Crowdsourcing Digitization: Harnessing Workflows to Increase Output

Gretchen Gueguen
East Carolina University
guegueng@ecu.edu

Ann Hanlon
Marquette University
ann.hanlon@marquette.edu

Crowdsourcing is a term that was coined by Jeff Howe in a 2006 article for *Wired* magazine to describe outsourcing a large task to a large group of people by harnessing the internet to bring them together. Our use of the term "crowdsourcing" taps into a smaller crowd – our patrons, as well as the crowd that are our very own co-workers by:

a) Capturing existing workflows, such as digitization for patron request, and using those to build robust digital collections
b) Allowing selection to be primarily driven by patron request
c) Creating a balance between decentralized and centralized staffing models

Our talk will be divided into four parts:

1. **The Wisdom of Crowds**
   How the project was conceived and developed, and how it succeeded

In 2005 two projects at the University of Maryland Libraries converged: one to create an "image management system" that would help the archivists, curators and librarians manage all of the digital files being created due to patron request, exhibits and preservation efforts; the other was to create the Office of Digital Collections and Research to support the teaching and research mission of the university by facilitating access to digital collections, information, and knowledge. These two efforts combined to create a digital object repository using the Fedora object repository that would be partially populated by the items digitized by archives staff. As the Fedora repository was developed, policies regarding metadata and digitization were created and a temporary MS Access database was created to capture metadata as images were scanned. Ultimately these images and metadata records would be ingested into the repository and future images and records would be added directly through an administrative interface.

Once the Fedora repository was in place, we found that, through the course of scanning for all of these individual requests, we had the makings of at least two robust digital collections and the potential for many more. We realized that by simply harnessing our existing workflow in a new way, we could create useful digital collections. These collections also had the benefit of being the result of the selections of many of our patrons and employees over time -- they represented materials that we knew were used and would continue to be used.

2. **The Madness of Crowds**
   How the project almost failed, why, and how we brought it back from the brink

Despite the project's successes, there were naturally bumps along the road. In particular, the project suffered from the lack of a plan for metadata and digitization training and the fluctuations caused by the development of the infrastructure of the repository itself. Many of the earliest digital objects, created before documentation was fully developed, suffered from incomplete or incorrect metadata, improperly devised file names, and image quality deficiencies. Still others were created after documentation was

available, yet issues persisted. These issues of quality are quite apart from the fact that a backlog of at least 5,000 images and metadata records accrued while the Fedora repository was being developed. Finally, once we were able to begin the process of ingesting these objects, we needed a way to organize them into collections or provide some other way for users to browse through what could have essentially been a random selection from across the libraries.

3. **Crowd control**
   The methods we used and lessons learned in our efforts to use crowdsourcing to build digital collections

The problems with the project came down to an inadequate plan for this distributed workflow. While the idea of capturing what is truly an existing workflow (patron scanning requests) is an elegant one, there is no question that it also adds a burden. Indeed, it can slow it down the original workflow considerably - not something desirable when you are dealing with a patron who themselves has a deadline, or simply a reasonable expectation for good service. Eventually we realized we needed to do three things:

1) *Create a bottom line and a clear set of standards:* We needed to sort out what quality factors to prioritize. Completeness? Standardization? Other easy to measure rubrics? Eventually we were able to set up some basic control and review processes to improve the quality of the images we were producing.
2) *Don't underestimate the value of personal communication:* Over time we realized that documentation and simplified directions were beginnings, but not enough. We were expecting people to take on new responsibilities, and needed to give them more support for the day to day decisions they would face.
3) *Plan for things like repository organization and backlogs of work:* Knowing that we would have a backlog would have allowed us to plan a regular schedule of maintenance on this group of records. It would also have helped us think more about the overall organization of the repository so that we could better manage that backlog. Finally, a sense of the pacing of a backlog would have alerted us to when there truly was a problem (i.e. if too many records were piling up) and helped us analyze places where we needed to be more efficient.

4. **Attracting a Crowd: Critical mass for the masses (and serious researchers)**
   Why it's smart to build digital collections this way and our ultimate goal

The concept of crowd-sourcing, and really, the concept of capturing the existing workflow in patron requests, means that collections are being built in an ostensibly neutral manner. Nothing is ever really without interpretation or bias, of course, and nothing could highlight that more clearly than the very pronounced bias toward sports in our collections. But this concept and practice of "neutral collection-building", as opposed to collection building based on selection, captures items that have value to someone other than the curator and builds an online resource in a way that more formal selection can't. It is similar to the concept of Wikipedia versus Encyclopedia Brittanica - the sheer breadth of subjects and knowledge contained in Wikipedia far outweights that of EB and has made it a fundamental source for systematic research - even if you do have to check your sources later.

That is ultimately where we see our crowd-sourcing scheme leading. By focusing on ways to streamline the process of building digital collections, and involving as many players in the process as is possible and effective, digital collection building becomes a core function of the library. Eventually, digital collections begin to build to a critical mass, so that researchers can come to the web to conduct original research using primary sources, and not just to see what treasures a library might hold.