The born-digital deluge: documenting twenty-first century events.

Rachel Howard
University of Louisville, rachel.howard@louisville.edu

Heather Fox
University of Louisville, heather.fox@louisville.edu

Caroline Daniels
University of Louisville, c0dani01@louisville.edu

Follow this and additional works at: http://ir.library.louisville.edu/faculty

This article was originally published in Archival Issues, volume 33, issue 2, in 2011.

ThinkIR Citation
http://ir.library.louisville.edu/faculty/1

This Article is brought to you for free and open access by ThinkIR: The University of Louisville's Institutional Repository. It has been accepted for inclusion in Faculty Scholarship by an authorized administrator of ThinkIR: The University of Louisville's Institutional Repository. For more information, please contact thinkir@louisville.edu.
THE BORN-DIGITAL DELUGE: DOCUMENTING TWENTY-FIRST CENTURY EVENTS

BY RACHEL HOWARD, HEATHER FOX, AND CAROLINE DANIELS

ABSTRACT: With digital recording devices readily available to most people, events are documented and shared on-line in real time by the “person on the street.” The ease of creation and dissemination belies what archivists know will be the long-term challenges of organizing and preserving collections of born-digital information. While other processes require little modification, the inherent fragility of digital content and the ease of depositing files call for a substantial modification of established procedures. In this article, three University of Louisville archivists discuss their approach to the acquisition, copyright transfer, file naming, selection, description, and preservation of born-digital content donated by the local community to document Louisville’s August 2009 flood.

Introduction

On the morning of August 4, 2009, three hours of heavy rainfall backed up Louisville, Kentucky’s aging sewer system, resulting in flooding several feet deep in parts of the city. The city’s iconic racetrack, Churchill Downs; the University of Louisville’s Belknap and Health Sciences campuses; and the main branch of the Louisville Free Public Library sustained significant damage. Bystanders using cell phone and digital cameras documented this disaster on sites like Flickr and YouTube. Archivists at the University of Louisville decided to capture and preserve these born-digital materials, created by the community, to document this event and also to create the university’s first on-line collection comprised completely of born-digital images.

There is abundant literature about processing, describing, and organizing collections of digitized images for on-line access, but born-digital collections have received less attention. Aspects of this project were similar to previous work with analog and digitized materials, but the digital origins of this collection had a significant impact on
the project staff’s choices, particularly in the areas of acquisition and preservation. This study illustrates the challenges encountered in maintaining control over, providing access to, and preserving the born-digital images donated by the local community to document Louisville’s August 2009 flood.

**Literature Review**

While archival literature has addressed developing electronic records management programs and dealing with manuscript collections containing digital materials, it offers relatively little on the topics of acquiring and displaying digital materials collected from disparate sources to document a community event. Elizabeth Dow provides a valuable source of practical advice for accessioning collections containing digital materials, and Michael Forstrom provides a significant amount of detail for those seeking to address issues of authenticity. As Susan Davis notes, a large proportion of “collecting repositories” lack policies in important areas such as acquisition, preservation, and access to digital materials, even though they accept such materials. But given the ubiquity of devices capable of capturing images digitally, surprisingly little has been written to guide those seeking to acquire and provide access to born-digital collections. Michelle Caswell, a notable exception, discusses the value of images (including video) captured in the moment, even on cell phones, by individuals who are related only through witnessing particular events. Her work focuses on the events of 9/11 and on efforts to document human rights abuses. She provides valuable insight about appraisal, noting that the September 11 Digital Archive preserved all submissions—even those of such low quality as to be unintelligible. She suggests that the perceived lack of storage constraints for digital materials explains archivists’ increasing willingness to preserve all submissions, but she also notes that storage constraints do, in fact, exist. She argues that dubious, irrelevant, or otherwise suspect submissions may be rejected. Caswell also examines the ease with which digital materials may be altered; this means that the repository must either trust the donor or collect metadata that aids in determining authenticity. While donors may alter the contents of their files (analog as well as electronic), this possibility applies particularly to multiple donors, many of whom are essentially anonymous, rather than to a single donor with whom a repository may develop a relationship. She warns that file types received in such a project are varied, and the repository may need to adjust its systems in order to ingest material.

In recent years, spontaneous historic events, such as the 2007 Virginia Tech massacre, the terrorist attacks of September 11, 2001, and Hurricane Katrina in 2005 have sparked born-digital collections. A panel presentation entitled “Documenting Tragedy: Special Collections on the Front Line and on the Front Page” at the American Library Association Annual Conference in July, 2009, included Virginia Tech’s 4-16-07 Prevail Archives, a project also mentioned in a February 2008 Library Journal article. To commemorate 9/11, The Library of Congress established both a multimedia project documenting “person on the street” reactions through its American Folklife Center, and a Web archive “preserv[ing] the web expressions of individuals, groups, the press
and institutions in the United States and from around the world in the aftermath of
the attacks in the United States on September 11, 2001.\textsuperscript{11} The George Mason Center
for History and New Media, in partnership with local institutions, established the
September 11 Digital Archive\textsuperscript{12} and the Hurricane Digital Memory Bank: Collecting
and Preserving the Stories of Katrina and Rita,\textsuperscript{13} both of which use electronic media
to collect, preserve, and present those events.

Although these Web sites offer valuable documentation, they provide little informa-
tion that other repositories can use when developing projects to document their own
local events with born-digital materials. The issues inherent in providing on-line access
to digital images—particularly from disparate sources—are addressed specifically by
two of the archivists interviewed for Caswell’s article, who provide commentary but
little practical advice for other archivists. While some approaches for handling analog
materials may be adapted to digital materials with little significant modification, digital
collections present new and unique problems of preservation, authenticity, descrip-
tion, and copyright transfer. This article addresses what to expect—and what might
work—when documenting local events with digital materials.

\textbf{Soliciting Born-Digital Donations}

The University of Louisville Libraries’ collections include photos of local people,
places, and events taken by professional and amateur photographers. Material related to
the 1937 flood of the Ohio River, which affected Louisville and other towns along the
river from Illinois to Pennsylvania, have always been of particular interest to research-
ers. In 2007, community members commemorated the flood’s 70\textsuperscript{th} anniversary with a
new publication using images from the university’s special collections and archives.\textsuperscript{14}
The resulting publicity produced another round of community members’ donations of
photographs and newspaper clippings and a surge in reference requests on the topic.

When flash floods hit Louisville on August 4, 2009, awareness of the efforts of
Virginia Tech, the 9/11 Museum, and others to create commemorative collections,
coupled with the surge of interest in the 1937 flood, inspired the university’s digital
initiatives librarian to create a digital resource for the community that would support
Louisville’s ability to reflect on its own history—even before these events ceased to
be contemporary.

The digital initiatives librarian, in conjunction with the acting director of the Univer-
sity Archives and Records Center and the head of the main library’s Special Col-
lections department, decided to solicit born-digital donations from the community.
Regardless of the potentially poor quality of the material donated, the university was
most interested in the “person on the street” perspective. As the project staff expected,
community members’ images and comments captured the flood as it was experienced
by a broader population than was reflected in the news media. As with any other
acquisition or a retrospective digitization project, copyright, metadata, storage, and
preservation posed particular concerns. Given the ephemerality and insubstantiality
of the digital files (despite their apparent ubiquity), staff reevaluated and in some cases
reinvented established procedures.
The University of Louisville Libraries’ special collections units normally seek the transfer of copyright when acquiring analog materials. Copyright ownership provides significant advantages, including the right to create digital copies and post them on-line (barring other legal or ethical complications). The language of the standard deed of gift that accomplishes this transfer often intimidates donors, but can usually be explained and even modified. In this case, however, the relationship with donors proved to be more tenuous, and the normal exchange between donor and repository over copyright transfer and licensing sometimes broke down. The archivists anticipated that digital donors would be more interested in donating materials if the process were as easy as possible. The delivery of the images to the archives was certainly easier than with analog materials—donors just had to attach files to an E-mail and hit “send,” as opposed to packing and mailing prints. Electronic mail saved a visit to the archives, and the need to park and navigate an unfamiliar campus.

Project staff also suspected that potential donors—being technologically savvy enough to snap pictures of the flood with their cell phones—would be familiar with sites like Flickr and Facebook that enable individuals to post their own images on-line. This orientation toward sharing had the potential to make copyright enforcement difficult, since donors were likely already providing on-line access to their images elsewhere on-line. This contrasts with the way analog materials are shared, particularly those from amateur photographers, simply because sharing photographic prints requires more effort.

With all this in mind, the archivist modified the standard deed of gift, making it shorter and less intimidating to donors. The libraries’ legal expert approved the revised deed of gift (see Appendix A), which was then posted on-line. Project staff also created a form to collect essential metadata, and posted it with the deed of gift on the donation Web page so that this information would be submitted with the images (see Appendix B).

To make donating as easy as possible without overloading the limited amount of space available in an individual university E-mail account, project staff used a service account, (an E-mail account with shared access allowing central management and storage of electronic communications or data). Next, they drafted a press release, which the campus communications and marketing department distributed. This was all accomplished within approximately a week of the flood.

A day after the press release went out, a news reporter at the local National Public Radio affiliate interviewed the digital initiatives librarian, and colleagues reported seeing the E-mail address included on local news broadcasts. Almost immediately, born-digital images captured with cameras and camera phones began to arrive via E-mail, compact discs, and external hard drives. Thirty-three donors deposited 1,228 digital images and 14 digital videos. Six images were downloaded from Facebook and 184 harvested from Flickr with permission. Donors included university students and staff, faculty, and community members.

As anticipated, donors were less interested in filling out the deed of gift and supplying metadata, which required staff to do considerable work to follow up and ultimately secure permissions from 30 of the 33 donors. While most archives have experience with analog materials that have been “left on the doorstep” without affording the
opportunity to collect a deed of gift, a greater than usual proportion of these digital donations required significant follow-up in order to obtain even a simplified license. As the project progressed, the staff determined that it was not necessary for the university to hold copyright in order to accomplish the goals of preserving the documentation of this event and making it available on-line and on-site in the archives. In hindsight, the options of a non-exclusive Creative Commons license\(^\text{16}\) or an electronic signature on a PDF should have been promoted to potential donors. The staff did accept Creative Commons licenses, but making this option more obvious at the outset might have simplified the process for all concerned. Donors who submitted their files electronically might also have been more likely to submit their permission forms electronically. Subsequent analysis also indicated that the archives could have taken advantage of the news reporting clause of the Fair Use doctrine.\(^\text{17}\) This clause identifies “news reporting” as a fair use of copyrighted material, and arguably the on-line digital collection was intended to report on the events of August 4, 2009.

Most donors preferred sending files by E-mail, which raised another unforeseen difficulty. Although the service account itself had ample space, the university’s system blocked E-mails with attachments over a certain file size. Some donors sent their files in multiple E-mails to avoid the file size limitations. Other donations may have been missed entirely because the system did not notify either sender or recipient when it blocked oversize attachments.

To process the digital donations, project staff established a server directory for the donations and created a subdirectory for each donor. Donors who had not yet granted permission were flagged in the folder title so that project staff would wait to create metadata for them. If the donation came via E-mail, a copy of the E-mail was saved in the folder, along with any other electronic documentation (e.g., the metadata form). Paper copies of signed permission forms were retained and filed with other donor files. This system worked well at the folder level to keep track of groups of files from individual donors. At the item level, however, the bulk of the donations were auto-numbered.\(^\text{18}\) Folder level control sufficed during the gathering phase, but individual files required more organization.

After considering the need to preserve original filenames, project staff concluded that the requirement for control trumped any purpose for maintaining the original filenames. Keeping track of over one thousand auto-numbered digital images depicting flooded streets and basements necessitated renaming. It would have been cost-prohibitive to manually rename each of the files, so the donors’ names were automatically appended to the beginning of each of their files via free downloadable software called \textit{Bulk Rename Utility}.\(^\text{19}\) Once the files had been organized, project staff began selecting the images to be described and posted on-line.

\textit{Selection for On-line Access}

Although the libraries made a commitment to long-term digital preservation of all donations, early on the project staff decided to post only selected items on-line, in order to minimize the labor of item-level metadata creation while also minimizing
duplication. Staff employed sampling, a method that has long been applied to analog materials, to reduce the volume of content to be retained. Kepley describes two types of sampling, subjective and statistical. Both have been employed with large twentieth-century collections. Sampling allows the archivist to limit the resources needed to store and preserve such collections by reducing their size. Born-digital images pose another dilemma. They require an intensive expenditure of resources on description, usually item-level description, if they are to be posted on-line. The staff therefore chose to provide access to “samples” of the collection in the digital library, while providing an on-line finding aid describing the collection as a whole. One-sixth of the collection, or 212 images and three video files, were selected for the on-line collection, but all images are accessible on-site to a user upon request.

Project staff based this sampling on a number of criteria. They aimed for the widest possible geographic representation and therefore included photos received from every location. Most donations depicted the university area and were donated by staff and students. This was no surprise, considering the university area was one of the hardest hit in the city. Of the small number of donations received from points farther west and south, more were included per donor than from better-represented areas.

Due to high-capacity storage and photographers’ ability to shoot multiple photos, donations often included multiple copies of virtually the same image. Donors would dump everything onto a flash drive or CD and give it all to the archives. In such instances, project staff chose the version that included the most information (e.g., visible signage). Finally, the most dramatic sites garnered the most attention and thus the most donations. For example, nine of the 33 donors contributed photographs depicting the flooding of the building that houses the admissions office. Significant time was required to consolidate these images and select the best from among them.

Description

Once selections had been made, the process of description began. A finding aid would describe the entire collection, with item-level metadata created for the items selected for on-line presentation. While large-scale digitization projects are currently underway that provide minimal (if any) item-level metadata, these appear to be predominantly organic collections that can at least be described on a folder level. Any effort to create “folders” based on anything other than the photographer’s name would have been arbitrary. Furthermore, the opportunity to discover connections between materials that documented the same locales and similar events over time required item-level metadata. The University of Louisville employs the Dublin Core-based CONTENTdm digital media management software to provide access to its digital collections, and the librarians have compiled a data dictionary defining mandatory and optional metadata elements. According to the dictionary, the source field is not a mandatory element for born-digital materials, but project staff used this field to include detailed information about the capture device, when available. They also took advantage of an optional “abstract” qualifier to the Description element to provide contextual information received on some of the acquisition forms and through communication with donors. This field
was labeled “Photographer’s comments” in the on-line display and the decision was explained in the “About the collection” portion of the Web site. Project staff also gathered and incorporated metadata from blogs and social networking sites to enhance description. On his blog, one donor remembered traffic lights flashing over a completely flooded intersection. He noted that the picture does not convey the strangeness of what was happening. Another donor submitted a video documenting this phenomenon, and the description fields for each item cross-reference one another.

As noted previously, files not selected for item-level description are available to researchers on-site. They are foldered by donor but are not described in greater detail. This follows the archives’ standard practice of providing folder-level rather than item-level control of collections.

**Digital Preservation**

These born-digital collections are backed up on secure file servers, or caches, at six geographically-distributed locations thanks to the libraries’ membership in the MetaArchive Cooperative. The MetaArchive Cooperative’s distributed digital preservation network uses a technical framework based on the open source LOCKSS (Lots of Copies Keep Stuff Safe) software not merely to back up the files, but also to check constantly each file against cached copies and provide repairs whenever necessary.

The MetaArchive assumes the existence of digital files over which the contributing institution has stewardship. Project staff quickly realized that they also wished to preserve related Web-based content over which they had no curatorial control, such as the aforementioned student’s blog post containing photos and description of his experience. While this material would not become part of the archives’ collections, it was similar in content to some of the documentation created during the 1937 flood: the homemade newsletters and journals held by the libraries’ special collections units. The archives did not have the resources to harvest and preserve Web-based content on their own, but saw value in its long-term survival. Archive-It (a subscription service from Internet Archive) provided a solution. Their “Spontaneous Events” collection includes archived Web sites relating to the shootings at Virginia Tech; the Chilean and Haitian earthquakes of 2010; and Hurricane Katrina, among others.

Archive-It was the only service available that had the capacity to harvest and preserve Web content, but for an archives on a small budget, their fees posed a challenge; a long-term subscription was not feasible. Using a one-month Archive-It subscription, staff preserved blogs, Facebook groups, Flickr sets, state of emergency information, and news reporting related to the flood. This meant that for one month, Archive-It harvested content relating to the flood of 2009, and they will preserve that content indefinitely. Unfortunately, YouTube videos could not be preserved: YouTube includes a robots.txt file in their site, which signals to Web crawlers seeking content that they should exclude this site from their harvest. Archive-It later worked out a deal with YouTube to allow harvesting and preservation, but by that time, the libraries’ subscription had ended.
Conclusion

This experience, for which the archives staff was not completely prepared, contained important lessons that can be applied to future work with born-digital manuscript collections. While many of the processes involved were identical to those applied to analog collections, the born-digital nature of these materials required a different approach to acquisition and preservation, in particular. Arrangement and description were handled in much the same fashion as analog collections are managed. Future collections may be more complex and include series that contain both analog and born-digital content. However, the concepts are the same: the archives staff considered the content of the images and their origins, as opposed to their file format, when determining how to arrange the files.

The project confirmed the staff’s expectations about donors of born-digital content. In this case, donors were more likely to complete forms that were simple, required little or no explanation, and could be submitted electronically. It was also easier for the archives to obtain a license to display and preserve the content in some cases, and this was in fact acceptable for the purposes of this collection. These donors were also likely to submit anything, or nothing, in terms of metadata, including filenames. While digital files can be accompanied by significantly greater metadata of various kinds, the bulk of the information received for this project pertained to the capture device itself. Technical metadata are useful for preservation but hold little descriptive value; descriptive metadata, which are essential for access, are unlikely to come from the donor. In this regard, these born-digital materials differ little from their analog counterparts—and the ready availability of a form that can be submitted digitally did not increase their metadata submissions.

If archivists are interested in collecting documentation of local events, they must be prepared for born-digital materials that arrive in a variety of decidedly non-archival formats and via a variety of conduits. The archivists who undertook this project remain convinced that born-digital materials, however “messy” they may be, are worth preserving to document events that typical community members experience. The long-term value of the materials documenting the flood of August 4, 2009 will be determined after these events recede into history, but the collection has already served a more immediate purpose: the first reference request relating to this on-line collection came from a government agency looking for material to document its report on the flood.
ABOUT THE AUTHORS: Rachel Howard is the digital initiatives librarian at the University of Louisville Libraries. She holds an M.L.I.S. from the University of Washington and a B.A. in history from the University of Notre Dame. She has worked with digital collections for libraries, museums, and archives, including the Library of Congress, Smithsonian Institution, Cornell University, University of Washington, and Museum of History and Industry (Seattle).

Heather Fox received her M.S.L.S. from the University of Kentucky and has worked as an archivist at the Filson Historical Society, as a data wrangler at the University of Louisville Archives and Records Center, and as a project archivist for the Speed Art Museum. She currently serves as project archivist at the Kentucky Historical Society. She is a board member of the Kentucky Council on Archives and the Kentucky State Historical Records Advisory Board.

Caroline Daniels is university archivist, director of the University Archives and Records Center, and co-director of the Oral History Center at the University of Louisville. Previously, she worked in Tufts University’s Digital Collections and Archives. She holds an M.S. in library and information science with a concentration in archives management from Simmons College, and an Ed.M. from the Harvard University Graduate School of Education.

NOTES

1. Although many readers may be familiar with the terms, we feel it is important to highlight the distinction between digitized and born-digital. According to the SAA on-line glossary of archival terminology, “born-digital information is distinguished from digitized, the latter describing a document created on paper that has been scanned (and possibly transformed into character data using OCR). A document created using a word processor may be described as born digital.” (Society of American Archivists, Glossary of Archival and Records Terminology, <http://www2.archivists.org/glossary/terms/b/born-digital> (5 August 2011).) An image created using a digital capture device may also be described as born-digital.

2. For example, we know donations to the August 2009 Flood Collection came from the following devices: iPhone 3G, Nikon Coolpix L18, NIKON D200, Canon PowerShot G6, BlackBerry 8310, Sony Handycam, and Canon EOS Digital Rebel Xti. Not all devices were identifiable.


15. A Web page was created to provide information and access to the deed of gift and information form, available at <http://louisville.edu/library/giving/2009Flood.html> (5 August 2011).
18. For example, file number 5855_1112192331375_520726375_2372903_1982696_n.jpg depicted the flooded intersection of Third Street and Kingston Avenue.
28. According to Webopedia, “Robots.txt is common name of a text file that is uploaded to a Web site’s root directory and linked in the html code of the Web site. The robots.txt file is used to provide instructions about the Web site to Web robots and spiders. Web authors can use robots.txt to keep cooperating Web robots from accessing all or parts of a Web site that you want to keep private.” See “Robots.txt,” *Webopedia*, <http://www.webopedia.com/TERM/R/robots_dot_txt.html> (5 August 2011).
Appendix A—Deed of Gift

DEED OF GIFT

I, ____________________________, hereby donate the materials described below to the University of Louisville Libraries as an unrestricted gift, to be administered in accordance with its established policies. I assign and transfer to the University of Louisville Libraries such rights and copyrights as I possess in these materials. Further, I agree that this material may be made available to researchers without restriction.

DESCRIPTION OF MATERIALS

Please check the following
Items not retained by the University of Louisville Libraries will be
☐ Discarded or disposed of as deemed fit by the staff.
☐ Returned to donor.

Donor information:
☐ I grant permission to use my name as the donor for exhibits, description and publicity.
☐ I wish to remain anonymous.

Signed: ____________________________ Date: ______________

The University of Louisville Libraries hereby accepts the material described above and under the conditions specified.

Signed: ____________________________ Date: ______________
# Appendix B—Photo Information

## August 4, 2009 Flood - Photo Information

Please fill in the fields below. You can copy and paste the form if you are submitting more than one image. If they are sufficiently similar (same place/time), you can describe them all at once.

<table>
<thead>
<tr>
<th>Field name(s):</th>
<th>Name of digital file. Example: IMG007.jpg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Photographer:</td>
<td>Type name of photographer here.</td>
</tr>
<tr>
<td>Title:</td>
<td>Provide title for image(s).</td>
</tr>
<tr>
<td>Location:</td>
<td>Provide as much detail as possible about where the photo was taken: address, intersection, or neighborhood.</td>
</tr>
<tr>
<td>Date:</td>
<td>Date photo was taken.</td>
</tr>
<tr>
<td>Time of day:</td>
<td>Time photo was taken, if known.</td>
</tr>
<tr>
<td>Other:</td>
<td>Any other details you’d like to provide.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Field name(s):</th>
<th>Name of digital file. Example: IMG007.jpg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Photographer:</td>
<td>Type name of photographer here.</td>
</tr>
<tr>
<td>Title:</td>
<td>Provide title for image(s).</td>
</tr>
<tr>
<td>Location:</td>
<td>Provide as much detail as possible about where the photo was taken: address, intersection, or neighborhood.</td>
</tr>
<tr>
<td>Date:</td>
<td>Date photo was taken.</td>
</tr>
<tr>
<td>Time of day:</td>
<td>Time photo was taken, if known.</td>
</tr>
<tr>
<td>Other:</td>
<td>Any other details you’d like to provide.</td>
</tr>
</tbody>
</table>