Categorized | ChapterNews

Crowdsourcing and Linked and Open Data

Posted on 12 December 2011. Tags: 2011 #4 Winter, ChapterNews, Crowdsourcing, DBpedia, Magnum Photos, Mechanical Turk, metadataAuthority Control, Museum of the City of New York, National Endowment for the Humanities, Open Data, Pratt, Tagasauris

John Tomlinson | Twitter | mail@johntomlinson.com

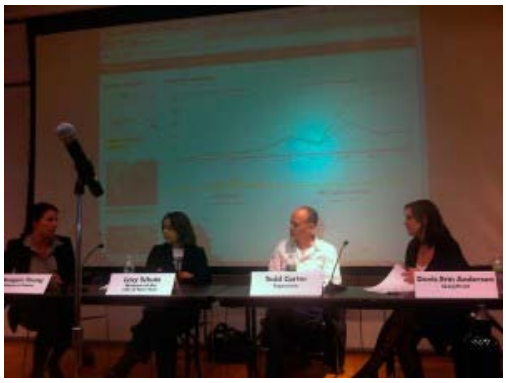
John Tomlinson is website manager of SLA@Pratt and lead organizer of this event. He is completing work toward an MLIS degree at Pratt's School of Information and Library Science, with interests particularly in knowledge management and information sharing for nonprofit organizations and socially oriented businesses.

Crowdsourcing and Linked and Open Data

What if you had 500,000 photographs, of which 200,000 lacked subject metadata, with more photos added each week and a team of just four cataloguers? And what if your organization's collection contained many original and iconic news images, yet Internet searches turned up copies from other sources? Magnum Photos faced these challenges.

What if you worked at a museum that was creating digital images of tens of thousands of objects and print photographs with only a few cataloguers to apply metadata? And producing each digital image took only a minute or two, while applying keywords from your organization's lexicon took 15 minutes? The Museum of the City of New York faced these challenges.

An emerging approach to meet such challenges, and its implications for information professionals, was the focus of a panel discussion called *Crowdsourcing & Linked & Open Data: New Ways to Make Collections Visible*. The event was organized by SLA@Pratt, the Pratt Institute School of Information and Library Science's student chapter of SLA, on October 14, 2011.



Meagan Young of Magnum Photos (left) speaking as Lacy Schutz, Todd Carter, and Davis Erin Anderson look on, with dashboard for Magnum's Tagasauris tasks shown live on the screen. Photo by Sharon Middendorf/Tagasauris

Increasing Sales

Following welcome and introductions from SLA@Pratt President Davis Erin Anderson and me, Meagan Young of Magnum Photos shared her approach to annotating a large digital image collection. Young, a Senior Web Product Manager with particular expertise in social media and data-driven design, started with some background on Magnum. The company is a cooperative of photographers founded over sixty years ago, with its members continuously adding to its collections. But its digital offerings were simply not being catalogued quickly enough. With its current cataloguing staff they probably never would be. The result was poor findability not only for potential clients, but even for Magnum's own sales and marketing staff. Moreover, the lack of findability of photos was threatening to undermine the confidence of member photographers in adding to the collection.

In finding a solution to these problems, Young set three goals: improving search, improving economy, and

improving workflows. The solution she put in place used crowdsourcing – harnessing multiple people to apply small bits of information to images.

"There are misconceptions about crowdsourcing – that it's a Wild West approach that can't produce quality results," Young said. There is, she admitted, some truth to that, but only because "the wisdom of crowds can't be realized without proper systems." Magnum got a proper system in place, using the services of Tagasauris, a crowdsourcing and media annotation start-up led by Todd Carter. Tagasauris provides structure to workers recruited through Amazon's Mechanical Turk service (and, for Magnum, volunteers from its Twitter followers). The job of adding metadata is broken down into small tasks that even non-experts can perform. By having multiple people do each task to confirm their choices, or even having some workers vote on the results of tasks completed by others, the quality of metadata is kept high.

The results were impressive – with much richer metadata improving the findability of images. Sales increased by about 8% over the course of a little more than a year, and the portion of traffic referred to Magnum's website by search engines increased from 1% to 38%.

Authority Control

Lacy Schutz, Director of Collections Access at the Museum of the City of New York, faced similar challenges. The museum's digital collections were growing – projects she managed had produced digital surrogates of about 100,000 objects. But with only a few cataloguers and funding (mainly from grants) limited, her team simply did not have the time to add sufficient metadata to all their digital images. Moreover, the metadata added to images – which was derived in part from Library of Congress Subject Headings – didn't always meet the needs of visitors searching the museums website. A visitor might search for "car" instead of "automobile," for example. That could be resolved through adding (many) related terms to the museum's lexicon (thesaurus), but an even trickier problem relates to the way potential image licensees might search. An advertising agency might want to find photos that felt "happy" or were predominantly blue – and MCNY's website could not meet that need.

Schutz met Carter of Tagasauris last year, and the two of them spent time talking and thinking about the problem. Crowdsourcing metadata seemed a good approach for the museum, but a key issue for Schutz was maintaining some consistency within the terms used. Schutz, a Pratt SILS, said she "had lexicons on the brain and asked Todd 'what about a controlled vocabulary'?"

That concept was not part of Tagasauris' offerings, but they found a solution. They connected the service to the online thesaurus DBpedia, which is derived from Wikipedia (itself a crowdsourced resource). This allows not only consistency, but brings more semantic relationships between terms, greatly improving the quality of the metadata. The Museum recently got a grant from the National Endowment for the Humanities to test this approach for a part of the collection as a possible model for other cultural institutions.

Web-Scale Knowledge Work

Carter provided details of how Tagasauris works, but first tried to put the problems information professionals face in context. "Four million photos are added to Flickr each day, and three billion to Facebook every month. We're facing a tidal wave of information and have to find a way to deal with that. We're all drowning in a sea of data, and need to find ways to organize it, to add value to it so it becomes useful information and knowledge." Crowdsourcing done right, he explained, can be a "force multiplier" for knowledge work, taking it to a scale that can address the information explosion.

He talked about some great thinkers whose ideas are reflected in this work. One was Henry Ford, whose assembly line multiplied human labor. He described management thinker Peter Drucker, who said that knowledge work is most important of all. And he pointed to Alan Turing, who developed the concept of algorithms for computing. Through algorithms, computing power can be adapted to supplement human action, with machines doing tasks for which they are better suited. "We're seeing changes in the nature of work, with tasks coming to workers when they want it and when they are interested in it, and better division of labor between people and machines."

Tagasauris enables breaking knowledge work into little pieces. Computers do some tasks, such as determining if an image is in color or black-and-white. The same image might then be passed to crowdsourced workers, asking them to decide whether the scene depicted is indoors or outdoors, or to identify objects in it. Or computer algorithms can be used to pull faces out of an image, but identifying them might be left to human workers. And by linking some choices to open metadata collections such as DBpedia, and offering terms from those sources as choices to human workers, the metadata that they apply can be controlled and connected.

The system is set up so that clients – staff at cultural institutions such as the Museum of the City of New York or Magnum – can design a series of tasks themselves, turn the tasks over to online workers, and end up with quality descriptions for their objects quickly and at reasonable costs. He had examples of the process running

live during the discussion, with over 2,000 people working on tasks for Magnum at the time.

Interestingly, both Magnum and MCNY have found the number of terms they are using is much larger now that they link to metadata from other sources, with much richer descriptions than were possible with the controlled vocabularies they had developed internally.

Sharon Middendorf, a co-founder of Tagasauris, added that they soon will make the service available to much smaller institutions, such as small historical societies and other cultural institutions, including through a service in which images in an organization's Flickr collection can be annotated and the metadata taken back out easily.

Q&A

The discussion then turned to questions from Anderson and the audience. Several focused on why workers do online tasks, and how they are recruited. Workers who come through Mechanical Turk are paid, but Carter said he believes that the tasks are (and should be) interesting. People seem to enjoy looking at and working with great photos or images of cultural objects. The Twitter followers Magnum uses are volunteers. They may value getting a look at great historic or new images few other people have seen, and given their clear interest in photography might be best suited for more difficult tasks. Aurelia Moser, a Pratt SILS student, remarked about the social aspects of this, raising possibilities for using social media to create communities working together on these sorts of tasks.

Krissa Corbett Cavouras, a recent Pratt graduate who manages medical research, asked if this approach could be applied to non-photographic collections, such as large quantities of text or even datasets from scientific research. Carter felt that it could be, but doing so would require very careful design to address issues of confidentiality and privacy. It might also be hard to make such tasks as interesting as working with photos, which could create challenges in recruiting.

That related to one of the last questions of the evening, from Kimmy Szeto of the SUNY Maritime College Library, about the implications of crowdsourced approaches for information professionals such as those of us in SLA (and in library schools such as Pratt). Both Carter and Young responded by pointing to the care and expertise needed in setting up work processes for crowdsourced workers. A lot of planning, careful testing and good management are needed to make jobs interesting and efficient while meeting the needs of the institution. That takes skill.

Schutz pointed out that all the cataloguers on her staff asked to be involved in this project, bringing their expertise to bear on process design. They also are very much still needed for difficult cataloguing tasks for which novices simply don't have the expertise. And she helped end the discussion with a blunt statement: "Do we get degrees such as an MLS so we can do basic photo tagging for our whole career? That undervalues our education. Figuring out how to use what we know and engage with technology to improve collections and access, at scale, is far more interesting."

Leave a Reply

Name (required)

Mail (will not be published) (required)

Website

Submit Comment

Upcoming Events

JOBLOG

Event Calendar

Chapter News

Follow Us!

Categorites

Select Category

Recent Events Gallery

Archives

Select Month