Technology and community: Why we need partners, collaborators and friends

Kate Zwaard, Library of Congress | Oct. 14, 2016 | Charles E. Young Research Library, UCLA

KATE ZWAARD: ***[00:07, Slide 1]:*** I just want to start with a quick caveat that I'm not here talking on behalf of the Library of Congress when I'm expressing my personal opinions. I'm using experience that I learned there. I'll be talking about a few things that I think, and it's not the official position of the Library of Congress, just to kind of state that. A little bit about my background is I'm a software developer and I've come to libraries as sort of my adopted home that I love very much, but my undergraduate degree is in journalism and statistics so it's – news is really where my heart is so I'm really excited to be here today to talk to you about this.

[00:47, Slide 2]: So, we have a little bit of time together, we've got about half an hour. My talk is pretty dense with information so I'll be looking at your faces – please use your bodies to indicate to me how it's going and if there's something that I'm talking about that's not quite landing; I can explain a little bit more. And so, I'm going to start a little bit with talking about my team, which is brand new, just to give you some context into what I do at the library and how I think that we can help. Excuse me I had bronchitis for three weeks so I'm a little bit rusty.

[01:24]: And then I'll go into some things that I think about in terms of emerging standards, tooling, partnerships and friendships that I think can help us all move together in this space. We first start with my team; we're very small. Like I said, we're not quite a year old yet but we're interested in doing some exciting things. And before I get into the nitty gritty about what we're working on, I'd like to start with a quick story.

[01:58, Slide 3]: So Henriette Avram — who in this room has heard of this person?

[02:03, Slide 4]: Great. Awesome. I'm so excited. She was born in 1919 in New York. She did two years of pre-med at Hunter College and then left to start a family. When she was about 30 years old she started programming, which I think is really cool because there are still some people who will tell you if you haven't been writing code since you were] a baby that you'll never make it as a programmer. And I'm here to tell you that they're wrong. She did, and she changed the world. So, Henriette invented the MARC cataloguing standard. Who in the room does not remember using a card catalog? Everybody remembers these. So, she invented the format that made it possible to do precision searching and searching at a distance. I loved the way card catalogs smell but I don't really miss anything else about them. Search is much, much better. And she did this in the late 1960s — and to give some context for the other nerds in the room, this was a time before relational databases, before computer networking and before character and coding standards were really mature. So, this was at a sea change.

[03:08, Slide 5]: And what I think is really interesting about her story is that she did this with the power of cross-discipline fertilization. She was not a librarian by training. She was a software developer. And she came into the Library of Congress with that perspective. And she saw the future of how computers were going to be used around information. At that time computers in libraries were accounting for accounting, right, they weren't used for information sharing. And because she had that unique perspective she was able to cause this sea change that has affected all of our lives

here today. And I think about this in context of this room here and what we might be able to do with sharing our different perspectives and cause the same kind of sea change.

[04:00]: So a little bit about my group — National Digital Initiatives — we have a couple of goals and I want to tell you about this because there's been some talk today, and yesterday, about the Library of Congress and what the Library of Congress can do, and, I'm not in the Copyright office. I don't make acquisitions decisions. But, I do have the power to encourage some collaboration to incubate new projects. So, if that's the kind of thing that you're interested in, I want to talk about that as a possibility.

[04:40, Slide 6]: So our first goal is to maximize the value of the digital collection. We have a lot of stuff at the library Congress; a lot of it's digital. We want to think about how we can squeeze all the juice out of the orange. So how can we encourage digital humanities-type of research? How can we reach out to universities and make sure that they're making use of our information? How do we reach life-long learners? How do we talk to journalists to make sure that they're using our resources when it's applicable to them and they understand the value that reference can bring for their stories? That sort of stuff. We also want to encourage creative reuse.

[Slide 7]: So the Library of Congress has a bunch of photographs up on Flickr, and, one of the neat things that kind of came out of this project was that people started tagging the awesome mustaches. And it's fun, right? It's fun and it brings attention to the collection but it can also serve a scholarly importance, a purpose that's of scholarly importance, too.

[05:30, Slide 8]: The Library of Congress has on loan the Rosa Parks Collection, which is a huge treasure trove of manuscripts and photographs documenting her personal life and her work on behalf of civil rights for African Americans. And it contains this hand-written note to her friend, which says, I had been pushed around all my life and felt at this moment I couldn't take it anymore. And this gives us a window into her as an American hero in that moment.

[06:00, Slide 9]: But other pieces help us understand her as a human; as a friend; as a family member; and that's really valuable, too. And that's why I love her pancake recipe. I've heard that it makes really great pancakes, but it also shows us that she was a person; and that we have the moral obligation to make those kinds of decisions in the moment, too; that she wasn't an icon frozen in time.

[06:26, Slide 10]: The second thing that my little team is doing is to incubate, encourage and promote digital innovation. I used to call us semi-permeable membrane; but that got a lot of weird looks. So, my staff told me I could not use it anymore. Don't tell them.

[06:41]: But I think there's a lot of exciting work going on in the Library of Congress. It's really big. And it's hard for people to get a glimpse into some things that are often going on in the production-side of the world. And I'd like to help that information flow out a little bit better. But also, I want to help information flow in. So, there's a lot of our staff who is focused on the production-nature of their work. They've got work to do. They don't have time to go to conferences. And what I can do is help introduce those folks to some people out in the field doing work that's similar that they can benefit from.

[07:15, Slide 11]: Also, I think we could act as a catalyst. Who loves chemistry?

I love chemistry too. So, I'm sneaking in an energy diagram here. This is, as you all know, an energy diagram of an exothermic reaction. And exothermic reactions are spontaneous; which means they happen automatically, but not if it requires an activation energy. And I think about a lot of our products — our projects — as things that are like that too; that we're excited about, that we're crowdsourcing in libraries, right? We want to do it. But, we need a little bit of push to get off there to get over that hump. We're not enzymes; I mean, they enable life, so I can't say that we're that good.

But what we can do is provide a little bit of technical expertise; provide some context to other people; show folks some open-source projects that might be relevant. So, that's the effort that I think can get us over that little hump.

[08:12, Slide 12]: So I'll tell you a few things that we've done in our first year. One of the things is we co-hosted DPLAfest which is really exciting. The other thing is we co-hosted a hackathon that Matt Weber and two of his colleagues organized; it was super fun. One of the things I really loved is that they were very intentional about opening it up to programmers and also scholars and librarians. And when I talked earlier about that cross-discipline fertilization I really was struck by the value of it there. Having librarians in the room who understood the datasets intimately was hugely helpful to the scholars and I think it really illustrates the evolving shape of reference. It was super cool and I hope we get to do it again.

[08:56, Slide 13]: We're also working on a digital scholar lab. So, the Library of Congress is a little bit different from most large libraries; we're not attached to an academic institution so we don't have students and we don't have faculty members that we can partner with. But we do have the John W. Kluge center; so that's a place where scholars come from all around the world to be in residence at the library for a little while to use the libraries resources but also to publicize their research and share their information with policymakers. And we've been thinking a lot about how we can enable more digital scholarships-type of research, more digital humanities; and we've engaged with two outside experts — Dan [inaudible], whose name has already come up, and Michelle Gallanger — to write a report about the infrastructure that might be necessary to do things like that. And following that we would like to do a small-scale implementation just as a proof of concept.

[09:54, Slide 14]: We've started a library innovation fellowship this year. So, we have selected two candidates who are charged with doing something innovative and interesting with the digital collections. We hope to advertise this a little bit more widely in future years. Biking will not be required. That's just a very weird coincidence, so, if you're more of a walker or a driver it's going to be fine.

[10:20, Slide 15]: And the last thing is, we just hosted this big event, "Collections as Data," Laura and Matt were there, and it was really fun and it was really great. The live-stream video is online, so, if you're doing some manual data entry at your desk I encourage you to check it out. All of the speakers were great. They were just phenomenal. But it really explored the contours of how we can use our collections as data. And we intentionally broadened it from digital humanities; we want to include all walks of digital scholarship including social science. But, you know, the ethical challenges, the technological challenges; it was a really great event.

[11:01]: So, I'm going to pivot a little bit. I think Edward can attest that I handed him my slides before I came here, right?

[11:15, Slide 16]: OK, because I feel like I'm doing a little bit of cold reading; like, there's a lot of themes that have come up over the last couple of days that are in here, and I really don't want you think I'm copying anybody, OK? But it's neat and there's some fun things that come up again and again and I really encourage you, if you're ever going to give a talk like this, go first. So, as you might imagine, I think a lot about the tension between innovation and sustainability. I think about how I like to say that the Library of Congress is a building carved out of marble by revolutionaries; that we've been doing innovation and sustainability for a long time. But there's this perceived tension here; and how can libraries maintain the qualities that make them endure, but still be agile and innovative. And, in that way I think we're natural partners with news; which has a bias in sort of the opposite direction. News for the most part is not dominated by the sort of large, sustained, durable organizations that we've heard from today. They're dominated by small, agile kind of fleeting organizations. Community — the bad news is, community news is dying. But the good news is, it always has been. You know, the national landscape is littered with the death of small newspapers. So, this is not new. We can do this, right?

[12:38, Slide 17]: So I realize this is obvious to all of you because you're here; but I'm going to get into the meat of my talk now, which is historic partnerships, emerging standards and new opportunities. Thinking specifically about the Library of Congress, we have a different relationship with printers than most institutions. Obviously that's through that mandatory copyright deposit, and it's commemorated in the beautiful Jefferson Building. If you ever get the chance, if you're ever in D.C., please stop by. I think it's the prettiest space in Washington.

[13:12]: We have this long, historic partnership that has endured many changes in copyright laws, and that legal relationship has created personal relationships that have resulted in other adjacent goods. For example, when The Washington Post recently moved buildings they donated some print material to us that was outside of the copyright law, but, I think we need to figure out how to maintain those relationships in the digital world where we're not tied by those same rules.

[13:46]: I also want to mention that we are doing a pilot with the news — it's the News Media Alliance, formally the NAA — to test the technological ability of the library to accept and for them to create e-prints, which is an alternative to microfilm. So that's going on right now; we're working with Gannett, with The New York Times, and with the Washington Post and Wall Street Journal, which is pretty exciting.

[14:22, Slide 18]: I'm going to talk for a second about the National Digital Newspaper Program. This is not online news; this is like the opposite of online news. But I think it's cool for two reasons: One is to show — you all know the importance of saving news, because you traveled here and you're sitting in these uncomfortable chairs, and are enduring my talk — but, it's nice to have tangible examples that you can use to talk to other people. And we can look to our history to have those. And the other is to show this historic partnership and the other benefits that it has had outside of its product. So, the National Digital Newspaper Program is a historic partnership between the Library of Congress and NIH and hundreds of state partners all around the country to digitize historic newspapers which are ingested at the Library of Congress and provided at chroniclingamerica.loc.gov.

[15:10]: It's really neat; and I think one of the cool things that you find out of these historic archives are very personal. So, people use these for personal ancestries and family trees, but also some really cool academic research comes out of it.

[15:35, Slide 19]: NEH recently issued a data challenge where they gave some prize money to people doing digital humanities-style research with the OCR that we make available. And this is one of the really interesting projects that came out of it. Lincoln Mulligan, who's a researcher at GW, did some text analysis around Bible verses, and showed what Bible verses were being used at what time; tracked them over time; and which ones were adjacent. And I think this is really interesting because it shows the kind of knowledge that we would never have had if we hadn't preserved these newspapers. But also, it's the sort of research that is not directly competing with the use of those newspapers. So, it's almost secondary knowledge in that way.

[16:14, Slide 20]: So I want to talk a little bit about push versus pull. And we talk about a lot of this at the Library of Congress; what should our mode of getting material be in the future. It used to be that we just open the loading dock and things came, right, so that's the pull model. But the digital world is a little bit different for various reasons. And I think when we're talking about the current state-of-the-art in preserving online news, most of what we're doing is pulling. So, I think most of what is actually happening right now is through web archiving.

[16:56, Slide 21]: As most of you know, the Library of Congress is a very large web-archiving program. We archive a lot of born-digital news sites, so sites that don't have a print surrogate or whose print surrogate is kind of secondary. Examples of these are Verge, Vox, Salon, the Daily Beast; and these are sprawling sites. You know, you think about BuzzFeed and how big it is. We try to crawl those regularly and really in-depth. But what we miss when we do that is the stuff that goes fast. So, an approach that we're taking recently is using RSS feeds. This was pioneered by the Icelandic

National Library. And what that allows us to do is get the stuff that's brand-new right away. And we're doing a very shallow crawl using the RSS feed, and a deep crawl using the site. I mention this

because RSS has a standard — it's not a new standard but I think this is kind of a new application of it in the way that I think is really fun.

[17:53, Slide 22]: The next standard I want to talk about is IIIF — who here has heard of IIIF, who here knows what it is — OK. I feel like that's good. That's actually way more than I thought. I really encourage you to check out the talk "Everything you always wanted to know about IIIF but were afraid to ask" from DPLA.

[18:53]: So the news folks gave us web developers in libraries Django, right. Django is hugely used in libraries and it came out of news. I think this is a thing that we can give news folks. If you're ever looking to replace reptile server, check out IIIF. It's a really cool standard that allows you to use an API to request different-sized images, to request metadata; what it was intended for was to enable annotation and shared image viewers, but that's not really important today. What's important today is what it gets us in terms of integrating libraries and news, which is virtual repatriation. So, what we can do is use our viewer to show an image from the news site giving them the hits, which may or may not be important but it's just one more tool in our tool box. The other thing that it gives us is structured URLs; which, if we enter an agreement with a news organization to ingest their site that's a thing that we can use to get the images in a way that is automated that makes sense. I don't know of any adoption in the news community of this but I do know JuxtaposeJS, which is an image viewer that you can use to compare images, is very similar; shares a lot of features in common with Mirador. There's definitely some mutual interest here that I think might be worth exploring.

[19:50, Slide 23]: StoryCorps — who here's heard of StoryCorps. OK great. That's perfect. Then I don't have to explain it.

So, you may have heard that StoryCorps recently got some money to make a mobile app — if you haven't downloaded it, you should, it's really cool — so that you can experience the StoryCorps experience at your Thanksgiving table and capture the oral histories of your loved ones from the comfort of your own home. When they started building their app they engaged with us at the Library of Congress to figure out a way to archive that. We work with them to develop their API, which is public, and we are using a script to ingest that content to the library and preserve it. We wrote an iPress paper; if you're interested in learning more, you can check that out. But one of the really interesting things about engaging with this — and this is kind of news adjacent, right? So StoryCorps ends up on NPR, which is news. But I think that technical parallels are very similar. One of the interesting things was in working with them, talking about checksums and fixity was so foreign to them. They kept talking about their API as a stream. And we don't think of it that way; we need discrete chunks that are frozen in time.

[21:00]: So making decisions about — were we going to take comments? Were we going to take likes? Were we going to take metadata? How final is final when things get tagged? So, there's a distribution of times where when a new thing is published, it gets tagged, and where's the sweet spot in which we should take it? So, that kind of conversation is really interesting. And the mutually beneficial relationship between the two groups was kind of neat to explore.

[21:37, Slide 24]: I like to think about how we can encourage media organizations to donate their back catalogs or sell, because as you heard in the print world we didn't get those for free either.

[21:50, Slide 25]: And how can we encourage libraries to accept this stuff? Because there are technical hurdles there, too. Let's start with publishers. I think one of the ways that we can encourage more publishers to donate or sell their back catalogs is to talk about the scholarly use of this material and explain that it doesn't really conflict with regular readership. We do this all the time with our manuscript collections. We take people's manuscript collections, which are of value to them, and we

take them into libraries and we make gift agreements that place a box around that so that we're not opening up to the world. And I think having some sample gift agreements before we start the conversation with publishers could really be useful. I think there's also, from the library side, can be a little bit of a cultural issue, too.

[22:46, Slide 26]: So, we're looking at Wired "explainer" on how to use Snapchat, which is a toy for children from Wired which used to be a magazine for people in the know [laughter]. And when you think about the news that we all read, which is getting increasingly more fractured, there are some things that we all experience; everybody in Washington reads the Washington Post, but lots of people in Washington read Wonkette and FiveThirtyEight and lots of people in Washington don't know what those are. That's becoming more and more true. And I don't really have a solution for this yet; that when you're talking to your acquisitions folks explaining to them what the cultural significance of some of this material is. But, I think it's an interesting challenge.

[23:31, Slide 27, Slide 28]: In terms of technological challenges, The New York Times talked about their data migration and you heard how hard it is. Anybody who's gone through it once does not want to do it again. But in libraries, what we're talking about is doing that a hundred times, a thousand times, a hundred thousand times if we're going to take these archives on. It's not really scalable.

[23:52]: So, we've got to think of some alternate ways of taking this material in. We need to think about ways to take this material in that makes it usable. Libraries are not a hole for dropping books in and they're not a hole for dropping data in, either. We want to preserve this to make it useful. And I've got a couple of ways that we can think about making that doable.

[24:26, Slide 29]: One is to put Jupyter notebooks on top of raw database dumps. So, thinking about making this material available to researchers who are using digital humanities and data-style research. I think Jupyter notebooks are a really fun way to do that. It tricks people who are afraid of code into coding; you can be like, if you're afraid of code just type things into this web page. I wish I had more time to talk to you about that but I don't.

[24:53, Slide 30]: Another thing is to rely on virtualized hosting. So, we do this a lot now with our scholarly articles. The Library of Congress will accept an archive and will preserve it. We make it actually available to our researchers through database subscriptions. And, similarly, if there's a media organization that was going out of business or wanted to donate their archive to us what we could do is host it on a virtual app virtual server and ingest it, and that way we could provide access in a usable form.

[25:32, Slide 31]: And the last thing I want to talk about is the importance of friendship. Tech folks in libraries and tech folks in news share a lot in common. We use really similar stacks and in some cases the same exact web frameworks. I spend a lot of my time on Twitter, and a lot of my Twitter friends that I hang out with are from news organizations just because we have the same kinds of problems. We have large back catalogs we want to leverage; we have the same problems of digital in a print world or print in a digital world; and the same kind of skill issues, where we're basically physicists in the 50s in libraries, right? It's not enough to know physics. You have to learn how to write code. You have to learn how to use technology. So, making friends I think is really good; talking to journalists at the one on one level and recognizing that working together is like a muscle. The more cooperative projects we do, the easier it is to do more of them. Thank you.