The Digitization of Archives: In Case of Emergency or the New Normal? An Overdue Conversation with Peter Hirtle (Transcript)

Thai Jones

Welcome to another episode of Overdue, I am Thai Jones, curator for American history at Columbia Univrersity's Rare Book and Manuscript Library. Today our guest is Peter Hirtle, archivist and copyright scholar, with whom we will be speaking about some of the big changes that Covid ushered in to special collections

Last spring during the initial lockdown, the Internet Archive announced the emergency library, the legal repercussions of which are still ongoing. Peter and my colleague, Lina Moe, talked about the thorny legal issues around digital lending more generally. We spoke also about how some libraries and institutions may have too tight of a reign on their collections, like when materials in the public domain still have terms and conditions put on their use.— what happens if those are ignored? Can a library take you to court?

We also touched on the issue of monetizing collections.. Libraries and museums have very valuable, sometimes priceless artifacts— but they also have huge labor costs and are dependent on institutions or grants— what kinds of revenue streams might be available for libraries to pursue? If the special collections in Nevada have decided they can't commercialize images of the rat pack, who can and should try to make money from their collections? And what happens when venerated cultural images get put on refrigerator magnets or billboards?

Peter is deeply knowledgeable about the conflicting demands and desires within libraries to care for, make available, but also control their materials. I hope you enjoy this conversation as much as we did.

Lina Moe (LM)
Peter Hirtle Welcome to the podcast. I'm so eager to talk with you today about digitization, and how it is impacting special collections, as well as the frameworks that guide access and reproduction. And that I think have become particularly important over the past seven months of the pandemic, as we have shifted towards digital access more and more. So to start with the issue of digitization, I want to ask you about both the benefits and the possible harms of it. And I'd like to start with the benefits. There's a clear connection that you point out between digitization and increased use and access. If we put things on the web, they'll be found and used. And you give the example of the University of Michigan's Making of America collection, which uses I believe, 19th century American periodicals. And you say that the online collection has 5000 page views a day, and this is this is dated, so it's probably more now. But this is a rate that I hazard is many, many times that of the borrowing of the physical copies before digitization. But you also make the point that putting things on the web changes how hard copy collections are used. So do you have a sense of not just how web usage drives increased hardcopy usage, but how digitization drives different kinds of uses of the material.

Peter Hirtle (PH)
In terms of usage? Surely, the COVID experience has pointed out how utterly important digitization has been and the fact that we've been able to make materials from libraries available in spite of the quarantines and the closures of so many of our reading rooms.

In this regard, the National Emergency library from the Internet Archive is a real eye opener, as is the Hathie trust separate to make its in copyright materials available. One of the most interesting things I don't have the exact figures at hand, the but I believe I recall reading it with the Internet Archive, that they have found that individual titles 10, the usage of individual titles hasn't been that great, but that the breadth of the materials that have been used is exceedingly broad that lots and lots of titles are accessed. And one of the things they have discovered is that many books are just consulted very briefly. And of course, it's the sort of thing that you would never go and visit a remote library with suddenly being able to do word searches have digital access means that the breadth of material I can find is so much greater. So I think we're seeing materials being used more broadly. That the tremendous increase in access, and we're starting to see some changes in research, I think that maybe we don't have people sitting down yet. and wanting to read the 400 page tome in Latin from the 16th century online, but much more sitting down and saying, well, let's take advantage of word searches, the kind of broad availability to, to look and see if there's something that might resonate with us.

LM
And I love your observation that digitization changes the relative position of a text in scholarship and general awareness. So you say that in hardcopy, the material may seem obscure, but when digitized, it can move more towards the center of the canon. And I think that digitization allows us to cast a wider net and then bring it in to scholarship.

PH
Yep, I was just talking virtually with a historian of 18th century America had been working on a project that relied on a lot of printed sources. And I asked whether that historians research was changed by the ready availability of digitized books, newspapers, other materials, and the response was, in some ways, not so much because so much of this had been microfilm in the redex project. So in principle, that historian could have gotten access to the material through microfilm, but in practice, the historian also found that doing word searches in those online databases made the process of running Search easier and may have led to factoids topics, other materials that it would have been easy to have overlooked if you were just sitting down scrolling on a microfilm reader trying to read text at that point. So not a complete game changer, but certainly, an assistant to research and having a broader range of material available online is really exciting.

LM
And it seems like there's also a sense that digitization doesn't just change the resource research practices of traditional scholars, but it also broadens the user base. So high school students

using digitized materials. And I can imagine, I mean, I can testify to the fact that sometimes it's uncomfortable to try to get a reader's card at a library. There's a kind of formal process and it can be intimidating, and not to mention the expense of getting there. But I wonder if you could talk a little bit about about that the expanding user base?

PH
Yeah, I think there is one obvious one is that genealogists have leapt into the full text resources, especially things like full text newspapers, but also the range of 19th century books looking for material and people who might have been intimidated or had difficulty in finding things in government documents, case law, other material. I mentioned in a way back in my FLIR conference paper, how much lexicographers were taking advantage of digitized material, the Oxford English Dictionary and people like that. So I think there is a broader range of people looking at the digitized resources, the K to 12. I mean, that's is a tough one. But I my sense, and I haven't looked at this formally, but my sense is things like the History Day projects, projects that are based upon primary sources have been enhanced by having digitized materials readily available.

LM
Okay, so with all of this in mind about the, I'd say significant benefits of digitization, I would love to talk to you about some of the drawbacks, or the possible harms. And you have written about several. One is that the erasure of printed books, gets rid of something aesthetically pleasing, and a tangible connection to the past. And the substitution of physical objects with digital objects is made possible because what most users want to do is read the text, not perform a close reading of the text as object, but you write that the substitution is also an identity crisis for special collections.

PH
Sure, I guess I was just echoing the concerns that Terry Belanger had expressed even earlier that if we cut down on the number of people coming into a repository, then it's quite likely that administrators are going to sit down and say, Well, why are we spending money to support a special collections, that isn't getting researchers to come into it, but I don't think sitting down and saying, Okay, then we shouldn't make our materials available digitally, we should force people to try to come in and actually work with the physical book, that's a no-win situation as well, because we're going to get to a point where people just gonna say, Hey, I can get 80% of what I want done online. And I don't want to spend the time and the money to do that other 20% of actually coming and looking at the physical items that are obscure, or I can't get access to online, it'll take the rare specialized scholar who want to be able to do that. So I think the trick is to sit down and recognize that the kind of rationale and reason for Special Collections is slightly altered by digitizations of materials, especially printed books, and that, therefore the focus needs to change if we are going to continue to have them be valuable components of our libraries. And I think they are and I think it's quite easy to make those changes.

LM
I want to ask you a bit of a thorny question for, for someone in my role. Which is that? I wonder if digitization changes who stewards the collections once they are digitized. So digitized Special Collections might not seem like they necessarily need to be separated from general collections. That is they can be added to digital libraries more broadly. So does this have an impact on the role of the curator versus the reference librarian? And does that matter?

PH
Well, you know, no, even 20 years ago, we were saying that all academic libraries are going to become special collections. And the general reference librarian is going to, and the special collection is going to be a special collections library. And because there isn't any distinction I used as an example, when I was at Cornell University, we had a microfilm set of in kuna bowls. And they were kept in the general collection with the microfilms that are in the general collection. But that made no sense because the people understating It was good in that it didn't require the kind of special oversight and handling of that were found in the Special Collections Reading Room, it's not valuable materials in that sense. But it also meant you didn't have the expertise of the Special Collections librarians to help anyone who was working with it. And conversely, there might be a collection of Oh, the microfilm of presidential papers, that that were housed in special collections that could just as easily be made accessible in the general collection. So I never understood how to draw those distinctions. And I think it continues to be an issue with the making of America project from the 19th century, it was always an interesting, you know, that it would get to the point where the specialists in the general collections would try to answer the questions immediately, and then refer that to Special Collections if if there was a problem, and it sort of worked, but it was in very much an artificial distinction. The problem is to make sure that everyone is aware of the contexts and strengths and weaknesses of any digitized research resource. I'll give you an example. I know of one historian who is working with early English books online. And I am under the impression did not realize that that was not all 16th and 17th century and 18th century books, but rather just a subset of those books, because that was just one of the online databases available through the the general collections, the library didn't have anyone who would explain the context and limitations and scope of that resource. And if you're trying to make an argument that says that 30% of the books in early English books online, use this word, and 70% don't. And in reality, that resources only 20% of all the available books that might shape and the basis of your argument, you need to know that in terms of being able to understand context, and that's where I worry that especially election, librarians may not understand the scope and nature of online collections. General collection reference librarians may not understand the nature and and scope of Special Collections materials that are available online. And and that's a problem, isn't it?

LM
Yeah, you're putting your finger on a tension between the need for subject expertise. And the need to understand the scope of these services that are that are offered by ProQuest or Gale... How to understand what it is actually that is that is on these collections.

PH

Yep. So we can turn general collections librarians into Special Collections librarians or we can teach special collect collection librarians to understand the scope of the digital the whole universe of materials available, and not just focus on what's in the vault behind the reading room.

LM
This past spring, the Internet Archive got into some trouble for its emergency library. But in general, it seems like it provides a platform for the kind of integration that you're imagining. So what do you think the role of the Internet Archive is in this channel? They're all movement towards digitization. And are there appeals and drawbacks of that site that you think about?

PH
The Internet Archive has done some things that are just fundamental importance. For one getting money to support digitization, special collections, that's really good. And they're experiments with controlled digital eye lending and the National Emergency library are appealing. What I really appreciate about them is their willingness to push copyright law in ways that other institutions can't do or unwilling to do. And by pushing it, perhaps in de facto being able to change it. So those are the big pluses. As a collection, I am actually much more fond of the HathiTrust Digital Library, because it's combining material from so many libraries. So building the shared collection that way. And I guess I'm enough of an archival professional, that I appreciate the underlying library structures that are in the hottie trust that if you have a multi volume work, be it a volume or a serial, it's easier, not easy, but easier to trace the all the copies and the volumes in the Hathi trust than it is in the Internet Archive it one of the other concerns I've had with the Internet Archive is their willingness to take down materials upon complaint. I mean, that's a strength that they will do lots of things that might be pushing the envelope in terms of copyright law, but if anyone objects, poof, the material goes away. And in some cases, I think that can be disadvantageous. And the last thing we have to worry about is the sustainability of all these things. The Internet Archive at Cornell used to do a digital preservation workshop. And one of the exercises that students in the workshop had to do was examine the Internet Archive. And one year, the participants came up with the analogy of the Internet Archive being a gentleman's library. And I think that's an accurate one in the sense that it's been funded so much by Brewster Kahle, his commitment to the issue, but do we have the endowments and funding that will ensure that its holdings continue on forever? So lots of really, really good, exciting things. We're all indebted to it.

LM
That's interesting how you point out that Internet Archive, on the one hand, is willing to venture out on a limb. And on the other hand, when that limb proves sort of wobbly, scurries back and retreats quickly, and it might be because it only has a single funder and doesn't have endowed resources to fight those legal battles, which in the spring promised to be significant when they suspended the single lending structure that they have.

PH
Yep, that's going to going to be really interesting to see how that case plays out. I mean, it's

LM
Are you following the Internet Archive case? And what's happening with it right now?

PH
Oh, of course. Well, they're the briefs have come in. It's still in its very early stages. And don't hold your breath. I mean, after all, the hottie trust case took about 10 years to get settled.

LM
What do you think are then the big issues of the Internet Archive?

PH
Well, there's two issues. One is the National Emergency library itself and whether COVID was enough of a specialized experience to warrant saying that the access it was providing was a fair use or not. And then the second one is the control digital lending overall. And I think it's that latter issue that is going to be more problematic for them.

LM
You think that the general lending structure is going to be problematic for Internet Archive or for problematic for I guess, digitization moving moving forward?

PH
No, I think for the for the law. I mean, thanks to the Google Books and HathiTrust cases, we know that libraries pretty much can digitize things without much worry for preservation purposes. It's when you're trying to provide access to that material, that things get dicey.

LM
That's actually a great transition to what I want to ask you about about copyright. So I think that the first part of our conversation outlined how digitization has benefits and drawbacks, it might include new users, it might allow new kinds of research. And it's been especially accelerated during COVID times. But digitization doesn't equal free access. And so I want I wanted to ask you about some of the ways in which Digital Collections can be restricted. They have been restricted. And maybe they they shouldn't be so restricted. And you write, that there's a confusion between property rights and copyright. And before we talk about some of the critiques you make, I wanted to ask you, when do you think collections can be legitimately restricted? And how should that be done?

PH
Hmm, that is an excellent question. So I think there are collections that may have privacy issues associated with them, or sensibility issues, that when there are donor restrictions on materials, that all makes sense, then that material should be restricted in some fashion. And when I teach about this, I suggest that there can be a range of restrictions based upon the nature of the material that some things you'll just have to sit down and say they're not available or were so controlled that you have to actually come into the repository to do this. I can remember learning

that the Norman Mailer papers at the Ransom Center at Texas, many of the born-digital materials were at the time were on T space servers, and theoretically could be connected to anyone in the world. But you needed to go actually into the reading room and work with the materials in the reading room. Same thing with the Salman Rushdie papers at Emory, they were in electronic form, but you had to be used that way. The other extreme is making it open available.

For something in between, I'm rather fond of the idea of what we call the virtual Reading Room, UC Irvine was one of the first to do that, where people have to agree to the reading room rules that we use in a physical Reading Room have any limitations agree not to release any private information, and then are allowed remote access to that private material. For books, it may be a different set of criteria. I often in teaching workshops, say just as important as what is legal is who's going to get angry so that if you sit down and say I maybe you'll have a collection that is in the public domain, and there's no problem at all with it. But if someone thinks that they have rights in it, or they should have some say over it, and are going to kind of raise a stink, well, then you have to decide if you want to fight them and potentially face a lawsuit over it, the one that you might win, but still is a nasty process, says Kim bursley. We have lots of material that may technically be protected by copyright. But it's extremely unlikely that anyone would ever object, the copyright owners would ever object to the material being online. And if for some reason they do that, you just say Oops, sorry. We'll take it down right away, and hope that no harm has been done. And and so you handle it that way.

LM
You've raised so many important issues there from whether it's better to err on the side of making things accessible with the understanding, you might have to retract some of those items at some point. I want to come back to the idea of the virtual reading room and how that differs from click through agreements or terms and conditions. But to hold on on one of the points you were making about when public domain collections might not be free use. I think, in broad strokes, your work is very convincing both ethically and from the administrative perspective that public domain collections should be freely accessible. But other than situations in which someone might claim copyright or raise a stink about the legal right of a library to put Collections Online. Are there other situations in which you would be cautious about potential cases of misuse and try to guard against that misuse or is the possibility of misuse, just the price of doing business and something libraries have to be prepared to accept.

PH
One of my favorite articles, out of the legal literature has a title something like keeping the world safe from naked chicks on refrigerator magnets isn't quite right. But the idea being that isn't a horrible thing to have a work of art suddenly be turned into kitsch. And I can remember speaking to a museum director once, who said, Oh, wouldn't it be awful if somebody took a painting from one of our from our museum and put it on a billboard with a Nike shoe on it. And we have an obligation to guarantee the integrity of works in our museum. And that would be horrible misuse. And I guess I've just never have seen that for libraries and archives we make the material available, it's important that we serve as a source for the authentic original material that if

somebody is complaining wants to come back, they can check and see what the original was like. But we are not responsible. If another user sits down and crops in in a way we don't like or don't have the colors quite right or misquote from an item or presents an item to make an argument that we disagree with, I don't see that as being our role.

LM
You do raise the interesting counterfactual of in order to track misuse, you have to know where the image came from. That is you have to know that the image came from, from you, as a library, say that a user clicked through the agreement at Cornell and downloaded the image from the Cornell library and then misused it. Whereas the user could have gotten the image from some other place on the internet, and therefore not be liable to the contract of the terms of agreements. But in order to know that you have to track where the users are getting their images, which is something that very reasonably you say libraries don't want to be in the business of doing.

PH
We've been very careful about protecting the ability of researchers to read anonymously. And I think that's an important value for us, it may be helpful here to talk a little bit about what how restrictions actually happen in detail. Let's talk about public domain works. And if you've digitized that and put it online, there is no copyright. So there's no risk of copyright infringement.

LM
I want to step back just for a minute, because I think that this is such a complicated issue. So there's often a confusion between property rights and copyright, which is that Special Collections own the physical object without owning the intellectual property, which could belong to the author or the estate of a deceased author. Or it could be in the public domain. What does it mean for an object to be in the public domain?

PH
Well, it's a it's a real interesting question. And there's there's no clear answer. Some people view the public domain as being the absence of any of the rights of copyright. Other people say that all the rights belong to the public, but probably the simplest way to think about it is that copyright specifies a set of six exclusive rights that belong to the copyright owner. That's the right to reproduce the right to prepare derivative works the right to perform a work publicly a right to display a work publicly, or right to distribute a work. And when a work enters the public domain, the copyright owner no longer has the exclusive right to exercise any of those rights. So anyone may reproduce the work, anyone may prepare a derivative of that particular work, anyone may display the work publicly or distribute the work once it's in the public domain.

LM
And because anyone may do this, this is why the click through agreements are the terms and conditions that libraries use to either guide or restrict digitized images of objects in the public

domain are so controversial, and why a click through agreement is enforceable by contract law and not by copyright law,

PH
right. So if you want to control downstream use, you have to establish a contract with the researcher. And that can be the rules for the use of a collection that people sign when they show up in a physical repository. Or it could be online terms that they have to agree to. And what that would say is that you can get this image for your personal study, scholarship and research. But any commercial use has to require the approval of the repository and potentially a payment of a fee to that repository, or approval or a particular credit line. And I think there are a lot of problems with that particular approach. It's possible to do it. When I first started looking at this, there was some question about whether click through licenses were legal or not, whether they were contracts of adhesion, which are normally looked down upon contracts are supposed to be negotiated between two parties. I think, unfortunately, the consensus now is that yes, you can have a click through license or binding terms on a person that they have to obey. But for precisely that reason that you just pointed out, you may come to me and say, I'm going to want to use this image in your repository, and I'll read your terms for downstream use and pay it and then someone else has that same image and puts it up online for free, I have no way of tracking whether it came from you or not, I have not signed an agreement with them. If they publish the image or, or even sell it, there's nothing I can do, because I don't have a contract with them. And so you, as a good citizen, are limited in what you can do, but someone else is free to do what they want. And I think that my personal take is that rather than a contractual agreement that is in force, if you want to sit down and say, you know, we would like to know about uses, we would prefer that you not use this for commercial purposes, I would respect that. But Google Books used to put a statement like that in the front of their scan books. And the Internet Archive said, Yeah, they're all public domain and just went ahead and ignored that and got copies of them and put them into the Internet Archive, you can sit down and say, Okay, well, they're they're all freely available. Google hadn't created a contract, it was just a request. And even if they had created a contract, there's nothing that they could have done with it about it. So. And the other the other thing about the enforcement is that you have to be willing to enforce one of these contractual agreements. And I had assumed that, in reality, very few libraries and archives would want to take a researcher to court over violating the terms of a researcher agreement until a few couple years ago when when Harvard did it. And it reaffirmed that these agreements can be binding, but I think most of us would not dare do that.

LM 33:06
Right, well, I want to shift to a final set of questions about creating revenue streams from collections or monetizing them. And an underlying tension in libraries is the contradiction between, with some exceptions, the wealth that's held by special collections in the form of their materials, and the often underfunded day to day working in so special collections from shortages of staff that create lengthy backlogs to facilities that aren't adequate for preserving materials. And in general, auctioning of materials is not popular, either for museums or libraries. So rather than selling physical manuscripts or rare books, though, I could imagine situations in

which this makes sense, libraries might try to monetize digital assets. So how is it possible to do that?

PH
Well, there are people who have done it most often by partnering with commercial outfits that then make the ministerial material available on a restricted basis. I can remember reading a report from the American Antiquarian Society that indicated that a substantial part of their annual budget was coming from the licensing fees that they were receiving from redex for their digitized newspaper products and other digitized collections. But there was also the problem that those who are sold to libraries and the number of libraries that can buy these things is rather limited. And so once it's sold, then there's no more licensing fees coming in. So it's pretty tricky to have an ongoing, maybe a one time shot in the arm, but not an ongoing stream of money. By and large, there haven't been many studies that I know of about monetizing the collections; that were done, some done quite a while ago about in museum worlds. And it turned out that even though museums have extensive licensing programs, the ones who were actually making money at it were a relatively small number. When we were at Cornell, we did away with any restrictions on use of public domain material. And that was in part because we realized that boy, the amount of time we were spending, trying to manage permissions deal with permission requests, the kind of paperwork and the billing and everything else was costing more in staff time than we were actually getting back in return. Or if there was a cash return, it was actually a very, very small amount. And I was pleased when Michelle Light spoke about this a few years ago, she's discovered the same thing at Nevada, Las Vegas, where she was working at the time,

LM
Michelle makes a great point that she's at an institution with a lot of potentially commercially viable images, one would think, the Rat Pack or Frank Frank Sinatra in residence. But even there, they did not find usage fees to be a good source of revenue.

And you're, you're increasing your liability when you're doing that. Because Do you have to worry about clear publicity rights for Frank Sinatra was the photograph itself a copyrighted photograph from a news photographer that's still protected by copyright. So you do run a little bit of risk, especially when you're trying to commercialize your materials. So I think it's a much safer approach is down and say, let's just think about value added. And what can we do that makes make our collections available for free, as the public domain material is freely accessible? give everyone a usable copy? But are there other value added ads that we can do that makes for a product that's worth subscribing to on its own? Or has added services that make it valuable? And will the value added make enough to pay for the expensive offering a product like that?

So before, before, I asked you about some of the maybe creative things that you think are viable value added programs that libraries might charge for, just to clearly distinguish, you say that scanning fees, or the fees for reproducing an image should be considered separate from usage fees and usage fees might be charged to someone who wants to republish image in a book that's going to have a commercial market. And in general usage fees don't seem to be to be

worth it. They're just not a good revenue stream. What do you think about scanning fees? Do you think that even charging a fee for reproduction should be reexamined in certain cases?

PH
No, I'm quite comfortable with the idea. That's a kind of value added service that I'm talking about to make a high resolution copy of a photograph that we may be have on as a PDF for medium resolution TIFF online that requires equipment expertise time. And so I think it's quite reasonable to charge for, to doing that. And maybe even incorporating some of the cost of the fact that you've been preserving it over time, then, you know, those kinds of infrastructure cost should being incorporated into that. But recognize that once you let it out, if somebody wants to take that high resolution image, and throw it up onto flicker, there's not much you can do about it. And so then at that point, you might also want to think about whether Well, if it's done, should I be doing that myself? Or should I be putting up my immediate resolution and saying, oh, and I can make high resolution available? One of the interesting questions is, let's say it costs $100 to make a high resolution scan for the first time, and then reproducing that scan and distribute sending it out to a researcher might be $10 if somebody else wants it, well, do you charge $100 to the first and $10 to any subsequent users? Or how do you handle that and that's a problem. Special Collections have faced for an awfully long time. Back from the microfilming days even

LM
right because the the labor required to produce that image is real the in the in the first place. But it doesn't mean that the that the first user should maybe bear that entire burden or or should they?

PH
Or should they? I don't? I do not have I have not come to. I can see arguments on both sides on that one.

LM
Well, something, Peter that you'll have to continue working on. So thank you so much for all of the time today that you've taken to talk with me. And just to end with the last question. Are there any developments or creative projects in special collections that you've seen during COVID that have made you particularly excited?

PH
Well, certainly, you know, the most exciting thing is the expansion of access to materials, I think we've all realized how valuable it is to sit down and be able to look at the vast amount of material that's been digitized and have access to it to the to the full text rather than just being able to do word searches. And I think it should just make us redouble even more our efforts to do everything we can to make sure that our cultural heritage is as freely available as possible.

LM
Thank you so much Peter Hirtle for being on the podcast.

PH
Thank you, Lina. It's been fun.

Transcribed by https://otter.ai