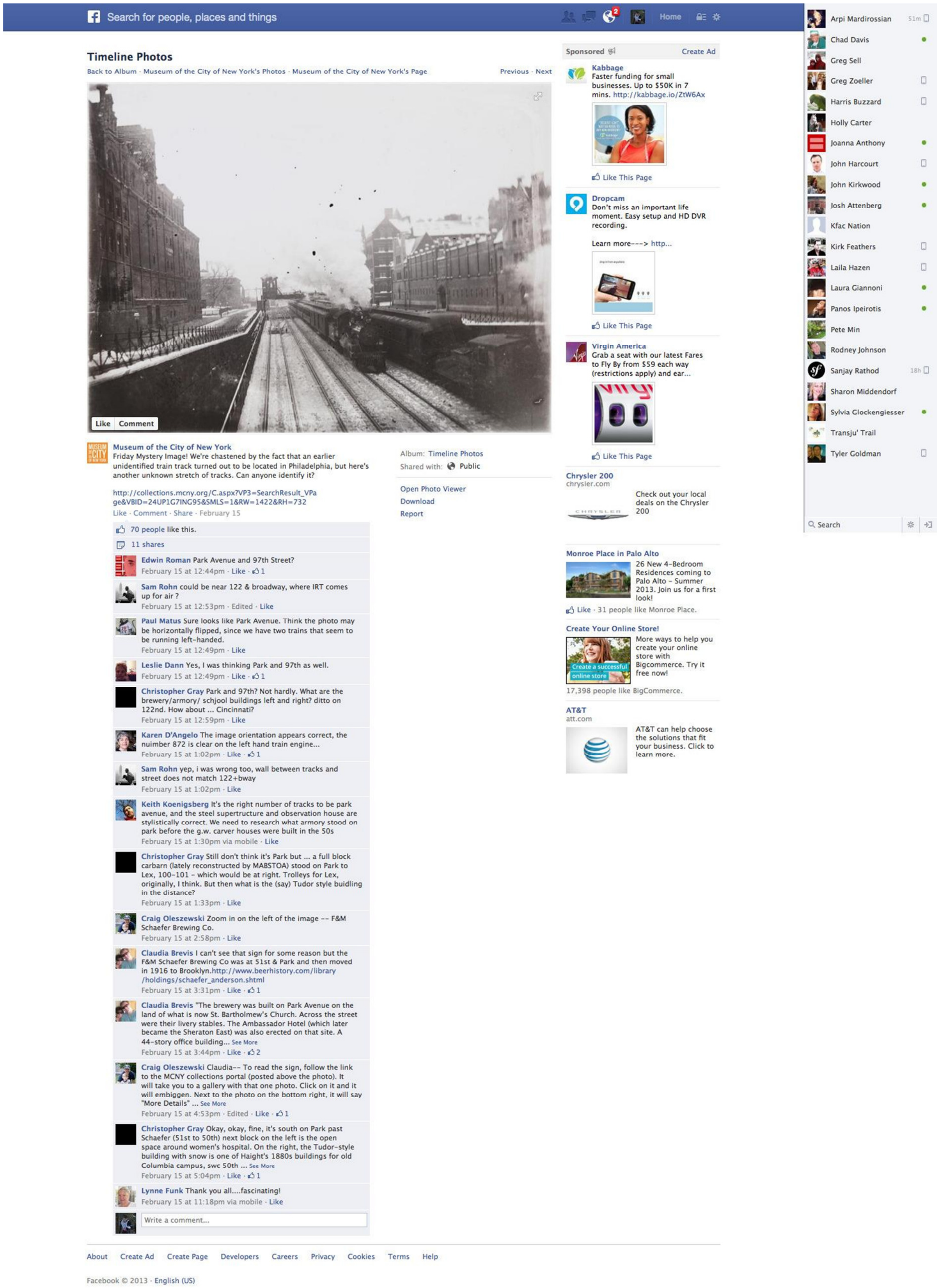
Image 12: Screenshot of a Tag Improvement task running against already submitted tags

Stage 2.3 Use Tagasauris Social Media Suite to tap into "Crowd Knowing"

Beyond the scope of the original Digital Humanities Start-up grant proposal, Tagasauris prototyped a social media suite, i.e. an application to engage the Museum's social network followers in the tagging process and capture their input across all interconnected micro-tasks and associated actions. In the future, the Museum could syndicate tasks to its social network followers instead of or in addition to crowd-sourced workers by posting a photo on Twitter or Facebook with a URL directing followers to each of the required micro-tasks. Upon logging into a stand-alone application using their Twitter or Facebook ID, followers would be able to annotate images directly. Social network laborers would work for free, but would receive points for their work, with the top scorers displayed on a "leaderboard" on the application interface and, optionally, within their own social media feeds.

We have collected empirical evidence that social media followers who are engaged with the Museum, and presumably are more informed about its collections, are capable of producing specific and high-quality metadata. A platform solution that could tap into crowd knowing and intrinsic motivation represents a significant opportunity for humanities and cultural organization to scale these sorts of annotation projects.

As an early proof of concept, in 2011, the Museum launched its "Friday Mystery Image" project. Each week, the Museum posts a photo from its collection on Facebook that is insufficient in metadata and supporting information. Refer to *Image 13* for a snapshot of a recent "Friday Mystery Image" post.



Search for people, places and things

Home

Timeline Photos

Back to Album · Museum of the City of New York's Photos · Museum of the City of New York's Page

Previous · Next

Like Comment

Museum of the City of New York Friday Mystery Image! We're chastened by the fact that an earlier unidentified train track turned out to be located in Philadelphia, but here's another unknown stretch of tracks. Can anyone identify it?

http://collections.mcny.org/C.aspx?VP3=SearchResult_VPage&VBID=24UP1G7ING95&SMLS=1&RW=1422&RH=732

Like · Comment · Share · February 15

70 people like this.

11 shares

Edwin Roman Park Avenue and 97th Street?
February 15 at 12:44pm · Like · 1

Sam Rohn could be near 122 & broadway, where IRT comes up for air ?
February 15 at 12:53pm · Edited · Like

Paul Matus Sure looks like Park Avenue. Think the photo may be horizontally flipped, since we have two trains that seem to be running left-handed.
February 15 at 12:49pm · Like

Leslie Dann Yes, I was thinking Park and 97th as well.
February 15 at 12:49pm · Like · 1

Christopher Gray Park and 97th? Not hardly. What are the brewery/armory/ schjool buildings left and right? ditto on 122nd. How about ... Cincinnati?
February 15 at 12:59pm · Like

Karen D'Angelo The image orientation appears correct, the nuimber 872 is clear on the left hand train engine...
February 15 at 1:02pm · Like · 1

Sam Rohn yep, i was wrong too, wall between tracks and street does not match 122+bway
February 15 at 1:02pm · Like

Keith Koenigsberg It's the right number of tracks to be park avenue, and the steel supertructure and observation house are stylistically correct. We need to research what armory stood on park before the g.w. carver houses were built in the 50s
February 15 at 1:30pm via mobile · Like

Christopher Gray Still don't think it's Park but ... a full block carbarn (lately reconstructed by MABSTOA) stood on Park to Lex, 100-101 - which would be at right. Trolleys for Lex, originally, I think. But then what is the (say) Tudor style buidling in the distance?
February 15 at 1:33pm · Like

Craig Oleszewski Zoom in on the left of the image -- F&M Schaefer Brewing Co.
February 15 at 2:58pm · Like

Claudia Brevis I can't see that sign for some reason but the F&M Schaefer Brewing Co was at 51st & Park and then moved in 1916 to Brooklyn.http://www.beerhistory.com/library/holdings/schaefer_anderson.shtml
February 15 at 3:31pm · Like · 1

Claudia Brevis "The brewery was built on Park Avenue on the land of what is now St. Bartholmew's Church. Across the street were their livery stables. The Ambassador Hotel (which later became the Sheraton East) was also erected on that site. A 44-story office building... See More
February 15 at 3:44pm · Like · 2

Craig Oleszewski Claudia-- To read the sign, follow the link to the MCNY collections portal (posted above the photo). It will take you to a gallery with that one photo. Click on it and it will embiggen. Next to the photo on the bottom right, it will say "More Details" ... See More
February 15 at 4:53pm · Edited · Like · 1

Christopher Gray Okay, okay, fine, it's south on Park past Schaefer (51st to 50th) next block on the left is the open space around women's hospital. On the right, the Tudor-style building with snow is one of Haight's 1880s buildings for old Columbia campus, swc 50th ... See More
February 15 at 5:04pm · Like · 1

Lynne Funk Thank you all....fascinating!
February 15 at 11:18pm via mobile · Like

Write a comment...

Album: Timeline Photos
Shared with: Public

Open Photo Viewer
Download
Report

Arpi Mardirossian 51m
Chad Davis ●
Greg Sell
Greg Zoeller
Harris Buzzard
Holly Carter
Joanna Anthony ●
John Harcourt
John Kirkwood ●
Josh Attenberg ●
Kfac Nation
Kirk Feathers
Laila Hazen
Laura Giannoni ●
Panos Ipeirotis ●
Pete Min
Rodney Johnson
Sanjay Rathod 18h
Sharon Middendorf
Sylvia Glockengiesser ●
Transju' Trail
Tyler Goldman

Search

About Create Ad Create Page Developers Careers Privacy Cookies Terms Help

Facebook © 2013 · English (US)

Image 13: Screenshot of a "Friday Mystery Image" post on the Museum's Facebook page

The Museum invites its followers to identify the location where the photo was taken or any other relevant information about the subject it depicts. These posts often spur a dialogue among its social media followers and result in new information about the photo that the Museum would not be able to obtain any other way. The level of engagement and cooperation that this project illustrates points to promising outcomes with our proposed social tagging project.

The Museum has applied for an NEH Digital Humanities Implementation grant to continue its collaboration with Tagasauris, which proposes the continuation of the work plan outlined above--refining this approach, and implementing it in a way that feeds back into the Museum's collections database and digital asset management systems. The plan requirements include the use of the Tagasauris annotation platform; technical consulting services by Tagasauris; the use of the Museum's collections management platform, which utilizes a software called Cortex; and technical consulting services by Orange Logic, the digital asset management company that runs Cortex.

Long Term Impact

While the scope of this grant was purposely constrained to the area where the need was most pressing--improving digital record annotation capabilities with opened-sourced ontologies and crowdsourced workers--the long-term implications of the project go well beyond the enhanced ability to tag collections with descriptive metadata at scale.

In the short term, this system will allow organizations to make their collections more accessible with rich, descriptive metadata. In the longer term, one can envision a next-generation, knowledge-driven platform that delivers on the promise of a ubiquitous online cloud of seamlessly interconnected networked cultural heritage information.

Today, knowledge bases are powering a growing array of applications. Well-known examples of knowledge bases include the Internet Movie Database, DBpedia, and Freebase. For the Museum, a knowledge base would encapsulated all the concepts, instances and relationships depicted and associated with the Museums collections. The Museum would use this knowledge base to power a range of cultural heritage applications, including query understanding, deep search, and dynamic information-driven historical narratives.

Implementation of the tagging solution on a large scale, such as through the project proposed for a pending NEH implementation grant, would allow the Museum to integrate the crowd-sourced metadata and the LOD to digital entities in its online Collections Portal, surfacing the assets to a much wider base of users, bringing additional awareness to the collection and institution, and engaging users in tagging through social media sites such as Facebook and Twitter.

The project has had a profound effect on public perceptions of the Museum of the City of New York. The Museum's participation has brought its staff and collections to the attention of major players in the world of linked open data in libraries, archives, and museums. Because of her involvement in this work, Project Director Lacy Schutz is currently serving on the advisory board of a NEH Digital Humanities

Implementation project at Dartmouth College, Mary Flanagan's "Metadata Games: Improving Access to Humanities Artifacts." On May 2-3, 2013, she will be presenting on a panel at the Digital Asset Management New York conference entitled "Should you Set your Valuable Assets Free? Easier Said than Done; How to retag and enhance digital assets so they can roam the viral web and drive merchandising, marketing and other commercial opportunities." Later in May 2013, Schutz is participating in a symposium, "Digital Cultural Heritage and User Experience," that will focus on a collaborative IMLS-funded project between Pratt School of Information and Library Science, Brooklyn Museum, Brooklyn Historical Society, and Brooklyn Public Library. The Museum is also participating in a joint project between the METRO New York Library Council, New York Public Library, and New York University to create a confederated online resource of archival material related to New York City History.

The Museum has also contributed selections from its digitized photography collections to Artsy and the Google Art Project. Artsy is an online source for art, built on the Art Genome Project, an effort to map common characteristics (tags like medium, but also movement and concept) across individual art works and thus enabling new paths for discovery and learning. The Google Art Project is a far-reaching effort to provide access to the world's art collections. None of these opportunities would have been possible without the visibility this project has brought to the Museum.

NEH funding for this project has also laid the groundwork for new public-private partnerships exploring the intersection of the humanities and digital technology. As a result of the lessons learned through projects like this, Panos Ipeirotis, Tagasauris co-founder and New York University associate professor of information, operations and management sciences and George A. Kellner Faculty Fellow, along with researchers Serge Belongie of UC San Diego and Pietro Perona of CalTech, received a $1.5 million Focused Research Award grant from Google. With this grant, they will develop a program for integrating crowdsourcing with machine learning algorithms to improve *t*he ability to search and identify visual media.

Similarly, Tagasauris is funding mediaGraph, a web-scale semantic repository that seamlessly interconnects networked audio-visual content. Within mediaGraph, Tagasauris plans to place linked audio-visual media in an explicit historical context in order to provide a more complete and illustrated description of historical events. This knowledge-first approach to cultural heritage materials has the potential to enable access to objects in new and novel ways including personalized, information-driven narratives and dynamic historical sequences.

Grant Products

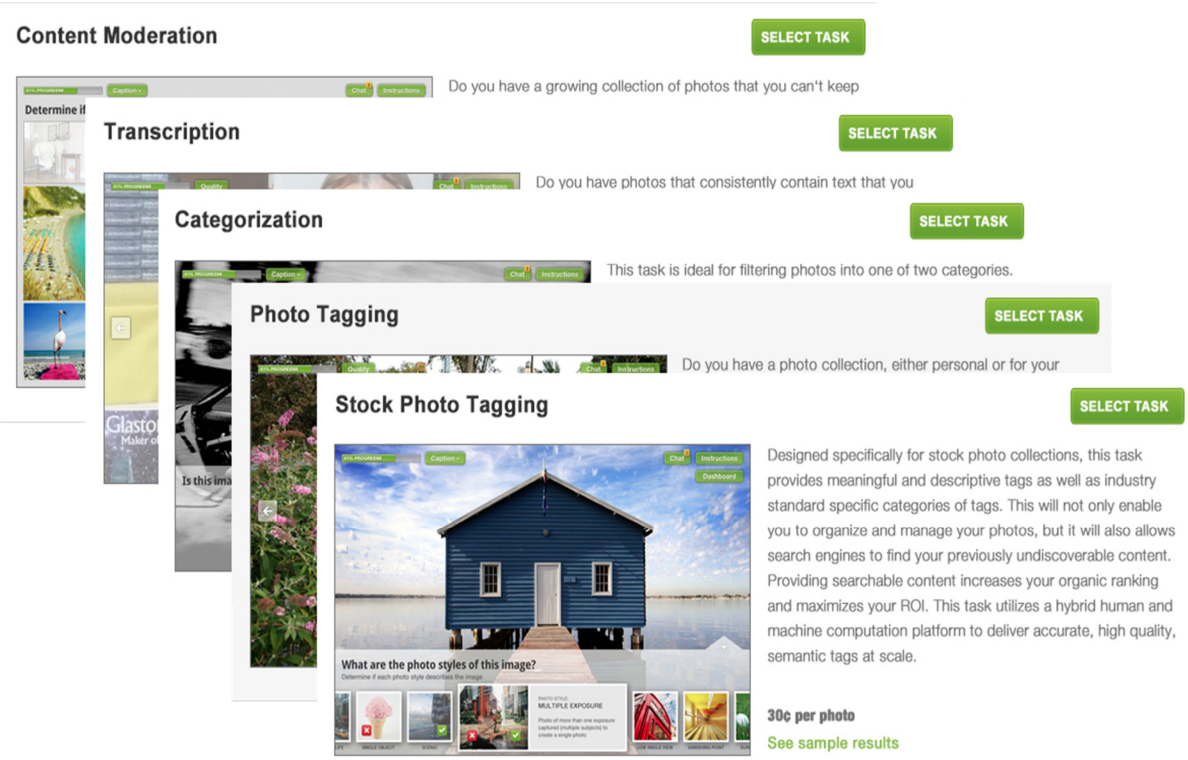
As described under Project Activities, the project resulted in the development of the following grant products:

- Subject authority as LOD on Freebase
- Micro-tasks
- Workflows
- Dashboard
- Image Detail Page

In order to facilitate access to all of these deliverables, Tagasauris has launched photo.tagasauris.com, an easy to use software-as-a-service website for annotating images with Freebase entity tags. Individuals and institutions can set up accounts at test a variety of standard workflows offered by the service at no cost. The service is available on a metered pay-per-use model and discounts of 20% are available to not-for-profit organizations.

The site gives collections managers for humanities and cultural institutions at all resources levels the ability to annotate media in their collections using the same open-sourced ontologies and crowd-sourcing workflow process developed during this project.

Workflow tasks implemented by Tagasauris for commercial clients are also accessible through this service. Workflow tasks from this project can be combined can be reconfigured and combined in new ways and published into the user accounts of humanities and cultural institutions on photo.tagasauris.com or accessed via a RESTful API as custom workflows providing annotations and other output, rather than relying on the workflows established by the Museum.



Content Moderation

Do you have a growing collection of photos that you can't keep

Transcription

Do you have photos that consistently contain text that you

Categorization

This task is ideal for filtering photos into one of two categories.

Photo Tagging

Do you have a photo collection, either personal or for your

Stock Photo Tagging

What are the photo styles of this image?

Designed specifically for stock photo collections, this task provides meaningful and descriptive tags as well as industry standard specific categories of tags. This will not only enable you to organize and manage your photos, but it will also allows search engines to find your previously undiscoverable content. Providing searchable content increases your organic ranking and maximizes your ROI. This task utilizes a hybrid human and machine computation platform to deliver accurate, high quality, semantic tags at scale.

30¢ per photo

See sample results

Image 14: Screenshot of canned tasks made available to via photo.tagasauris.com software as a service application.

The Tagasauris service also exposes RESTful APIs for programmatic access to the Tagasauris system from any third-party digital asset management, enterprise content management, media asset management, software application. mobile device, or information technology system.

These capabilities support organizations of all sizes and technical capabilities in accessing and orchestrating crowdsourced workers, and socially engaged followers in a way that is driven off the matrices above, strengthens those areas where the Museum needs the most help. Following a presentation by Todd Carter for the California Archives/University of California system on April 27, 2012, 347 people have created accounts to create workflows and begin testing.

Tagasauris is committed to actively maintaining and improving the service offerings available through this site, as well as expanding the functionality to support future uses.