The Electronic Media Review

Electronic Media Group

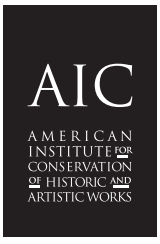
Volume Three 2015

Papers presented at the Electronic Media Group session of the 41st AIC Annual Meeting, Indianapolis, Indiana, 2013, and the 42nd AIC Annual Meeting, San Francisco, California, 2014.

Jeffery Warda and Briana Feston-Brunet, Managing Editors

Edited by Helen Bailey, Briana Feston-Brunet, Karen Pavelka, and Jeffrey Warda

Layout by Amber Hares
(Original design by Jon Rosenthal, JonRosenthalDesign.com)
Typeset in Trade Gothic LT and Myriad Pro



American Institute for Conservation of Historic and Artistic Works
Washington DC

The Electronic Media Review was published once every two years in print format by the Electronic Media Group (EMG), a specialty group of the American Institute for Conservation of Historic and Artistic Works (AIC), until 2013 and published online only thereafter. *The Electronic Media Review* is distributed as a benefit to members of EMG who held membership during the year of the issue. Additional copies or back issues are available from AIC. All correspondence concerning subscriptions, membership, back issues, and address changes should be addressed to:

American Institute for Conservation of Historic and Artistic Works
727 15th Street NW, Ste. 500
Washington, DC 20005
info@conservation-us.org
http://www.conservation-us.org

The Electronic Media Review is a non-juried publication. Papers presented at the EMG session of the AIC Annual Meeting are selected by committee based on abstracts. After presentation, authors have the opportunity to revise their papers before submitting them for publication in *The Electronic Media Review*. There is no further selection review of these papers. Independent submissions are published at the discretion of the EMG Publications Committee. Authors are responsible for the content and accuracy of their submissions and for the methods and materials they present. Publication in *The Electronic Media Review* does not constitute official statements or endorsement by the EMG or by the AIC.

Presented at the Electronic Media Group Session, AIC 42nd Annual Meeting, May 28–31, 2014, San Francisco, CA.

FILLING IN THE GAPS: FINDING YOUR WAY TO CONSCIENTIOUS CURATION AND PRESERVATION OF BORN DIGITAL COLLECTIONS AND OBJECTS

JASON EVANS GROTH

ABSTRACT

The Online Computer Library Center has released several reports since 2012 that attempt to demystify born digital for archivists. The positive response offers more proof that the world of archives is beginning to firmly face the challenge of how to deal with born digital collections and objects. The reports are just the starting point, however, and while they are incredibly helpful and provide an overview of mandatory tools for born digital curation and preservation, they are written for a broad audience and, because of that, do not address some of the finer points of this potentially confusing, time consuming, and, thus, often passed-over work. This article gives an example of a born digital workflow at one academic institution, and describes how it is very likely that no one report or tool will answer all of an institution's born digital questions, but that by being flexible, drawing from many sources, and understanding institutional requirements, it is possible to create a viable born digital process that is scoped appropriately, leading to quicker preservation of and access to materials.

INTRODUCTION

In an effort to help provide a clear path for archivists to begin work in a challenging and overwhelming domain the Online Computer Library Center (OCLC) began publishing a series of reports under the title "Demystifying Born Digital" in 2012.

The release of these reports and the active and positive response to them offer more proof that the archive and conservation world is not on the brink of, but is, rather, firmly facing the challenge of how to deal with born digital collections and objects. This comes as no surprise to many who have advocated for such awareness, but it is a reminder that plans need to be made immediately to start appropriately caring for this material for the long-term, while also making it accessible.

Media archivists predict that legacy video and audio formats provide preservationists a 10–15 year window to safely and completely migrate content to a more manageable form. Data—floppy disks, hard disks, USB drives, CD-Rs, etc.—is at least as volatile, if not more, than heartier media formats that came before it. Film, for example, can be examined and the content understood without a working projector by simply holding it to the light. While this is not an ideal situation for examining content, it can provide clues until an appropriate projector or other playback device is found. A floppy disk, hard drive, or magnetic card hides its content, like a magnetic cassette tape might. But unlike a magnetic cassette tape, a working 5.25" floppy drive that interfaces easily with current equipment is not a ubiquitous sight at thrift stores, especially considering that, on modern computers, an additional controller card is needed to read the data. And, unlike a cassette deck, 5.25" floppy drives are no longer being made en masse, and new stock is often prohibitively expensive.

Knowing that the clock is ticking can be motivating, and the OCLC reports attempt to foster motivation into a plan. They, however, are just the starting point, and overworked archivists who are, perhaps, not keen on learning command line operations at a born digital workstation, for example, may not put these quickly degrading, sometimes already obsolete and certainly un- or under-used collections at the top of their list to process appropriately and quickly. And even ubiquitous born digital materials—USB flash drives, external hard drives, DVD-Rs—are overwhelming, often disorganized, and erroneously thought to last forever if just kept in the box with the papers they came with, or on the shelf until the institution figures out a plan to deal with them. The OCLC reports outline tools and techniques for minimally viable born digital curation and preservation but, as their intended audience is a broad one, they cannot get as granular as an overworked and overwhelmed archivist may wish.

Additionally, the plethora of options may seem overwhelming to an archivist who is largely unfamiliar with the landscape, and while the reports are comprehensive, they are not case studies. They do, however, present the main concerns in a cohesive, understandable, and important way. In a way, the reports are like the Born Digital Processing "food pyramid," the backbone of the needs of a world that wants access to these items. The truth is, though, that even the food pyramid needs recipes to make it understandable and usable at a family level, and the same is true for the born digital reports. Recipes—case studies, testimonials, and reports on results—are imperative for archivists who are embarking upon this daunting frontier for the first time. For the majority of us there is no single tool or solution for born digital processing and access, and likely there never will be because of the diversity of our institutions. Understanding that this gray area is powerful and exciting is a step that general reports can only hint at; specific examples of active communication across institutions that are working on similar challenges are good first steps.

BORN DIGITAL PROGRAM, TAKE 2

I began my two years of work at the North Carolina State University (NCSU) Libraries Special Collections Research Center (SCRC) in August of 2013 as part of a NCSU Libraries Fellowship strategic initiative to start a born digital curation program. As a Fellow, I split my time as part of a home department (User Experience) and the SCRC. That means I got to devote a full twenty hours a week to born digital, but was still not full time in the SCRC.

Prior to my arrival, the SCRC had attempted to implement a born digital curation program, but was thwarted

by the same thing that thwarts so many: a lack of support for the chosen technology. The IT department could not devote the time and resources we thought we needed, so plans that involved IT building a workstation with our guidance and their expertise were hatched. This allows them to have a say in and an understanding of our technical requirements, but also allows us to have more control and flexibility in testing tools. Using sources like the OCLC reports as guidance, staff purchased write blockers, external drives, and a dedicated computer.

Those previous born digital program creation attempts were very well-documented in-house, including helpful descriptions about using open source tools, their strengths and weaknesses in our institution's context, and suggestions for more robust usage in the future. So the failure to implement a program actually led to success in that it laid the groundwork for contextual understanding of what a born digital project should look like at NCSU Libraries. This was not the full picture, however, as our institution's needs regarding born digital had evolved.

At this stage the commitment was set: I was hired and we had equipment. And while those may seem to be the main ingredients for a successful project, that, of course, is not the whole story. We needed our SCRC staff to buy-in, too, and we began building that buy-in through the implementation of a Born Digital group and then, even more importantly, learning how to talk about what we were doing with people who were not necessarily well-versed in the lingo of born digital. While words like write blocker and ExifTool make sense to people who are actively working on these problems, curators in other parts of the department may not. Learning how to communicate what these things actually do in a standard archival way was, and remains to be, important, as it enables us to approach institutional requirements with a common understanding. When others are enthusiastic and understand what these tools do, it helps the digital archivist scope born digital workflows appropriately for their particular institution, and which are in line with existing workflows.

Once effective communication levels are reached, an issue that often plagues similar technical work in more traditional library environments—*imposter syndrome*—is mitigated. Since there are very few specific recipes for this kind of work, it is a challenge to even decide what kind of questions need to be asked to reach the answers we need, or even think we need. Effective communication—and practicing this communication both locally and through online user groups, conferences, presentations, etc.—leads to a greater understanding of both the problems facing institutions regarding born digital and the possible solutions.

For this project we divided the two years into quadrants. The first, and most talked about, was processing and ingest. In the course of working inside of that quadrant we made the somewhat obvious discovery that access—our main goal at NCSU Libraries—was inextricably linked to processing and ingest. Thus, imagining how future access might work completely colored our way through the first quadrant. Other quadrants—arrangement and description, and sustainability—have changed throughout the course of the project, and our adapting to unforeseen needs shone light on how flexible an institution needs to be when implementing such a program.

What drives us is the need to free the content from its media "jail" in the most archival-appropriate way. We consider digital forensics to be the way: we gather all of the evidence we can, let forensics tools do the hard work of translating the bits into information we can use, and we do all of this so that we can, potentially, allow researchers unprecedented access to material. When an archive receives a donation of a box of papers, it is highly likely that the papers inside have either been arranged for the donation or that they have been moved from a different organizational container into that box. When we receive a hard drive straight out of a work terminal, however, we have the chance—if the creator hasn't changed it purposefully for the donation—to see inside the organizational mind of the person who gave us the drive and, in turn, let researchers see it, analyze it, and work with

it in ways we only wish we could do with other kinds of physical items. This involves providing access to disk images, for example, and that depends on the donor agreement. It is possible to plan for every step of this process so long as buy-in exists by all members of the team, enthusiasm regarding access is maintained, and an understanding of the kind of unprecedented, easy-to-get information afforded to us by digital objects is attained.

We came to these conclusions by reading general reports and then embarking upon environmental scans, striking up conversations at conferences, keeping abreast of professional work, but, most importantly, comparing all of this against what we thought we actually needed at our institution. Confidently approaching born digital with our institution's requirements makes the whole process easier to both sustain and grow. While this work is being done, we have been documenting it so that our future selves—whether actually us, or the librarians who will have taken our positions—can understand what it is we left them, and in what form. That is our job as librarians, after all: creating a network of information that can be accessed later so that understanding passes through time. Be transparent and foster repeatability. It's our only chance.

RELEASE THE KRAKEN (OUR BORN DIGITAL WORKSTATION)

Born digital is not an IT problem; it is a universal problem with IT solutions. One of the most important things you can do to ensure support for your program is to run it yourself with IT as background support. In other words, stop at nothing to get administrator privileges on your machine. This is important because your institution may find one free tool to be better suited than another for whatever your requirements may be. You can only properly test this if you can download and install freely. Your time is limited: be in charge. If you run into issues, work with IT to have them maintain an image of the machine as they delivered it to you. When you tell them that you ran into some trouble and all you need is to reinstall the image, they'll thank you for being organized enough to know the issue and that they don't have to learn new software to support your particular stack.

At NCSU Libraries we have purchased a relatively low-powered (for now) Dell running Windows 7, outfitted with free tools. We call it The Kraken because of all the tentacles running to and from it, those being the connections between the drives and write blockers when everything is plugged in (fig. 1). Those tentacles, in particular, are a WiebeTech Ultradock, a WiebeTech USB writeblocker,

Fig. 1. The Kraken: NCSU's Born Digital Workstation.

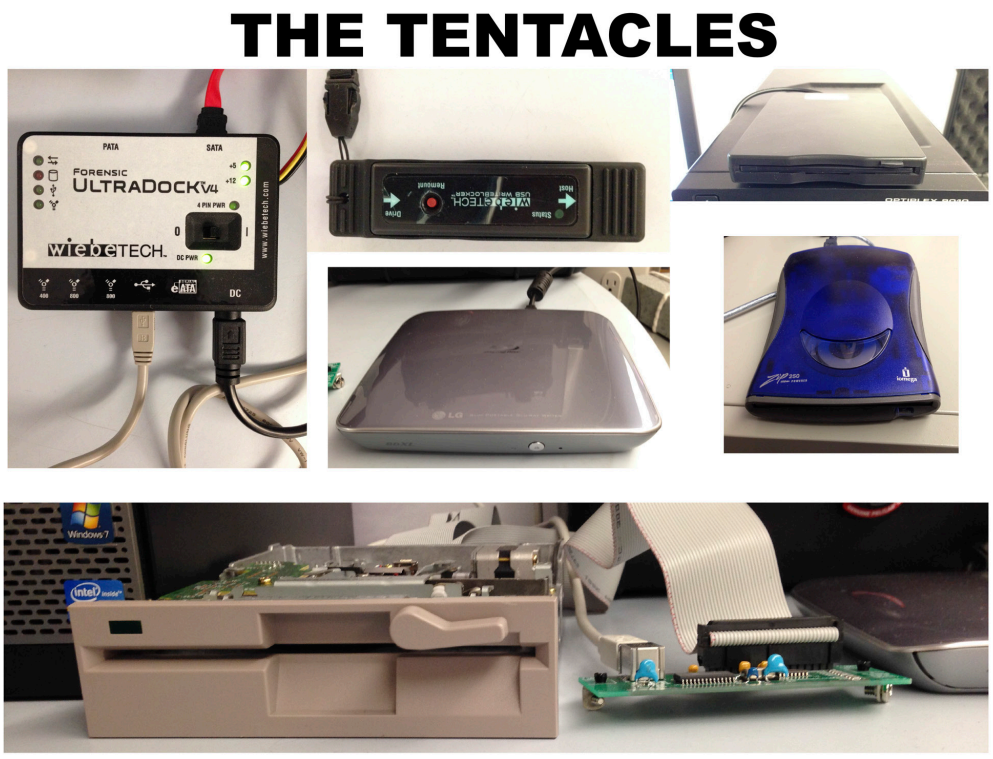


Fig. 2. The Tentacles: the main hardware used with The Kraken to process born digital objects.

a Teac 3.5" USB disk drive, a vintage Teac 5.25" drive with an FC5025 disk controller, an external CD/DVD/Blu-Ray drive, and a ZIP drive (fig. 2). We purchased these devices because they can take care of the majority of our backlogged digital objects. Your hardware needs may vary, however, as will ours. We have also worked with regional universities to begin setting up hardware sharing so that should we come across backlog anomalies, we do not have to purchase new hardware, but rather extract the data there and process it at our own institution.

We have created dynamic documentation that leads our processors through the steps we developed at NCSU. We used Google Forms to prototype our Choose-Your-Own-Adventure style form, which we call DAEV (Digital Assets of Enduring Value), which responds to answers with particular sets of instructions depending on media, donor agreement, etc. This follows our goal of making something that seems very complicated that much easier; despite the processor needing to use tools that are, perhaps, unfamiliar, the steps are clearly written and are led as if through a standard online survey. By the end of the form, disk images, and logical copies are made, metadata is extracted, personal and private information is searched for, viruses are scanned for, a readme file is generated to help future us decipher what's in our package, finding aids are updated, and all of the data is sent to storage with fixity checks put into place.

THE WORKFLOW

There is not a single recipe for everything, and it took trying out different media to figure out what hardware and software needed to be used in concert with that media to reach our goals. Our workflow is, thusly, determined by the media that needs to be processed. After DAEV knows what one is working on, it loads the particular path that is required to get the object processed correctly (fig. 3). Once an item has an accession number, it is given to the born digital processor. That person uses the appropriate "tentacle," attached to a write blocker (or using an appropriate write-blocking method) to connect to The Kraken. FTK Imager (the free version) is then used in

THE OBJECTS

Fig. 3. A buffet of content "imprisoned" on obsolete or fragile media.

most cases to make a disk image of the item (we use RAW because RAW comes to close to ensuring that the created file can be opened by widely supported applications versus forensics specific file formats). Once the image is created, we generate a logical copy of the image, for future access reasons. We then use many tools packaged in BitCurator to extract metadata, scan for viruses, and look for personal and private information. Metadata is created automatically, as part of the DAEV form, by the free tools we use for processing. We also take pictures of the object to provide more human readable metadata.

ACCESS

In all cases we generate RAW image files so that we can run the BitCurator reporting tool over it easily. In some cases we do not retain the image file, because that means storing empty space. We determine these cases based on how the object was given to us. If it's a hard drive that was simply removed from a working terminal, then we will retain the image. If it's an external hard drive that only has specific files copied to it for the sake of donation, we will retain the logical copies, since there will (likely) not be dependent information or hidden files that may shed light on the creative process.

We envision, eventually, sharing a logical copy of the data and the extracted metadata with our Digital Libraries Initiatives department who can use this information to generate a virtual disk browsing environment, linkable through our finding aid for the collection that the digital object came from. In other words, in the finding aid under digital objects, we have the chance to, for example, write "internal hard drive" and provide both a link to a sortable .CSV for local searching, in addition to a link that allows a researcher to browse the hard drive on their terminal, replicated exactly as the hard drive came in. While, in most cases, one could not actually access the files, they can see the file and folder lists and, by hovering over the filenames, could also see file metadata. This can potentially save time for the researcher and for our department, as the researcher can either identify specific files they wish to see ahead of time or find out that, perhaps, none of the files add to their work. By linking this virtual browsing environment to a logical copy of the files, previewing media files and text files would also be possible (subject to the donor agreement).

We have decided to let the bits describe themselves. In other words, aside from linking a virtual browsing environment to the finding aid (and potentially running word list programs to determine what might be on the disks, based on the lists they create, and then providing this in the finding aid), we do not plan on rearranging the files or describing the disks. The disks are contextualized within the collection already, and increasing use of automated tools helps achieve higher processing rates,

while reducing the time it takes to make these files available to patrons. If a patron wants to see the whole disk image, we will first ensure that this is in concert with our donor agreement, and then make arrangements for them to come to our reading room and use a non-networked laptop with the image mounted and indexed for easy searching. We will do this on a Mac, and utilize the already-familiar Spotlight tool for searching. The browsing environment allows them the chance to see the disk contents. All in all, we've decided to use tools to make arrangement and description easier (because the bits can do it), and then provide other tools to let the patron make the choices about what they want to see as easily as we can.

CONCLUSION: DON'T LET THE PROBLEM OVERWHELM YOU

Again, this is a universal problem with technical solutions. Prepare requirements with the help of colleagues and find the tools that fit them. Do this by talking to others working on similar challenges, who are at similar institutions, and who are willing to share their discoveries. Don't let the possibility that this work is challenging prevent you from even beginning it. Learn from the failure of others (and yourself), and document every step to make sure you don't repeat what doesn't work, while ensuring that what does work gets repeated. Use your favorite search engine to wade through previous discussions about problems that you're also encountering, and do not hesitate to ask questions. Workshops, articles (like this one), and reports help, but they can never address problems specific to your institution. It is possible to begin this kind of work with something as simple as a surplus laptop and an optical drive or a multi-thousand dollar setup; again, it depends on your institution, but both scenarios achieve results. Make sure you have administrator privileges. Remain flexible—there is no single solution that will work for everyone. Don't be afraid to make mistakes, because the biggest mistake of all would be to let this potentially very important information be imprisoned forever in its soon-to-be, if not already, obsolete media prison.

REFERENCES

Barrera-Gomez, J., and R. Erway. 2013. Walk this way: Detailed steps for transferring born-digital content from media you can read in-house. OCLC Research. http://www.oclc.org/content/dam/research/publications/library/2013/2013- 02.pdf (accessed 9/14/14).

Erway, R., B. Goldman, and M. McKinley. 2014. Agreement elements for outsourcing transfer of born digital content. OCLC Research. http://www.oclc.org/content/dam/research/publications/library/2014/oclcresearch-borndigital-content-transfer-2014.pdf (accessed 9/14/14).

Erway, R. 2012. Swatting the long tail of digital media: A call for collaboration. OCLC Research. http://www.oclc.org/research/publications/library/2012/2012-08.pdf (accessed 9/14/14).

Erway, R. You've got to walk before you can run: First steps for managing born-digital content received on physical media. OCLC Research. http://www.oclc.org/content/dam/research/publications/library/2012/2012-06.pdf (accessed 9/14/14).

Jason Groth
North Carolina State University