Why Collecting History Online is Web 1.5

Sheila A. Brennan and T. Mills Kelly
Center for History and New Media, Case Study, March 2009

It seems like only yesterday that we were transitioning from the first-generation, read-only web to the "read-write web" of Web 2.0, that fosters community and collaboration where users participate in online content creation. But does Tim O'Reilly's idea of Web 2.0 really work for the collecting and preserving of history online?[1] Not really. We, as digital humanists, however, are comfortable with that.

Now it is very common for websites from media outlets to museums to ask for input, comments, or stories from online visitors, but back in 1998 when the Center for History and New Media (CHNM) at George Mason University first engaged in collecting and preserving history online, such practices were new. The Blackout History Project (http://blackout.gmu.edu/), 1998, invited visitors to complete a lengthy on-line survey and asked contributors to provide a phone number so that a longer oral history interview could be conducted on the Northeastern blackouts in 1965 and 1977. Blackout offers a good example how a few historians were transitioning from traditional oral history to digital collection methods before the term "Web 2.0" was a glint in O'Reilly's eyes.

Even as target users were still getting comfortable with sitting down at the computer to type up a personal story or upload a photograph that would be shared in a public space on the web, CHNM pushed forward to experiment with digital collecting models including the Exploring and Collecting History Online (ECHO) project in 2001 (http://echo.gmu.edu), followed by the much larger, September 11th Digital Archive (http://911digitalarchive.org). We simply asked online visitors to tell their story.

Dan Cohen and Roy Rosenzweig emphasized in their book, *Digital History*, that collecting history through digital archives can be far cheaper, larger, more diverse, and more inclusive than traditional archives. This democratization however, does not mean compromising the quality of the historical work.[2] Through a number of different projects, CHNM uses digital media to create models and tools so that other organizations may build their own collections of images, texts, audio, and video files that encourage diverse audiences to be active participants in the saving and shaping of their own history. Like so many digital enthusiasts, we were convinced from the start that we were creating a model for a participatory "archive of the future" and as Web 2.0 model began to appear, we knew the work we had done already fit into some of the Web 2.0 principles as we sought to do more with digital archives.

We remain convinced that we (and others) are building archives of the future. It just turns out that achieving our goals requires more work than we thought and several of the core principles of Web 2.0 don't work when it comes to building these archives. Unlike Wikipedia, for instance, our collecting work could not succeed in an entirely digital and editable mode. For this reason and others, we have concluded that collecting history online floats in a world between the uneditable, didactic Web 1.0 and the completely open and editable Web 2.0, leaving us with a place we are calling, "Web 1.5."

What follows is a tale that we think offers some important insights into collecting and preserving history online that we learned from building the Hurricane Digital Memory Bank (http://hurricanearchive.org). We demonstrate how you can create a digital archive and encourage public participation without losing the integrity of evidence collected or compromising the privacy of a contributor.

Building the Hurricane Digital Memory Bank

Soon after Hurricane Katrina roared ashore on August 29, 2005, the staff at CHNM quickly realized that we were witnessing a very significant moment in American history. Television and newspaper coverage of hurricane victims stranded on rooftops, houses blasted from their foundations along the Mississippi coast, the displacement of tens of thousands of Gulf Coast residents, and the subsequent failures of all levels of government convinced us that we needed to act quickly to begin collecting the history of this terrible disaster. Hurricane Rita's arrival a few weeks later merely reinforced that we had a job to do.

Working in partnership with the University of New Orleans (UNO), we quickly launched the Hurricane Digital Memory Bank (HDMB) in an attempt to collect and preserve as much of the "instant history" of these events as possible—history that was being created and published by thousands of average people in their personal blogs, on photosharing websites, and YouTube. Our experiences with the September 11 Digital Archive had taught us a lot about collecting history online and so we expected that like the very successful earlier project, the HDMB would take off quickly and would rapidly become a central digital archive of original sources, many of which disappear almost as quickly as they are created.

The HDMB website launched in early November 2005, less than three months after Katrina's landfall. To get to that point we followed the rules we'd developed in the September 11 project. First, we needed to design a site that was easy to navigate, loaded quickly in a variety of browsers, and made contributors feel comfortable sharing their personal stories and images. The homepage explained our mission and provided the basic site navigation to browse, contribute, and learn more about the project. Without that ease and comfort of use, we knew the number of contributions to the project would be limited.

Our target audience was anyone who was affected by the 2005 hurricanes: survivors, volunteers, concerned citizens. We asked them to contribute their first-hand accounts of the storms, the aftermath of each storm, and how their lives had changed as a result of the fury of the two hurricanes that summer. Drawing on our experiences in other collecting projects, we created a short and simple process for contributors to share and upload on-scene images, podcasts, or other born-digital files they might have, or to copy blog postings or emails and submit them to our archive.

Using a mashup of GoogleMaps we also asked contributors to geolocate the content of their submission by entering a zip code or street address into the contribution form. Because this disaster forced the relocation of so many people, we thought compiling geolocation data would be a critical component for contextualizing this evidence in the future. Finally, we designed the site to W3C standards to be as accessible as possible to a variety of visitors.

Once the database and design were ready for public launch, we needed to drive traffic to the website. Our publicity efforts followed a process similar to one we used with great success in the September 11 project. In that earlier project we learned that potential contributors visiting the website wanted to see other contributions before they shared their own stories or uploaded other content. So, we seeded the archive with a number of detailed personal reflections and images submitted by University of New Orleans students and their families.

We then developed a series of relationships with local partners, asked those partners to promote the project far and wide, and then did our best here at CHNM to promote the project as well--writing about it on our various blogs, talking about it at conferences, posting notices on listservs, contacting bloggers and the conventional news media, sending out mailings, and so on. We also understood that while we could reach more people with a web interface there were still plenty of people who were unwired for various reasons. To combat that, we did two things.

First, we set up a local phone number through the online telephone company Skype (http://skype.com)that allowed those without connectivity, or for those wishing to talk through their ordeals, the opportunity to contribute via voicemail. These digital audio files were then uploaded directly into HDMB.

Second, we printed postage paid reply cards so that someone could pick up a card, write their story, and mail it in. We would then scan and upload it to the archive. With these efforts combined with the broad publicity campaign, our outreach team sent more traffic to the HDMB site.

Up until that moment, everything had proceeded according to plan. But to our surprise, all the national media coverage of the storm aftermath and the combined efforts of our staff here in Virginia and of our many partners along the Gulf Coast did not result in anything like the flood of contributions that we expected when we launched the project in a publicity push during Mardi Gras 2006.

To be sure, the lack of a flood is not the same thing as a flop. At this writing the HDMB database contains almost 1,300 personal reflections, more than 13,700 digital images, and more than 7,000 other files (everything from newspaper articles to PowerPoint briefings given by the National Guard units). With more than 25,000 digital objects in its archive (some not available for public browsing), the HDMB project is one of the largest repositories of sources on the hurricanes of 2005.

But even at 25,000 digital objects, the project did not live up to our expectations.

One reason our collecting work has been harder than we thought it would be is that our first big project—the September 11 Digital Archive—worked so well. As of this writing, the September 11 project contains more than 150,000 digital objects, including more than 40,000 personal stories and 15,000 digital images. Although the first 10,000 contributions to this project were hard to come by, by September 11, 2002, contributions were rolling in of their own accord. [3]All we had to do was manage the flow.

Of course, we knew that September 11 was a unique moment in American history--one that seemingly touched everyone living in America at the time--but even so we thought that other major national events would generate similar, albeit smaller, flows of contributions to our servers.

Moreover, we thought that as the average person became more connected to the Internet, the average person would be even more inclined to contribute something to HDMB.

Imagine for just a moment the differences between 2001 and 2005. In 2001, plenty of people were still using film cameras, only a small fraction of the population had a camera in their cell phone, and blogging had not yet become ubiquitous. By contrast, in 2005 it was increasingly difficult to purchase a cell phone without a camera, Flickr.com and other photosharing websites had already aggregated hundreds of millions of digital photographs, and bloggers were everywhere. Surely, we thought, with all that digital content floating around out there, with a well-coordinated publicity campaign, and with a great set of local partners, we would be building a much larger archive.

Lessons Learned

We sort the challenges of creating, managing, and sustaining a digital collecting history project into four main categories: collecting content; technical issues; attracting visitors to your site and building trust with potential contributors; and if your project is one focusing on a tragedy or disaster as ours was, allowing those most directly affected time to heal before they can share. For all four of these reasons, but especially the fourth, we found that we underestimated the amount of staff time that would be required for our project. If you learn nothing else from our experience, estimate the amount of staff time you think your project will require and then add 25 percent to that figure.

Each digital collecting project is unique when it comes to deciding what to collect, but you may find that those sources do not contain much contextual information. It is essential that your project adhere to common metadata standards used by archivists, such as the Dublin Core, but be aware that your contributors may not provide you with much of that data.[4] Without a metadata schema, however, your project is doomed to a lack of interoperability with other collecting projects that potentially threaten its longevity.

As might be expected with a digital project, we experienced a few technical issues that we know depressed the number of contributions we had expected. One problem we encountered was that for all the ease of use and various functionalities we had built into the project, the HDMB interface still wasn't as easy to use as interfaces our target audience was already using. For instance, we had no batch add procedure for digital images, which forced contributors to upload their images one at time rather than through a one-click batch uploader found on popular photosharing websites like Flickr.com. The lack of such a simple multiple image uploader definitely depressed the number of contributed images. On the other hand, a batch-add function does not allow users to annotate each image with unique metadata.

Similarly, some bloggers gave us permission to scrape their postings but would not re-copy their postings into HDMB themselves, because to do so they would have to upload each posting as a unique item in our database. Here the problem was external to us. None of the popular blogging platforms (WordPress, Blogger, etc.) make it easy for the author of the blog to export a subset of their postings that they could then upload. And uploading them individually was simply too much trouble.

The main lesson to be drawn from this part of our experience is that whatever contribution interface you design, it should be at least be as easy to use and efficient as those available elsewhere online. If your potential contributors find your site clunky and difficult to use, they will give up and move on.

This means that designing a website with a low barrier for entry is essential if you want people to contribute through a web portal. Make the contribution form clean and easy to fill out and make sure the main page of the site loads quickly and easily in all of the major web browsers.

In the early stages of our project we found that many people uploaded photos but only a handful shared their stories. In an effort to generate more personal accounts, we added a "quick contribute" form on the HDMB homepage to encourage more people to write a short reflection in response to a simple prompt like, "What will you miss the most?" or "Tell us about a hurricane hero."

At first we were pleased to see that the number of personal stories being contributed went up. However, this advantage was counterbalanced by scores of spam contributions that came in through this interface, consuming a lot of our time trying to outsmart spambots and deleting their fake deposits from the database. After one year of deleting dozens of spam entries each day, we removed the quick contribute form from the homepage. Early on we decided not to use user verification programs like Re-Captcha, because, again, we wanted to make contributing look and feel as easy as possible. [5]

Despite these problems, we did generate a lot of site traffic. Since its launch in November 2005, the site has had more than 2.5 million page views. One reason we believe that traffic to the site has remained strong, long after the events that are its focus, is that we have continued to receive a small but steady number of contributions. Always having something new for visitors seems to draw them back again and again. Considering how quickly "Katrina fatigue" set in around the country, we were very happy with the size of our audience.

Given the nature of the disaster we were dealing with, once we attracted visitors to the site, we had to build their trust and ensure that they felt comfortable sharing personal experiences for the historical record. To that end, we offered contributors many options, including anonymity and the option of submitting their story for researcher access only. We were very careful to ensure that none of the contributions that users wanted to keep private ever showed up in the public interface.

Maintaining control over their personal story turned out to be much more important for our contributors than we had expected. Following the hurricanes, many residents of the Gulf Coast felt that their lives had been taken over by others acting on their behalf and so it was very important to many of them that they retained ownership over their personal histories.

In HDMB, every contributor owns their digital submission and that submission may not be used for any public purpose without their consent. From the start of the project we made it very clear that contributors retained copyrights in an effort to gain the trust of those whose stories we desperately wanted to save.

Knowing of these concerns, we had many internal discussions about whether to allow both contributors and visitors to tag the items in our database, particularly since we wanted to encourage those building this archive to help organize it using folksonomies. [6]

Popular Web 2.0 projects like Flickr.com allow unlimited tagging of records and this ability for users to self-organize the data in the Flickr.com database is one of the many reasons why this particular project now contains more than 2 billion digital images. The Library of Congress is already experimenting with using Flickr to help develop more information on the Library's photographic archive and other institutions, such as the Powerhouse Museum, encourage its online visitors to tag thousands of objects in their digital collection. [7]

The HDMB team was divided on this particular question. Some argued that the more we allowed folksonomies to be created for organizing the archive, the more contributions we were likely to obtain.

Others argued for protecting users' contributions from tagging by others. In the months immediately following the hurricanes, the American people engaged in lively debates about race and class and what those categories meant for the response to the hurricanes. Those on our staff arguing against what might be called "open tagging" did not want any contributor to feel that his or her submission would be drawn into those debates. In this way, it was hoped that the HDMB would become a place where hurricane survivors felt comfortable sharing personal stories.

In the end, we arrived at a Web 1.5 compromise that allowed contributors to participate in saving their past and also to trust HDMB as a safe place that gave them the power to submit and tag their own submissions.

When it comes to figuring out how to drive traffic to your project--traffic that will result in contributions to your archive—expect to do more than launch a website and promote it online. We knew it was going to take personal outreach to generate the traffic and contributions we wanted.

To that end, our two project staff members living on the Gulf Coast developed partnerships with organizations that might help us achieve our goals--local and national groups, including universities, military units, media outlets, non-profit community groups, and museums. Because they were there on the Gulf Coast, our staff could speak to community groups, meet with the local media, attend events--everything from commemorations to Mardi Gras parades where they passed out drink cups printed with our URL and postage-paid reply postcards for collecting hand-written contributions.

The more effort our staff put into building these relationships and having a local presence, the more traffic and contributions we received. Given that the Gulf Coast is still recovering from the 2005 storms, if our funding allowed us to continue to pay staff to work locally, we would want our staff to continue their local collecting efforts. If you plan a similar sort of collecting project, learn from our example and plan to have staff members whose job it is to develop similar relationships in the "analog world."

At the beginning of the project we envisioned developing an automatic collecting tool to gather born-digital materials through periodic crawls that would reduce staff time spent crawling the Internet by hand. Unfortunately, developing such a tool proved to be a bigger challenge for our team than expected.

We were able to build a tool allowing us to search the Flickr.com database by key words, date, or Flickr user id, and provide the option to scrape the image and its metadata, (but not the discussions

associated with the image). This uploader selected all images in the Flickr.com database tagged, for example, "Hurricane Katrina" or "Hurricane Rita," that had a creative commons license and placed them in a holding space within HDMB for a team member to vet and then add to the archive.

Alas for us, just because an image is tagged "Hurricane Katrina" does not mean that the image is actually of something to do with Hurricane Katrina. It could be of someone's cat named Katrina. This tool saved us time but still required our staff to vet each image before we made it public in our project.

Here again, we underestimated our staffing needs for the project. Had we budgeted for more programming time to build automated processes for scraping blogs, digital photos, online video, and podcasts, we would have been able to collect more born-digital material without requiring staff time to do it by hand. But even if we had budgeted for more programming time, we still would have to vet what we collected--a human intensive activity.

Finally, we misjudged the intensity of destruction of these events. While CHNM and UNO went to work soon after the hurricanes struck to begin collecting, we found that many residents, former residents, and volunteers were not ready to share. The destruction along the Gulf Coast was structural, institutional, and personal. For so many, even today, dealing with the aftermath has been difficult and they are still in the middle of the recovery process, because that summer hasn't ended yet. For others, the magnitude of the destruction of their lives and their communities was so great that they find it impossible to put into words.

The length of our grant was for two years, but for this particular disaster, two years turned out to be not enough time. If, like us, your project is devoted to collecting the digital record of a tragedy, plan to spend more time than you thought you would.

We share these experiences with you because we at CHNM encourage others to start digital memory banks and archives and to be active participants in saving and shaping the past. Collecting projects need not focus tragedy, either. Projects can easily celebrate local or national events.

When building a digital collecting project, understand that it exists in this in-between place we are calling Web 1.5. We can ask the public to participate in the collecting and saving of the past, and ask them to help organize it by creating tag clouds of their own making, while at the same time we can protect those contributions from third-party editing by those who may disagree with a perspective that is unlike their own. And, for all the potentialities of online collecting and democratizing the past, remember that any project still requires a great deal of analog hands-on history work.

Endnotes

---

[1] Tim O'Reilly coined the phrase in 2005. "What Is Web 2.0," at http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html, that refers to a second generation of web-based communities and hosted services, such as social-networking sites, blogs, and wikis that facilitate collaboration and sharing between users. A more recent article parses out the differences between Web 1.0 and 2.0 and creates rubrics for categorizing websites. See: Graham Cormode and Balachander Krishnamurthy, "Key differences between Web 1.0 and Web 2.0" First Monday, Vol. 13, No. 6 (2 June 2008): http://www.uic.edu/htbin/cgiwrap/bin/ojs/index.php/fm/article/viewArticle/2125/1972.

[2] Dan Cohen and Roy Rosenzweig, *Digital History: A Guide to Gathering, Preserving, and Presenting the Past on the Web* (Philadelphia: Penn Press, 2005) and also found online: http://chnm.gmu.edu/digitalhistory/.

[3] For a case study of building the September 11 Digital Archive see: Cohen and Rosenzweig, *Digital History.*

[4] Dublin Core Metadata Initiative: http://dublincore.org. For those using CHNM's Omeka, http://omeka.org, software package, Dublin Core metadata fields are provided for each digital item in the archive.

[5] Re-Captcha is a free program that many organizations use for spam control: http://recaptcha.net/. As of this writing, we are planning to add Re-Captcha to the contribution form to make managing the archive easier now that the grant period has ended.

[6] O'Reilly, "What is Web 2.0."

[7] The Commons on Flickr began with images from the Library of Congress and now showcases images from the Brooklyn Museum and Powerhouse Museum: http://www.flickr.com/commons. The Powerhouse Museum's collection database is searchable through traditional searches but also has extensive folksonomies: http://www.powerhousemuseum.com/collection/database/.