Eric Hoyt

Bootstrapping a Digital Project? 5 Things to Consider

When I started publishing articles about the intersections of law and Hollywood, I found myself fielding questions for which I was ill prepared. Does ABC qualify as a fair use under copyright law? Can I deduct XYZ on my taxes? I had to explain that I didn't have a law degree, and I couldn't provide any official legal advice. All I could offer was my basic understanding of the law.

Recently, I've been getting less law-related queries and more related to building digital projects and digital archives. Many humanities scholars want to extend their work into the digital sphere through curating digital collections, creating media-rich versions of their arguments, and/or developing sofatware tools for educational use. In everyday parlance, the label of "digital archive" is sometimes applied to these initiatives—particularly if they contain a collection of digitized texts or videos that others can reuse. However, few of these initiatives meet the preservation standards and uniqueness definition of an archive in the traditional sense of the word. We're on safer ground if we use the term "digital project," which is less descriptive but still helps to distinguish what we're talking about from digital repositories, such as those archives using DSpace or Fedora.

Numerous open source frameworks are now available for the construction of scholarly digital projects. The options range from simple blog software (Wordpress, http://wordpress.org), to more sophisticated platforms for scholarly publishing and digital collections (Scalar, http://scalar.usc.edu/ and Omeka, http://omeka.org/), all the way up to integrated development environments for those brave enough to create their own PHP, Ruby on Rails, and Javascript apps (Aptana Studio, http://www.aptana.com/). The platforms are there waiting for you. In fact, they aren't just waiting—they are improving regularly, thanks to the hard work and commitment of open source developers.

Still, the process of designing and executing a digital project can seem daunting. I'm writing this essay from the standpoint of a humanities scholar who took the plunge a couple of years ago. As co-director of the Media History Digital Library (http://mediahistoryproject.org), I've had the opportunity to help build a digital collection of historic film and media publications that now exceeds half a million pages. The Media History Digital Library (MHDL) digitizes out-of-copyright periodicals, such as *Photoplay* and *The Film Daily*, and makes them openly accessible for public use. Since launching our website in September 2011, our digital volumes have been downloaded over 100,000 times and "streamed" online through the BookReader app many more times.

I've been fortunate to collaborate closely on the Media History Digital Library with David Pierce, the project's founder and director, and Wendy Hagenmaier, the MHDL's digital archivist. David and Wendy both have professional experience working at institutional libraries. Wendy recently completed a masters of science degree in information studies, and she now works as a bona fide digital collections archivist at the Georgia Institute of Technology Library. In contrast, I lack any formal training as a librarian or archivist. Aside from a part-time job one summer in college, I've never worked at a library in any capacity other than as a researcher and user.

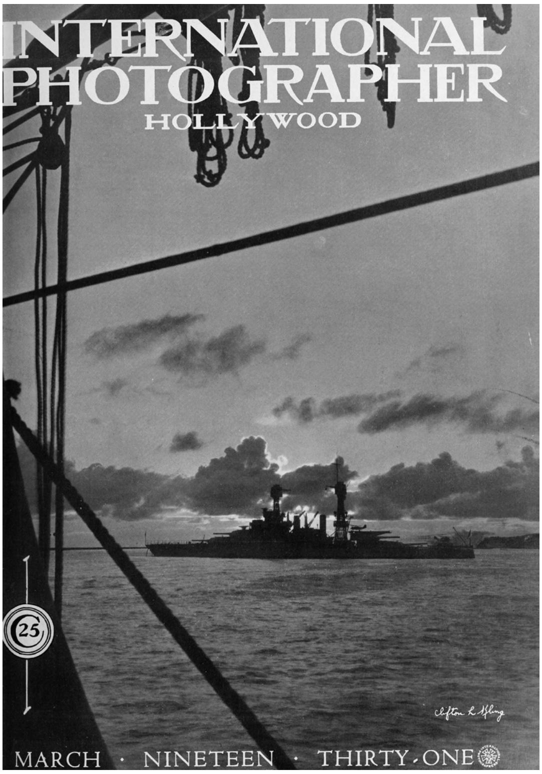

Figure 1. International Photographer is one of the historic magazines digitized by the Media History Digital Library, which has found an audience among academics, educators, and fans. Photo courtesy of the Media History Digital Library.

I cannot offer the perspective of an expert in digital archives or computer engineering. But I can offer my own point of view as someone who finds those fields tremendously interesting and who wants to more fully integrate them into the discipline of Film and Media Studies. I'm also a believer in "bootstrapping" digital projects, in finding ways to quickly move forward on a digital project, rather than waiting months or years for a grant agency or large institutional library to say "yes." I use the term "bootstrapping" for its general entrepreneurial connotation. However, the word carries additional significance in computer programming. To bootstrap is develop simple programs, then develop increasingly complex programs atop of them, scaling up the project in the process.

I worked with a small team to bootstrap the Media History Digital Library, and I'm currently taking a similar approach with our new search tool, Lantern. In both cases, I got in way over my head and learned a great deal in the process. The following are what I have found to be the five most important lessons, questions, and considerations for creating a digital project.

1. Who is your audience?

Walt Whitman remains one of the most widely read American poets. What if there was an online resource that contained all of his writings? The Walt Whitman Archive (http://whitmanarchive.org), edited by Ed Folsom and Kenneth M. Price, provides readers with remarkable access to Whitman's oeuvre—including digitized original texts and letters, poems marked up in XML, and translations in numerous foreign languages.

Film and media educators need to show movie and television clips in class, however, they tend make decisions about what to show in isolation and often lack access to many of the works they would like to screen. What if there was a community of educators who uploaded their clips and critical analysis to a shared database? Critical Commons (http://criticalcommons.org), founded and directed by Steve Anderson, serves this community of educators.

The Walt Whitman Archive and Critical Commons both effectively serve their respective audiences. Impressively, the Walt Whitman Archive finds ways to engage researchers, teachers, students, and Whitman fans alike. We've tried to reach a similarly mixed audience in creating the Media History Digital Library. The MHDL has found traction with film historians and devoted fans who were previously familiar with the historic periodicals on the site.[1] We hope our new full-text search engine, *Lantern*, opens up the project to a much larger group of students and fans who are passionate about film and media but who were not previously aware of the existence of the periodicals like *The Film Daily* and *Photoplay*. Film and media hold a great deal of popular interest, and we are missing a huge opportunity if we don't engage with this broader public.

Unfortunately, there are many Digital Humanities projects that never find an audience. Scholars don't spend time using them. Nor does the public. Unlike the Whitman Archive, Critical Commons, and MHDL, these are websites that no one visits besides the spam bots.

In building your project, think about whom you are trying to reach and the best ways to engage them. The academic audience can be one of the most difficult to engage. Scholars, on the whole, tend to be efficient with their time. They spend time seeking out the resources that they think will benefit their research. An article or dissertation can be quickly discovered, skimmed, and, if relevant, cited. If you choose an unconventional navigation design, your innovation may have the unintended consequence of alienating your target users, who have a difficult time finding what they are looking for.

I would also suggest thinking about how you can engage a wider public audience beyond academia. One thing to keep in mind: The public's barriers to entry to a web-based project are very low, but the barriers to exit are even lower. You need to immediately grab the audience and give them reason to stay on your site. Otherwise, they will find better ways to spend their time.

Placing the audience at the center of your design process takes a lot of work. But without the audience, what would be the point of creating the digital project?

2. What are the legal considerations?

Let's say you cleared the last question with no problems at all. There is an audience of academics and public web users who are hungry for your material. You know exactly the interface and strategy you need to reach them. You've digitized several terabytes worth of data that you are ready to put online. But wait—what gives you the right to put any of this material online?

Before you post any written or audiovisual works to the web, you should investigate the legal issues. In the case of unpublished archival works (such as letters), you'll need to review the agreement between the donor and archive and see what distribution limitations may be in place. In the case of published material, you'll need to investigate the copyright status.

If a work is copyrighted, and you still want to include it in your project, then you have three options. First, you can contact the rights holder about licensing the rights. Second, you can include the material in a way that qualifies as a "fair use" by using the material in a manner that is transformative and/or includes only a small excerpt in order to make a critical point. Third, you can upload the work in its entirety and just hope for the best. The first option may not be viable due to the licensing costs or the inability to locate the rights holder. The third option should be avoided in most circumstances, though there are compelling arguments pertaining to "orphan works" in which the rights holder absolutely cannot be found. The second option—fair use—may be your best course of action. Fair use enables the Critical Commons community to keep uploading and downloading clips of copyrighted content alongside users' critical commentaries. (And, indeed, Critical Commons as a project can be viewed as an argument in favor of film and media educators exercising their fair use rights and special copyright exemptions).[2]

The Media History Digital Library primarily works with public domain material. Consequently, we can digitize these out-of-copyright works in their entirety and offer broad access to all users. David Pierce, founder and director of the MHDL, has twenty-five years of experience investigating the copyright status of books and films. Before we scan anything, David researches the copyright status. They key question—at least for works published prior to 1964—is whether or not the rights holder chose to renew the copyright after the initial 28 years of protection.

If you are looking for guidance, there are online resources, such as the University of Pennsylvania's "Online Books Page" and Columbia University's Copyright Advisory Office, that contain lists of public domain works.[3] As Columbia's Copyright Advisory Office warns, though, the "list is not meant to be construed as a silver bullet to completely avoiding copyright" and the University does not make any "warranty that the materials are, in fact, without legal protection." If you are considering spending several thousand dollars on scanning a series of texts, then hiring a copyright consultant for a couple of hours of work could be a worthwhile investment.

3. What are the ethical considerations?

After you've addressed the legal issues, there are still ethical questions to consider. If someone else invests the time and money to digitize a public domain text, is it ethical to quietly copy the file and re-host it on your own site? David Pierce and I decided the answer was "no." As a result, the MHDL only includes works that we helped to digitize, works in which the digital collection gave us permission (such as the Prelinger Library's lengthy run of *Business Screen*), or works that are part of our same open, shared hosting environment at the Internet Archive. Additionally, in all of these cases, we provide attribution to the original collections and sponsors.

The ethical questions relating to unpublished works can even more challenging. To publish something, by definition, requires making it available to some public. But what about works that were not intended to be openly public? If researchers have to travel to a particular archive to view a collection of handwritten letters, then putting those letters online would certainly improve access. But is it ethically appropriate to make these private artifacts immediately accessible to the whole world? We might say "yes" in the case of a head of state but "no" in the case of a less public figure.

Archiving the heritage of indigenous cultures presents additional legal and ethical complexities. Copyright only protects tangible expressions, leaving intangible cultural heritage—such as practices, knowledge, and cultural spaces—beyond the scope of protection. Moreover, the Western IP system sees copyright as a temporary monopoly that stands only until the work enters the public domain where all may freely access and use it. This view is quite different from that of many indigenous communities that consider their rights over their culture to be perpetual and believe that access should stay limited. In their "Digital Dynamics Across Cultures" *Vectors* project profiling the Warumunga people of Australia, Kim Christen, Chris Cooney, and Alessandro Ceglia provide access to the Warumunga that is "provisional, barring access to specific images or performances, in a manner consistent with the logics or protocols of the Warumunga people."[4] Christen's content management system, Mukurtu (http://www.mukurtuarchive.org/), further advances these goals and allows indigenous archivists to set limits on what different users are allowed to see. Christen's work invites us all to consider the ethics of looking, access, and preservation as they relate to cultural norms. It's a valuable lesson no matter what type of digital archive you want to make.

4. What are your plans for data preservation and sustainability?

Preservation and sustainability are two different things, but both are vital to the long-term success of your project. You need to make sure the hard work you put in today will not disappear tomorrow. Furthermore, you should care about preservation and sustainability, because they will be crucial to your project's ability to attract grant funding. The NEH's Digital Humanities Grant applications, for example, require detailed plans both for data management and sustainability.[5]

The Media History Digital Library is more of an access project than a preservation project, but we still try to make good decisions about data preservation. First, we choose to work with widely adopted open source formats—XML for our metadata, Apache Solr for our search engine, and Ruby on Rails for our search interface. Second, everything we scan is saved within the Internet Archive's robust preservation framework, which mirrors the data across multiple sites and makes it easy for users to download their own copies (which also helps from a preservation standpoint).

Of course, there are many other good preservation options beyond the Internet Archive. If you are affiliated with a university, your library may already have a digital preservation repository, such as Fedora, in place. Talk with your librarian and find out how you can safely store your data using their system and what, if anything, your library charges for the service. And, in the meantime, remember the 3, 2, 1 rule that Dorothea Salo taught me: 3 copies; on 2 different formats (i.e. external hard drive and a cloud service); and 1 copy that lives off-site. You'll also need to audit your data at least once a year.

The sustainability side has both technological and human challenges. The technological challenges are those that come when you work with rapidly developing open source technologies. In developing Lantern, I have been amazed at the pace at which Ruby on Rails developers keep developing new "gems" (sort of like plug-ins) that enable new functionalities. However, some gems are dependent on specific versions of Ruby or other Ruby gems, and before you know it, you are trying to run Ruby 1.8.6 and Ruby 1.9.3 simultaneously in the same app (which, for those of you who haven't

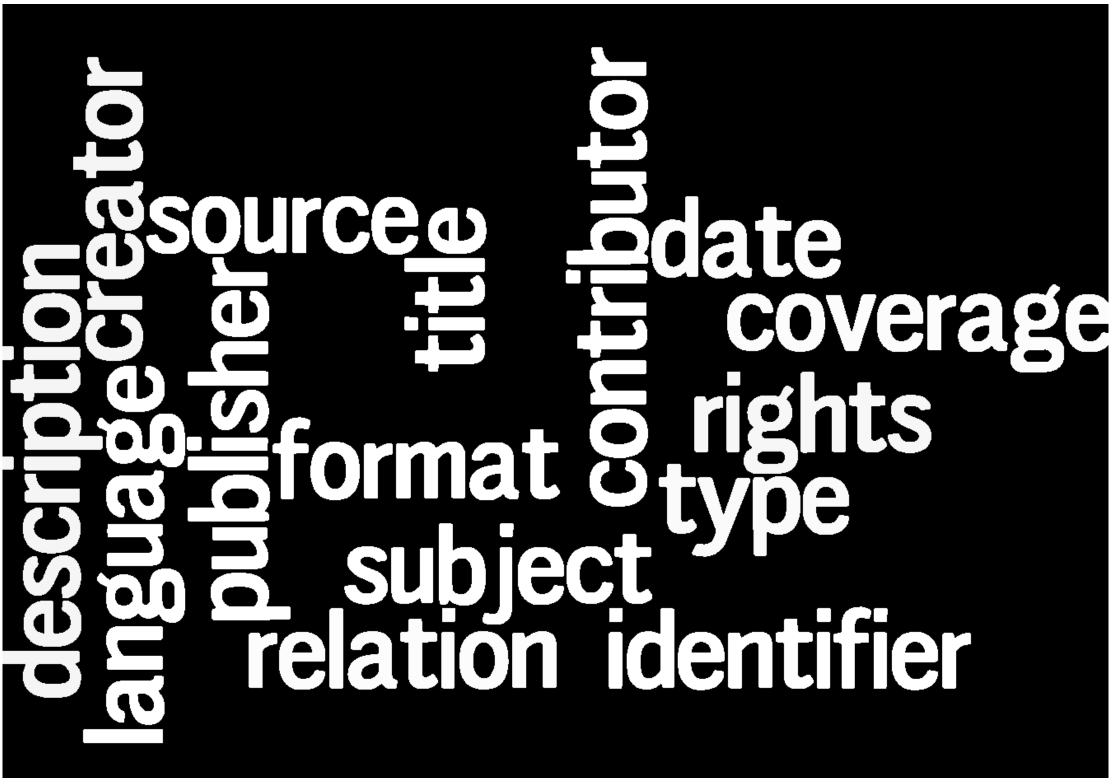

Figure 2. These 15 Dublin Core metadata elements may be the most important pieces of XML code that you learn. Image created by the author using Wordle.

tried, doesn't work well). Utilizing a development environment with version control is essential for any hope of staying sane—and keeping your web-based software sustainable for more than a few weeks.

As hard as the sustainability tech side might seem, it's actually easy compared to the human side. Who is going to spend the time adding new content to the site, updating the software, communicating new initiatives, and coordinating the digitization of more material? A three-year grant allows you to pay someone for a maximum of three years. What happens after that point? You need to be prepared for the possibility that you personally might have to take on most of these functions.

5. Learn to code and embrace collaborators

"I don't have time to learn how to code. I'm not interested in learning any computer languages. I'll direct my digital archive, but I'll find someone else to do all the programming stuff."

This is a fairly common attitude among academics who want to initiate digital projects. And it needs to change.

Let me begin by acknowledging that it is unrealistic to think you can take on a complex digital project all by yourself. The Media History Digital Library could not have digitized 500,000 pages of historic texts without close collaboration between David Pierce, Wendy Hagenmaier, and myself. Moreover, our entire digitization model has been built on the collaboration between collections, who loan materials for scanning, and sponsors, who pay for the scanning. We collaborate with the Internet Archive on the scanning process and data preservation. And there is no way our work on Lantern could have advanced without the input of Carl Hagenmaier, Andy Myers, Joseph Pomp, Pete Sengstock, and Derek Long.

Embrace collaborators—you'll need them. However, it is naive to think you'll be able to lead an innovative digital project without learning any code yourself. I know the prospect of learning computer programming can seem intimidating. When I first started trying to teach myself about programming, I got hung up on the idea that I needed to be proficient in one particular language, like C++, Java, or Python. My programming mentor and Lantern collaborator, Carl Hagenmaier, was the one who pointed out that the languages all follow a very similar logic. So rather than fixating on a particular language, I needed to learn the fundamentals of computing technologies.

I now feel fairly comfortable coding in Ruby and working on my Unix command line. More importantly, though, I can think through how to solve a problem algorithmically—that is, how to create a series of instructions to achieve the outcome I want. This gives me the ability to communicate much more easily with Carl and other engineers who know way more about programming than I do. We can collaborate on finding the right solution and design.

Similarly, it's important to learn about metadata standards and data management frameworks. You'll be able to better understand why librarians and archivists make the decisions that they do. Moreover, you'll be more likely to make good decisions yourself about how to store your data and metadata. A few hours of learning the best practices up front could mean the difference, a few hundred hours later, between having a large data set that you can easily share and having one that is unusable anywhere beyond your own project. Put another way, your education in XML is not complete until you learn the fifteen Dublin Core metadata elements.

If you aren't willing to learn the fundamentals of computing, metadata standards, and some basic code, then it will create real practical problems to achieving your vision. You won't be able to accomplish tasks on your own, so you will try to hand them off to someone else, with whom you'll be less effective at collaborating and communicating

However, there is another problem as well. When humanities faculty remain blissfully ignorant about the programming process, it devalues digital projects that aren't simply text-based. Tenure committees tend to look at these projects as less significant than a book, even when the usage metrics show that many, many more people are engaging with the digital project. Stephen Ramsay and Geoffrey Rockwell have astutely argued that, "To ask whether coding is a scholarly act is like asking whether writing is a scholarly act. Writing is the technology—or better, the methodology—that lies between model and result in humanistic discourse." Greater computational literacy is vital if we are to demand—and achieve—the very best digital work that advances humanistic knowledge both inside and outside of the academy.[6]

Conclusion

I'll conclude with one last thing to consider—it's ok to learn as you go. You don't need to have all the pieces described above figured out when you start. Bootstrapping is, after all, about jumping into a project with imperfect knowledge and imperfect resources and learning as you go. You can't simply wait around for the optimal conditions to present themselves, because, generally, they never do. Start with an initial plan, prioritize what matters most to your project, and try to make the best decision possible at every step.

Your project is bound to change in ways you could never have anticipated. As I mentioned earlier, the MHDL is more of an access project than a preservation project. We scan the materials in a non-destructive way, then return them to their owners. However, this is starting to change. In 2012, we received three large institutional deaccessions of old magazines and books. We've gone from being purely a digital collection to now collecting some physical, print artifacts as well.

One deaccession included at least 20 years worth of motion picture indexes. The indexes are fascinating artifacts of an earlier generation's epistemological framework. I'm aware that my own digital work, along with resources like the IMDB and AFI Online Catalog, is making these sorts of indexes obsolete. Few users nowadays crack open one of those tomes to find the answer to a question. Meanwhile, the indexes occupy valuable real estate on a library's shelves.

I wonder what the next generation of researchers and users will think of the MHDL and Lantern. Will our own designs, categorizations, and assumptions seem similarly obsolete? I don't know the answer, but I hope we can keep the project and data alive long enough to find out.

Eric Hoyt is assistant professor of communication arts at the University of Wisconsin–Madison. He is co-director of the Media History Digital Library and project lead of Lantern. His articles on media, law, and culture have appeared in *Cinema Journal*, *Film History*, *Jump Cut*, *World Policy Journal*, and the *International Journal of Learning and Media*. In 2012, he completed his Ph.D. in Critical Studies in USC's School of Cinematic Arts. He is currently working on a book manuscript about the content libraries owned by the Hollywood studios.

End Notes

[1] As I pointed out in a 2012 discussion on Henry Jenkins' blog, the MHDL only exists because of the commitment of fans who saved trade papers, fan magazines, and other items regarded at the time as ephemera. "The Affordances of Technology for Media History Research," *Confessions of an Aca-Fan: The Official Weblog of Henry Jenkins*, 5 December 2012, http://henryjenkins.org/2012/12/the-affordances-of-digital-technology-for-media-history-research-part-one.html and http://henryjenkins.org/2012/12/the-affordances-of-technology-for-media-history-research-part-two.html.

[2] For more on fair use and media, see Patricia Aufderheide and Peter Jaszi, *Reclaiming Fair Use: How to Put Balance Back in Copyright* (Chicago: University of Chicago Press, 2011).

[3] Columbia University Libraries / Information Services Copyright Advisory Office, "Public Domain Resources," http://copyright.columbia.edu/copyright/special-topics/duration-and-the-public-domain/public-domain-resources/ (accessed 7 February 2013); University of Pennsylvania Library, "The Online Books Page," http://onlinebooks.library.upenn.edu/lists.html (accessed 7 February 2013).

[4] Vectors Journal Editorial Staff, "Editor's Introduction to Digital Dynamics Across Cultures by Kim Christen, Chris Cooney, and Alessandro Ceglia," *Vectors* 2, No. 1 (Fall 2006), http://vectors.usc.edu/index.php?page=7&projectId=67 (accessed 2 February 2013).

[5] National Endowment for the Humanities, Office of Digital Humanities, http://www.neh.gov/divisions/odh (accessed 5 February 2013).

[6] Stephen Ramsay and Geoffrey Rockwell, "Developing Things: Notes toward an Epistemology of Building in the Digital Humanities" in *Debates in the Digital Humanities*, ed. Matthew K. Gold (Minneapolis: University of Minnesota Press, 2012), 75-84.