Thanks for Sharing

A Case Study on Crowdsourcing Materials and Metadata for Digital Collections

Sarah Sweeney
Northeastern University Libraries
...

Digital Library Federation Forum
October 16, 2018

Context

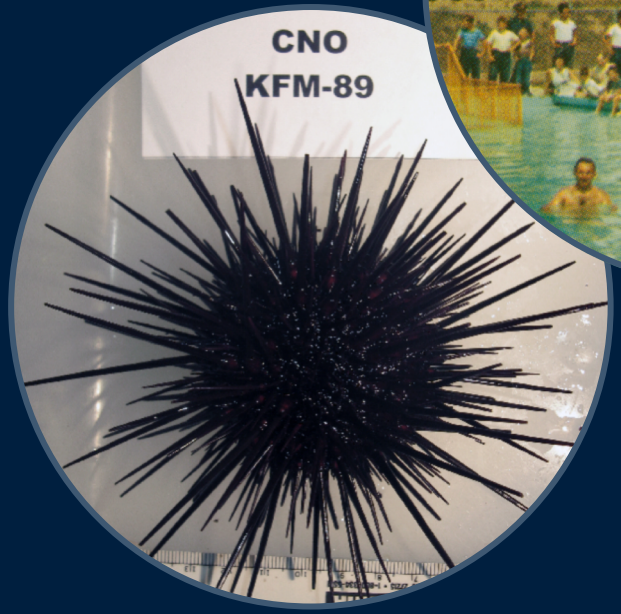
Digital Repository Service (DRS)

https://repository.library.northeastern.edu

1,588
collections

150,000+
digital objects

1,000,000+
downloads



Our Marathon

https://marathon.library.northeastern.edu

Project Team

Elizabeth Dillon
Ryan Cordell
Jim McGrath
Alicia Peaker

Content

22 collections
183 a/v files
630 documents
7,074 images

Our Marathon

Public Submission Elements

Dublin Core

- Title
- Description
- Date

Contributor information

- Current location
- Name
- Age
- Race
- Gender

Other

- Submission text
- More I might be forgetting...

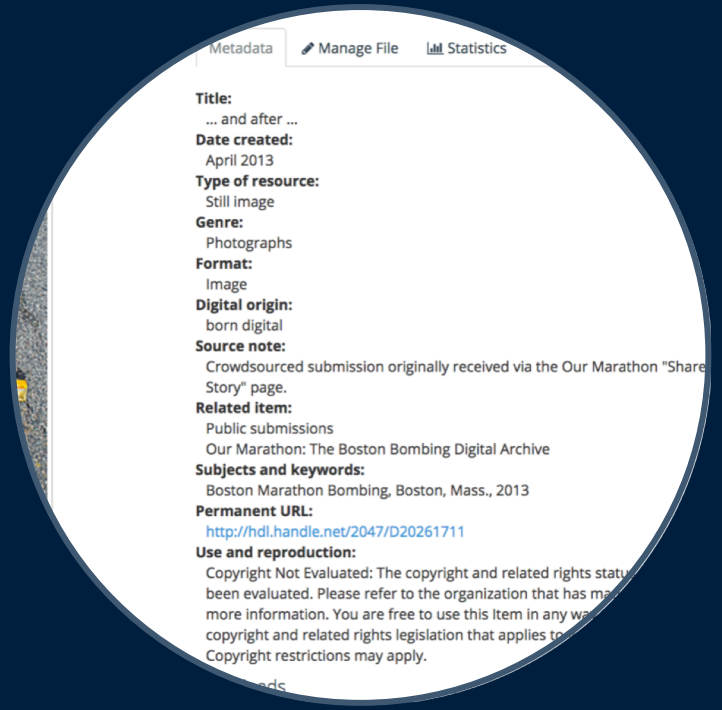
Our Marathon

Kept
(when asked to)

- Title
- Creator
- Description
- Contributor information
- Submission text

Added

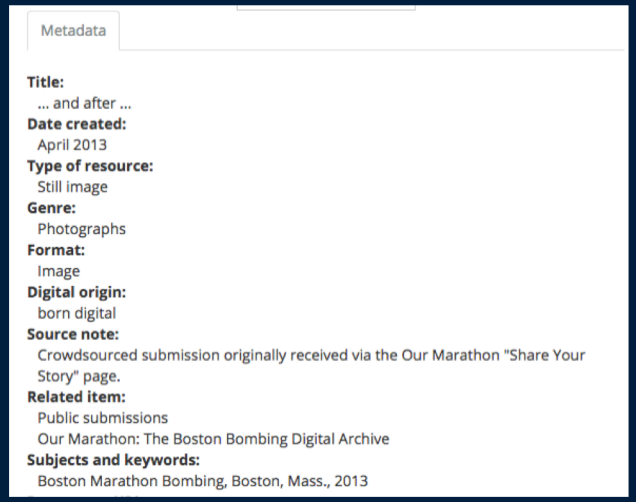
- LCSH
- Rights Statements
- Notes



Only fixed obvious errors

Our Marathon

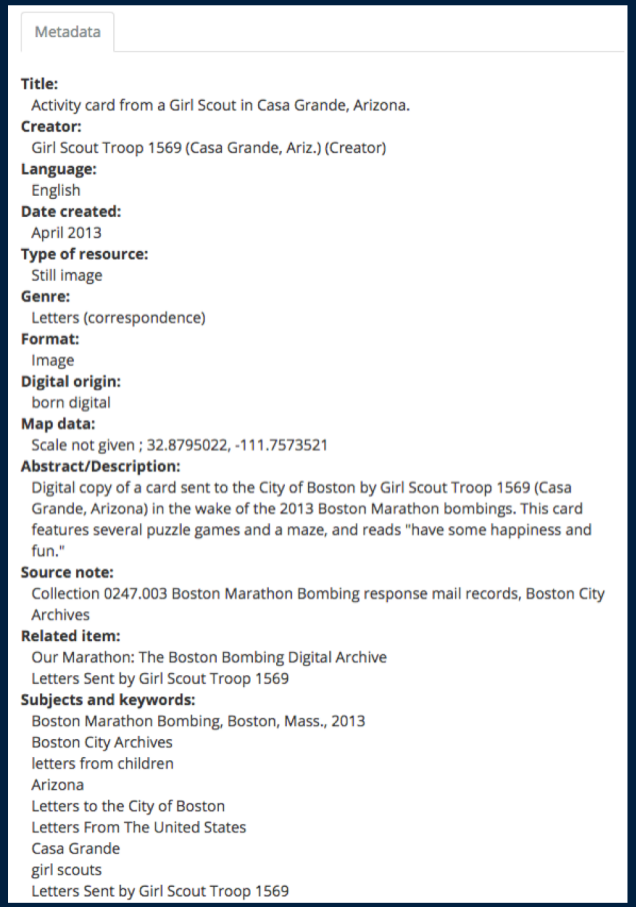
Public Submission



Metadata

Title:
... and after ...
Date created:
April 2013
Type of resource:
Still image
Genre:
Photographs
Format:
Image
Digital origin:
born digital
Source note:
Crowdsourced submission originally received via the Our Marathon "Share Your Story" page.
Related item:
Public submissions
Our Marathon: The Boston Bombing Digital Archive
Subjects and keywords:
Boston Marathon Bombing, Boston, Mass., 2013

Partner Submission



Metadata

Title:
Activity card from a Girl Scout in Casa Grande, Arizona.
Creator:
Girl Scout Troop 1569 (Casa Grande, Ariz.) (Creator)
Language:
English
Date created:
April 2013
Type of resource:
Still image
Genre:
Letters (correspondence)
Format:
Image
Digital origin:
born digital
Map data:
Scale not given ; 32.8795022, -111.7573521
Abstract/Description:
Digital copy of a card sent to the City of Boston by Girl Scout Troop 1569 (Casa Grande, Arizona) in the wake of the 2013 Boston Marathon bombings. This card features several puzzle games and a maze, and reads "have some happiness and fun."
Source note:
Collection 0247.003 Boston Marathon Bombing response mail records, Boston City Archives
Related item:
Our Marathon: The Boston Bombing Digital Archive
Letters Sent by Girl Scout Troop 1569
Subjects and keywords:
Boston Marathon Bombing, Boston, Mass., 2013
Boston City Archives
letters from children
Arizona
Letters to the City of Boston
Letters From The United States
Casa Grande
girl scouts
Letters Sent by Girl Scout Troop 1569

Art of the March

http://artofthemarch.boston

Boston Women's March, January 21, 2017

Project Team

Nathan Felde
Alessandra Renzi
Dietmar Offenhuber

Rescued ~2,000* discarded posters

Art of the March

~4,000
rescued posters

3,004
posters cataloged*

5,986
digital images

Art of the March

What materials were used to create the poster?

Paper, Cardboard, Poster Board, Fabric, Acid Free, Other

What is the poster medium?

Ink, Paint, Chalk, Pen, Other

What lettering style is used for the poster text?

Typeface (imitative), Block, Cursive, Decorative, San Serif, Serif, Other

Lessons

Crowdsourcing materials?	*very difficult!*
Crowdsourcing digitization?	*surprisingly effective!*
Crowdsourcing metadata?	*complicated!*

Lessons

- How much metadata do we need and how much can we expect to receive?
- Is the submitted metadata an artifact or just a starting point?
- How will the metadata provenance be documented?
  - Is the fact that it’s crowdsourced even important?

Lessons

… and after …

Permanent URL: http://hdl.handle.net/2047/D20261711

Metadata | Manage File | Statistics | Reindex

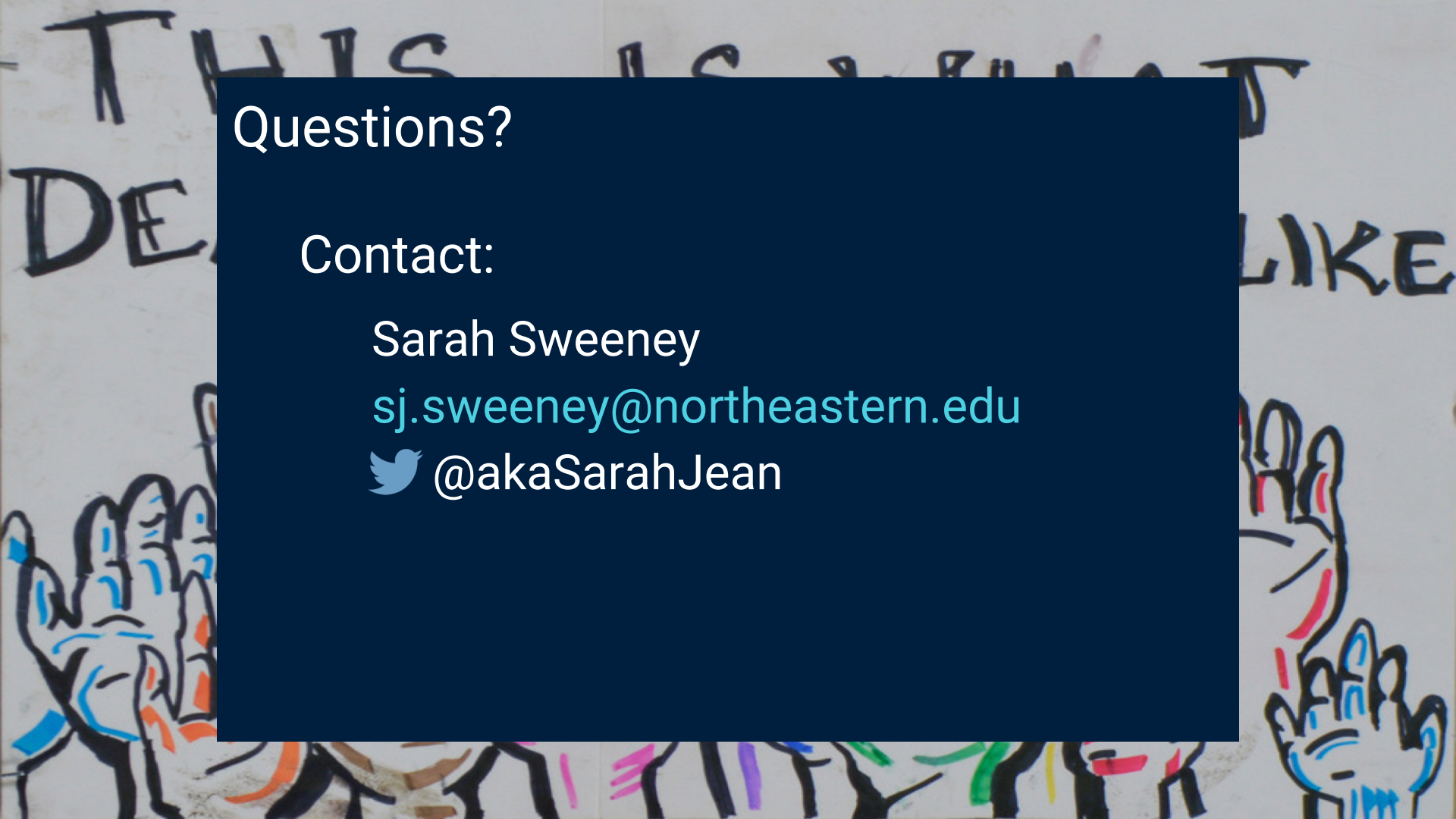
Title:
… and after …
Date created:
April 2013
Type of resource:
Still image
Genre:
Photographs
Format:
Image
Digital origin:
born digital
Source note:
Crowdsourced submission
Story" page.
Related item:
Public submissions
Our Marathon: The Boston Bombing Dig
Subjects and keywords:
Boston Marathon Bombing, Boston, Mass., 2013
Permanent URL:
http://hdl.handle.net/2047/D20261711
Use and reproduction:
Copyright Not Evaluated: The copyright and related rights status of this Item has not been evaluated. Please refer to the organization that has made the Item available for more information. You are free to use this Item in any way that is permitted by the copyright and related rights legislation that applies to your use.
Copyright restrictions may apply.

"Make a suggestion" *instead*

Downloads

Report a problem

Master Image | Save for later


Questions?
Contact:
Sarah Sweeney
sj.sweeney@northeastern.edu
@akaSarahJean

Digital Repository Service (DRS) : https://repository.library.northeastern.edu

Materials used:

- Bardow's Cottages : http://hdl.handle.net/2047/D20204197
- Strongylocentrotus franciscanus : http://hdl.handle.net/2047/D20264037

Our Marathon: The Boston Bombing Digital Archive : https://marathon.library.northeastern.edu

Our Marathon Public Submissions

- http://hdl.handle.net/2047/D20259798 or
- https://marathon.library.northeastern.edu/collection/neu:cj82qn382/

Materials used:

- .. and after … : http://hdl.handle.net/2047/D20261711
- Activity card from a Girl Scout in Casa Grande...: http://hdl.handle.net/2047/D20276407

Art of the March : http://artofthemarch.boston

Materials used:

- “Hubris”
- “Ovaries / Tina”
- “What Democracy Looks Like”
- “Rebellions Built on Hope”

Introduction

I'm going to share with you the experience we've had at Northeastern University with crowdsourced collections from two different perspectives:

- One from the angle of taking ownership of an archive of crowdsourced material and metadata
- Another from the angle of attempting to create metadata and digitize materials using crowdsourced help for a faculty-led archiving project

I'm going to talk mostly about crowdsourcing metadata, but I'll touch on collecting digital materials, as well.

A quick bit of background, for context: I manage the Digital Repository Service, known as the DRS, at Northeastern University Libraries in Boston, Massachusetts. We use a custom-developed Samvera repository to store more than 150,000 digital objects either created or acquired by the Northeastern community. This includes theses and dissertations, research publications, and monographs, but it also includes archival documents and photographs, as well as materials used in faculty research, like 2,000 images of ocean specimens collected around the world and more than 1,000 materials documenting the experience of visitors to the Catskills mountain resorts in the mid-twentieth century. All of this to say: our faculty and staff spend their time working and researching in a variety of different areas and disciplines and we do our best to support them by providing space for the materials in the Digital Repository Service, the DRS.

This often means working with project or research teams to transfer or migrate materials into the DRS, or help process large batches of materials so they can be made available and discoverable in the DRS. Like many institutions, we are interested in the possibility of using crowdsourcing to improve access to our materials, but we have not had much experience organizing crowdsourced submissions for our collections ourselves. We have provided assistance in a few projects for which crowdsourcing was a major component, and in some cases made the project much more successful than it may have been without the effort. I'm going to talk about two of those projects today. The Our Marathon project, which focuses on how we in the library became stewards of crowdsourced materials and metadata and what that meant for our metadata workflows. And the Art of the March project, which focuses on a crowdsourced effort to digitize and describe an entire collection and how complicated that effort was.

Project One - Our Marathon

Our Marathon is a crowdsourced archive of pictures, stories, and other materials related to the Boston Marathon. The project was started by two faculty from the Northeastern Department of English, Elizabeth Dillon and Ryan Cordell, just after the Boston Marathon bombing in April 2013. The archive was intended to be used as a space to collect and preserve digital artifacts related to the marathon, the bombing, and the experiences of those who were impacted by bombing and subsequent shut down of the city to capture the perpetrators. The Our Marathon project team, led by the co-directors Jim McGrath and Alicia Peaker, collected material for a few years by setting up a website using Omeka and accepting submissions from the public. Through a tremendous amount of outreach, they received more than 600 submissions in Omeka, including photographs, personal stories, social media posts, videos, memes, police scanner recordings, and others. They partnered with many local organizations, including Boston Medical Center and the Countway Library to house collections of digitized cards, letters, and messages of support sent by caring individuals all around the world. There's also a wonderful oral history series with runners and first responders created in partnership with WBUR, our local NPR station, which I highly recommend.

In 2014, management of the website and digital collections moved to the library's Digital Scholarship Group. As the five-year anniversary of the bombing approached, we met with Jim McGrath to discuss migrating the materials out of Omeka and into the Digital Repository Service. Although Omeka was still a useful platform, the site was no longer accepting submissions. Thinking of long-term preservation needs, we wanted to move the site and the materials into our local ecosystem, where we could maintain the files in the DRS and present them using WordPress, which is commonly used and well-supported in our library. From August 2017 to April 2018 we migrated the files and metadata and rebuilt the original Omeka site in WordPress. There were lots of elements to this project and lots of work from a lot of people, but I want to focus very specifically on the crowdsourced items and their metadata.

When participants submitted their photograph or story or video to Omeka, they were asked to submit descriptive metadata along with it: a title, a description, the creator's name, their name as the submitter, select a location to place the record on a map, among others. When we migrated, we had to map those values to MODS, the standard we use for descriptive metadata. When it came to migrating the metadata, we kept what we needed, ditched what we were told to, and did our best to respect the original data we were given.

But, this collection of materials is one we would like to contribute to DPLA through our regional hub, Digital Commonwealth, and the descriptive metadata as-is did not comply with their minimal standard for harvesting. So we worked with an excellent intern from the Simmons

University Library Science Masters program to review the records and update them so they were compliant. This raised questions for both of us and the library's metadata team about how to carefully edit the metadata for the more than 600 items submitted by the public through Omeka. If treated like any other digital collection, we would have made a lot of changes to the descriptive metadata, including error fixing and augmentation, like editing the titles to make them more descriptive, deleting some keywords and replacing them with more accurate terms, and doing our best to disambiguate or clarify some of the submitted names.

We ultimately decided that the submitted metadata for publicly submitted collections was itself an artifact of the submission and respecting that was an important part of the process. But, we also wanted to make sure the items were discoverable. We couldn't willfully leave a typo in a title or keyword, so unless the word seemed intentionally misspelled, those were updated for discoverability. So, we added LCSH subject headings, notes, and rights statements to the records and left the rest alone.

This makes the cataloging for the different collections in the Our Marathon project vastly different. The public submissions collection is minimally described, but the Oral Histories and Letters collections have fuller titles and deeper descriptions. We prefer to assert more consistency than that within a particular project, but in this case it seemed more important to preserve the information submitted by a user about their digital object than it was to make the cataloging appear consistent.

It's unclear at the moment how we should document these decisions. Right now it lives in our institutional memory, but it won't be long before the decision we made when working with this collection becomes a fuzzy memory for our future selves. Also, by choosing to respect the content of the metadata in the submitted record, did we just create a cleanup project for future metadata staff? Moving forward I'd like to consider using the MODS recordInfoNote field to document the description and its provenance, or find some other way of indicating the decisions we made around this collection.

Project Two - The Art of the March

The second crowdsourcing project I want to talk about is The Art of the March. This project focused not just on crowdsourcing metadata, but crowdsourcing digitization, as well. This led to the creation of a website for exploring the digitized materials and the research around those materials.

Boston Women's March was one of many world-wide protests held across the world on January 21, 2017. As the Boston March wound down, Alessandra Renzi, Dietmar Offenhuber, and Nathan Felde, three Northeastern faculty from the College of Arts Media and Design, came across a large pile of posters that had been discarded by protesters and collected by city cleanup crews to be taken to a landfill. The faculty recognized the artistic qualities of the posters and intervened. They were told they could have the posters if they could remove them within an hour. They mobilized quickly, renting a van and grabbing volunteers to stash what they estimated to be about two thousand posters.

The faculty team reached out to several campus groups that night to coordinate different facets of potential research activity related to the posters. Their primary goal for the project was to make the digital versions of the posters available in an archive to support research activity. The team was particularly interested in researching common themes, the emotional sentiments expressed in the posters, as well as the techniques used to create them. This lead to building a website for exploring the archive and sharing research related to the posters.

The faculty team reached out to the library for guidance on digitizing the posters, on where to host the digital materials, and how to catalog them so the themes, emotions, and techniques could be researched. The faculty team was also interested in collaborating with the library to make the archiving work, the digitizing and cataloging, a collaborative effort - not just among various Northeastern units, but collaborative in the Boston community.

But how do you get a community to digitize and catalog two thousand three dimensional objects of varying sizes and shapes? They organized and hosted a two-day digitizing sprint where volunteers could help photograph, describe, and sort the posters. A warehouse on the waterfront was rented for a weekend, volunteers arranged posters on the floor so they could be photographed en mass. They were then sorted by size and sent to stations set up for rapid photography.

Once a poster was photographed, it was given a label with a number and sent to a metadata station. A Google form had been designed for the event to capture as much information as possible. Volunteers could use their phone to enter what they could about the poster, including a title, a description, select a theme, describe the emotion, and even transcribe the text.

Overall the two-day event was a success. Every single poster was photographed, even though the number of posters was almost double the original estimate, which generated nearly 6,000 images with front and back captures. This goal was likely achieved because most attendees were interested in helping with the photographing activity, certainly more so than the metadata capture, which was a lot less successful.

With the exception of a few enthusiastic participants, volunteers found describing the posters to be tedious and less engaging than working with the photography. In fact, fewer than 5 people volunteered on the second day to contribute metadata. The form simply asked for too much information and it took several minutes to fill out the form for each poster. The form was edited and trimmed throughout the first day until we were only asking for a title, a few thematic elements, the poster number, and a transcription. The subjective descriptive elements were difficult to describe, as well as the elements that required some material knowledge, like the technique used to create the posters: Was the poster made of poster board or cardboard? Did the creator use ink, pen, or markers? Was a block font or a serif font used? Unfortunately the description portion could not keep up with the pace of the photography, which led to a large backlog. In the end, most posters got a title, but little else. Instead of relying on real time description, we ditched the metadata altogether and pivoted to a post-event plan for having Northeastern work study students describe the posters and provide metadata.

As a part of the same project, there was also a much smaller effort to collect submissions from the Boston community to contribute to the archive. We received fewer than 20 submissions and the ones we did receive were crowd shots or were posters from other marchers or other marches in other cities, not Boston, which was very disappointing and we ultimately dissolved this portion of the project.

These are two very different projects with different approaches to crowdsourcing, but let's look at them together.

Not surprisingly, soliciting materials to be submitted to a crowdsourced project is difficult. Stumbling on a large cache of materials with research value isn't exactly crowdsourcing, but I think you could stretch the definition to include it. But, if you're not lucky enough to stumble on a collection like the Art of the March project team did, then I suggest that when you try to solicit material for a crowdsourced collection, you should find an audience for the work and do lots of outreach. There was a lot of enthusiasm after the marches that weekend and I think the library could have received more appropriate submissions by being more explicit about what the goal of the submissions project was and being more intentional with our solicitation for material. The material sourcing portion of the library's work on Art of the March project really made me appreciate the detailed work done by the Our Marathon project team to find partners and reach out to the community, which was a huge contributor to their success.

I think the Art of the March project proved that crowdsourced digitization could be effective! There were some issues with focus, image orientation, and cropping, but otherwise it was incredible that a group of around 100 volunteers could photograph 4,000 items in two days. If you have the time, space, and labor, this seems like a serious possibility for collections with similar traits (you may not want to open up your rare materials for crowdsourced digitization, of course). But, an exception does not prove the rule. This method worked for the Art of the March

team, but I'm curious to hear about other projects that have tackled crowdsourced digitization with similar results.

Crowdsourced metadata is complicated. Ultimately, I think it's fair to say that crowdsourcing some light descriptive metadata is easy, but gathering the right amount of metadata to support discovery is very hard. Our decision not to alter much of the publicly submitted metadata was made possible because the Our Marathon project gathered just the right amount for their discoverability needs. The Art of the March approached metadata collection differently because that project had different needs and a much shorter runway for collection. In retrospect, we should have treated metadata gathering as step one of many. We should have asked participants to provide just a transcription, or just a description, or just two categories. That simple metadata could then have been used as the foundation for another longer-term crowdsourced project, or as a starting point for the work study students.

Looking forward, before we embark on our next crowdsourcing project, I'd like to make sure we ask a few key questions at the start, like:

How much metadata do we need and how much can we expect to receive?
Is the submitted metadata an artifact or just a starting point for future augmentation?
How will the metadata provenance be documented?
Is the fact that the metadata crowdsourced even important to us or the project team?

Crowdsourcing metadata is something we talk about regularly in my little circle of Northeastern. We have a great metadata team with lots of expertise who are able to describe many of our digital collections, but there are many more collections that would benefit from the expertise of the crowd, or from a particular community. We're experimenting with a few ideas about how to accept, vet, and update metadata in the DRS using input from the public. In particular, we're thinking about adding metadata submissions options to the information page for each item in the DRS, and in places where the DRS API is used to display DRS materials in WordPress. There are workflow implications of this process, of course, but ultimately I think that involving voices and expertise that comes from outside the library will be hugely beneficial for our digital collections. This is where I want to hear from you: What lessons have you learned from gathering crowdsourced metadata or materials? What about crowdsourced digitization? I'd love to hear about other experiences with that effort.