DLF 2020 Forum

Engaging with the Public Domain: Scaling up the Deep Backfile Project (Transcript)

November 2020

[JOHN MARK OCKERBLOOM:] Hello, and welcome. I'm John Mark Ockerbloom.

[RACHELLE NELSON:] And I'm Rachelle Nelson.

[JOHN:] And today we will be talking about how we've led a team of over two dozen people working at the Penn libraries to research copyrights from home during the COVID crisis. We're identifying serials in our collection that are in the public domain, so that they can be shared with the world.

Like many large research libraries, we have thousands of serials in our collection from the 20th century. Many of them don't get much use. And of course, since the COVID crisis started, most of our print serials get no use at all. We can use digital versions of some of them. But there are a lot more that are not available digitally, or that are only available by subscription, even though they're in the public domain. There are now electronic resources that can help us determine which ones are in the public domain. For instance, most serials have no associated copyright renewals, which can make US serials as late as 1963 in the public domain. Some don't even have copyright notices. And that can make US issues as late as 1989 in the public domain. Basically, we're trying to answer the legal aspects of the question "Can we digitize this?" that you heard about in the previous presentation.

In the project we're describing today, we're publishing information we find about the copyright of our library serials as structured data files linked to Wikidata. We're also publishing links to free online copies of those serials as we find them. With that information published as linked open data, we're hoping to grow a public commons of serial literature online. That'll be useful not only while many of us are stuck at home, but also long afterwards.

[RACHELLE:] John has been working on serial copyrights for a while now, but it had mostly been a solo effort. A few years ago, we got an IMLS grant to produce a complete inventory of serials published before 1950 that had renewed copyrights. We also produced a guide that tells you how to use this inventory and other data to determine whether a serial issue of interest is in the public domain. This slide shows a small portion of that inventory.

[JOHN:] After we determined what did have copyright renewals, we then decided to see what didn't have those renewals. We created information tables for different sets of serials, including a number of popular subscription packages. We developed a system for reporting on the copyright status of those serials and called it the Deep Backfile project. We started working through those subscription lists. But since the only people working on it were me and my assistant Alison Miner, who were both working on it part time, progress was slow.

And then the COVID crisis hit. And as you know, a lot of things changed. Our university locked down, so users no longer had access to our print collections. And a lot of our library workers could no longer do the work they used to do when they were on site. We saw that this gave us an opportunity to have them help us research copyrights of the serials we had at Penn, so that we can make them more widely available when we had that capability. So we made up a new table of serials we owned at Penn that might be in the public domain. And you can see a bit of that table on this slide. It's one thing to run a copyright clearance project with a couple of part time workers like we'd been doing. It's a much bigger challenge to make such a project work with dozens of new people, all of whom are suddenly working from home, and might not have any previous familiarity with what we were doing. That's where Rachelle came in, and she'll pick up the story from here.

[RACHELLE:] Because of the work from home mandate, two projects were offered to Penn Libraries supervisors as options for those who could not do their usual work remotely: transcription of the Marian Anderson papers, and serials copyright research. This meant our project needed to accommodate staff with varied skill levels (for example, library specialists and library assistants), and areas of expertise (for example, technical processing staff and public services staff). Supervisors were asked to provide an individual's planned participation, and an estimate of how much time per week would be dedicated to the project. The estimate of hours per week ranged from four to 14 hours. 12 supervisors reported that 25 staff from 14 departments would participate. The departments were from all sectors of the libraries: departmental libraries (for example, biomedical, music, veterinary medicine); technical services (acquisitions, central cataloging, global studies technical services); circulation and stacks; and physical processing.

[JOHN:] Given the wide variety of participants, we needed to design a process that could involve library workers who might not have expertise in serials or in bibliographic research. We also needed to keep the technical barriers to entry low. So we designed a questionnaire form that could be completed in any web browser, without any additional software needed. This slide shows a portion of the form. If you select a serial to work on in the Penn serials table we showed you earlier, you're taken to this form to fill out in your web browser. Some parts of the form, including the title and the ISSN of the serial, are automatically filled in for you. Other parts of the form let you select answers from a pulldown menu, or type in an answer as text. We asked questions about what countries a serial was published in; about what renewals, if any, are recorded in online copyright databases; about whether they can find copyright notices; and about what free issue content might already be available on the web. The form also has links that send off searches to our local Franklin library catalog, to the Copyright Office's catalog, and to other sites. Often you can find the answer to the questions being asked if you follow the link and see what's at the destination site. There's also a link on the form to a help file with more detail on how to answer each question.

[RACHELLE:] Staff and their supervisors were invited to an introductory workshop held via Zoom. John provided an overview of the project explaining its purpose, and I gave a step by step demonstration of filling out the copyright research form. The demo was meant to assure participants that no special skills were needed to do the research. A recording of the session was provided for those who are unable to attend. We received an unexpected plug for the project when the library's IT project manager

referenced it in the May Tech Tip entitled "How to navigate tabs in browsers while working on the serials copyright research project". This was great because those starting out on the project would most likely open multiple browser tabs while doing the copyright research. A second workshop was held six weeks later, where we reviewed the research process. Staff shared their experiences and asked questions. It was also an opportunity to verbally express our appreciation with the staff's participation. We have a link to our project documentation on our last slide.

We ended up with a variety of different kinds of help and documentation. John tended to write detailed textual documentation while I produce illustrated reference guides explaining the process in a more visually accessible way. The live demonstrations in our online workshops, where we also took questions, provided more interactive visual explanations. Interaction could also happen afterwards, as people sent us questions not only via the questionnaire form, but also by email. In addition, every participant received a message from John upon their first submission. We both also responded to individual questions. Communication with the supervisors also continued throughout the project. They were informed of how we responded to their staff's questions and received a weekly report indicating the number of titles submitted by their staff. We used "Penn+Box", a local version of Box, to manage and share documentation and workshop recordings among the project participants.

[JOHN:] After someone answers the questions we ask about a serial, the serial is marked in our table as under review, and their answers get put in a queue to be dealt with by a back end reviewer. Currently, well, that's me. After I verify the information that's been entered, a copyright information file is created for the serial. You see it here as a web page, but underlying it is a JSON file that a script automatically creates from the form responses, and that I edit as necessary. The JSON format allows much of the information here to be processed by machines, as well as by people.

I'll draw your attention to a couple of parts of the page. Near the bottom, there are lines that note my responsibility for the information on the page, and that also thank the person in the library that filled out the questionnaire. That acknowledgement of each person's work is important to us, and it's automatically put into the JSON file at the time of its creation. Another thing to notice are the links to Wikidata and Wikipedia on this page. That information is not present in the JSON file. It's actually coming from Wikidata.

When we add a serial to our copyright knowledge base, we also make sure there's a Wikidata item for it. If there isn't one already, we create one. In either case, we make sure it includes the ISSN we use for the serial, and we make a link to our copyright information using a special Wikidata property that was created in an earlier phase of our copyright work. It allows anyone or any program using Wikidata to find the information available at Penn for this serial. In turn, we can use the other properties people have added to Wikidata items to automatically find and link to things like Wikipedia articles on our serials. We're currently training some of our librarians to create and edit Wikidata items on our serials. Not only does that speed up this part of our project work, but it also gets our librarians used to working with Wikidata, which we're also planning to use for other metadata projects.

[RACHELLE:] Once one of our library workers has filled out the copyright questionnaire for a serial, and we verified it and created a JSON file and Wikidata links, our Deep Backfile tool will automatically update

to include a summary of the information we found. That update occurs not just on the Deep Backfile table for Penn Libraries holdings, but also on any other Deep Backfile table we have that includes that serial.

[JOHN:] As we record this, it's been about four and a half months since we welcomed library workers to this project. And they've accomplished a lot. 25 people have worked on it so far, and they've researched over 5000 serials. It's going to take a while for us to fully digest the work they've all done. But we've published information on over 1300 of the serials they worked on. We've also linked to free issues of many of them, not just at the usual mass digitization sites and open access journal collections, but also to many out- of-the way sites of some of the lesser known serials that our library workers found, and that we hadn't previously known about. But even more gratifying than these numbers are the responses we're getting from across the internet that this information is making a difference. We've heard from a scholar in Eastern Europe, now able to find issues of historic journals that their local libraries had either purged or not collected in the first place during Communist rule. We've heard from multiple authors who've been happy to learn that they can use articles and images from periodicals whose copyright status they hadn't been able to determine previously. We've even heard from some folks offering us issues that they own of some of these serials to digitize or upload.

We still have some challenges to cope with in our project. As we mentioned, our library workers are filling out our questionnaires faster than we've been able to verify them and put them into our knowledge base. But training and empowering more librarians to do more of the back end work, like we've been doing with Wikidata, should help. And eventually, we should get to publishing information on all the serials that are in our Deep Backfile table. It just might take a while. We're also discovering issues like canceled ISSNs and other metadata issues in our existing catalog records as we go through these serials. But we hope to have better metadata in our catalog as well by the time we're done. And of course, trying to work amidst everything else that's happening this year is not easy to begin with. We have to be mindful of that and be supportive of all our workers, some of whom have to deal with more challenges than others. We also need to think about our priorities. We decided early on that we weren't just going to deal with submissions on a first-in, first-out basis. We make sure every worker gets some of their submissions reviewed at regular intervals. And we try to catch misunderstandings early and answer questions as we see them, so no one gets left hanging with a problem for too long. We also try for faster turnaround for serials that we know get used frequently, or that are included in popular subscription packages. But we'd also like to make sure that overlooked titles, particularly those from overlooked communities, get the attention they deserve. I don't always know what those titles are. But I hope that others, whether colleagues involved in the Penn Libraries' Diversity in the Stacks initiative, or folks at other libraries, can help out here.

That's just one of the ways that people outside of this Penn Libraries project can engage with it. As we said at the beginning, once we know that a range of a serial is in the public domain, we could digitize it, or open access to previously digitized content. But so could others. Maybe your library could. The questionnaire forms we've shown are on the open web. Anybody online who's interested in a serial we list can fill them out, and we'll eventually get to their submission and include it in our knowledge base. All of the information we put in that knowledge base is CC Zero, meaning it's free and open for anyone

to use as they see fit, just as they can with the Wikidata items we create and link to. And the information in Wikidata is being added to and used by people in projects all over the world. We've also talked with some people about creating Deep Backfile tables for cereals that their projects are interested in. If you have a list of serials whose presence online would advance research, learning, social justice and the public good (to quote from the DLF's mission statement), and copyright information on them would be useful. please get in touch with us.

We want to close by thanking all the people who worked on this project. You can find a full list on our project page, along with more information and documentation on how we're running it. The URL for our project page is on this slide. And thanks to all of you for listening, and we'd love to hear your comments and questions.