Transcription
Section 108 Study Group, Public Roundtable #1
March 8, 2006, UCLA School of Law, Los Angeles, California

Topic 4: New Website Preservation Exception

Participants
Richard Pearce-Moses, Society of American Archivists
Cynthia Shelton, University of California – Los Angeles
Brewster Kahle/Michele Kimpton, Internet Archive
Sherrie Schmidt, Association of Research Libraries and American Library Association
Patricia Cruse, California Digital Library
Kathleen Bursley, Reed-Elsevier, Inc.
Jared Jussim, Sony Pictures Entertainment

Dick Rudick: Michele, will you introduce yourself?

Michele Kimpton: Sure, I'm Michele Kimpton, Director of the Web Archive at Internet Archive.

Dick Rudick: You know, listening to the last discussion, and recalling histories on how this next offer seemed simple at first, simpler than the other stuff we've been dealing with, we'll see. I remember when I attended my first meeting at the Library of Congress of the National Digital Strategy Advisory Board, one of the things we talked about with most passion -- this was like couple of years ago -- what's on the web? How is this part of our cultural heritage? It may not be the most elegant part of our cultural heritage but it is part of our cultural heritage and what's the archivist job? It's to make the record of what we did and what we didn't do. So, the web is a big piece of our culture and our heritage and think about it, the people who put this stuff up probably, certainly, are not spending a lot of time themselves thinking about how to preserve it, some might even try to destroy it after they thought about it. It's the kind of thing where the only people probably who could preserve it are professional archivists and librarians fulfilling that part of their world and it's a shame it changes so quickly. It's a difficult target for preservation and though it seemed enough in the Study Group that we ought to just look at web sites as an issue for a special provision in section 108, and then a lot of the commercial issues we've been discussing may not adhere fully to the web sites.

OK, let's go right to the question. If you have to say it's more important than that, to say given the above and whatever else you know about web sites, should there be

a special exception in 108 to permit the on-line capture and preservation by libraries and archives of certain web sites or on-line content? And if your answer is yes, what types of limits should be imposed? We talked about some, but we want to hear what you have to say so, how about it? Who would like to speak first? We're going question by question here and so we'd like ... you may think of other questions we should've asked, but as they use to say in the army, "wait for it." So, first question.

Cynthia Shelton: Yeah, to make a very basic and simple statement that I think is a foundation, is that this content on these web pages is not just out there for us to capture and so on for a research library that is now part of our collections. We select web pages, we catalog web pages, we put metadata around them so people can get adequate portals, so preserving this particular digital format seems like a natural progression of our mission. And just to give an example of the point you were making, we have our rights-holders come to us to see if we have copies of their web pages, so they're certainly not doing it, and they're looking to libraries to handle the same problem.

Michele Kimpton: Should the preserving be capturing, preserving web sites? Absolutely. Internet Archive has been doing this for ten years. Luckily Brewster was an innovator in this field because he had the foresight to think, "hey, primary source material is being developed and nobody is saving it." The traditional institutions, libraries and archives, were not storing it at the time so he had the background, the technology to start preserving it. Ten years down the road, 2006, we have sixty billion web pages from fifty million sites around the world. We capture two billion pages a month.

Dick Rudick: It's a lot of culture!

Michele Kimpton: It's a lot of culture. U.S. national archives, we work with U.K. National archives, I can name twelve national libraries that we're teaming with to help them develop systems, tools, and also world archiving services so that they can capture pieces of material that's

important for their cultural heritage. So it's absolutely imperative that we do this, because as I keep going the web pages disappear daily. There's numbers quoted that the average life of a web page is a hundred days, but there are some that disappear in hours, and an interesting small study done by The Center for Research Libraries was looking at a section of sites I prepared for them, about two years ago, and they looked at the material from their sites -- it was a couple hundred million pages over a couple hundred sites -- and they compared it to what was on the web today, and forty percent of those web sites were gone out of a select sample that the Center for Research Libraries was investigating. So it's really important, it's important, we need to do it and you need to do it broadly because you don't know today what's going to be important for researchers tomorrow.

Patricia Cruse: I would agree with everything that Michele says, and I think that for the University of California it's imperative that we're allowed to continue building our collections. There are several depository libraries within the University of California that have collected in paper format the output of the federal government and state government and local government; now those publications are primarily produced on line. For the University of California, if we don't continue to collect those materials in a web based world, we simply cease to be a collecting institution and we're just not building our collections any more, which is silly. Now, that's one example of the way that we need to continue collecting every single day, going out and collecting the output of the government and things like that.

But also a different flavor is collecting events, for example Hurricane Katrina, when there's something that happens out there that's really vitally important for researchers but also for the community at large, citizens, to capture the historical record of these types of events so these two things kind of overlap but they're a lot the same. You need to be able to go out and capture at will, and not have your hands tied and saying, "oh, you can't go out and capture that stuff" because it has a robot exclusion on it or you didn't ask permission to capture it.

You just need to be able to do it and build your collections.

Jared Jussim: These all sound good: "I'm going to preserve culture." I'm going to point out that when our great neighbors to the north went to have a conference on culture, it managed to invite every country in the world except the United States because we didn't have a culture minister, and they came up with a treaty basically that was designed to rob American product, so I add because one of the reasons for culture ministers is because they worry about loosing their culture. I don't worry about things like that because I know we have a culture, I know it's out there and lives by itself. It lives because people produce it.

Now, I have really no problem with normally going on the web and capturing whatever images you want and preserving them; it's the other side of the coin which contains copyright material that's protected and you make it accessible to people, that's my problem. I don't want, you know, somebody goes on the web and steals *Lawrence of Arabia* for example, puts it up there. Don't tell me you have to go down there and preserve to see what pirates are doing stealing our work. By the way, I'll give you plenty of examples, but if there's something else, by all means. And there's nothing wrong in looking but you know, even *Lawrence of Arabia*, although I don't want you doing it, and if you touch any Disney work I'll kill you. But having said that, you know, it's obviously a protected work and I couldn't keep your preservation on the fence, but my real problem is when you make it accessible and what the restrictions are.

Dick Rudick: Remembering the Federal Register notice, there are actually a bunch of questions here, it says the access does come later. We're focusing now on whether there should be a general exception, and here's the critical thing which we need some discussion on: If so, are there any limits? If I can ask to follow the question as we go around and continue to discuss this. Should this apply to everything, and if not everything, what wouldn't get covered? So we should cover that as well as the desirability of doing it.

Topic 4: New Website Preservation Exception

Mary Rasenberger: I mean the types of considerations that we put in the notice and background paper are, for instance, how do you carve out the types of things that Jared is talking about, commercial content or content that's itself an object of commerce. Which is kind of an awkward way of saying it, but any help you can give us in how you would say that or define it, for instance, media sites, New York Times, CNN on-line, where the information is what they're selling. Are these the kinds of things that should be carved out and if so, how?

Kathleen Bursley: I guess I remember the days when libraries built collections by buying things. I know that's in the past but I'm trying to put myself in the shoes of somebody who does such content publishing, essays or something, and has a collection of them and sells them, more the New York Times, which on its site has access, I think for seven days and free access if you're a subscriber to the New York Times for another month, and then you pay to download articles in general, though they do leave some things out, continuing interests or continuing stories. On the one hand, I think well, nobody is going to go to the University of California Library to try and find this when they can go to the New York Times site and pay five dollars and have it there because it's going to take them forever to find it in the university library and here's the article right here and you know it's the final version of it and so on and so forth.

I guess my real question on this is: Is this preservation? It sounds like copyrightable compilation of copyrightable works, and before when we were talking about the copy that we distinguished between preservation and access and obviously we couldn't come to any conclusions about that, there did seem to be a clear distinction between making the copy and then allowing access. But it sounds like, at least in the case of the two we're mentioning, that the collections are being built precisely to allow access -- not just to have a copy somewhere in case the New York Times runs down, but actually to make the collection available. I just wonder how that's preservation.

Topic 4: New Website Preservation Exception

Mary Rasenberger: The Library of Congress has done some event-based collections: The elections, 2004 election, on the events on Katrina, on the war in Iraq and a number of things like that. And we do not in the Library capture sites like the New York Times or media sites, unless we work out a special arrangement with them, for instance with Katrina, the Times-Picayune asked us to capture and preserve everything that they were doing in the days and months after Katrina, it includes things off the profit sites.

Lolly Gasaway: Lots of it that came down.

Mary Rasenberger: Right, these are things that by and large do not stay up there very long.

Patricia Cruse: Let me just add that if you don't have it, you can't preserve it, whether it's the original producer it disappears off the site. You have to get in order to preserve it.

Brewster Kahle: In terms of how to encourage collections, which is our theme, we get to talk about, what do we collect? Most of the Internet Archive's holdings come from a donation from an organization called Alexa Internet, a for-profit company, and it's donated with a six month time delay, and so basically it's not life plus seventy, actually, it's six months, which I think is kind of interesting. It's an object commercial edge and they've been perfectly happy with us for ten years. We'd like to encourage more donations but there hasn't been a flood in the last ten years of anybody else, in fact we tried very hard when "Go" went under, that was a Disney-bought Infoseek, which was one of the great search engines of the 1990s, and when that was going down, we couldn't get anybody on that site to feel comfortable donating a collection. We've also lost out on other collections: Apple Computer had all the CD ROMs that were done with "Quick Time" which is a technology for making CD ROMs back before the on-line thing and, hey, decided to destroy fifteen thousand CDs rather than risk donating those to a library or an archive. We talk with Google all the time, "why don't you donate your stuff"? Well, you know, you

don't have deep pockets, we do. I mean, it doesn't matter if you indemnify us, what's that going to be? So, if we can make it easier for digital materials to be donated to a library or archive, that will be helpful.

I would emphasize your point of going to say on-line rather than web because things change all the time, web is pretty much the finest, h.t.t.p., but there's also r.t.s.p., there's blogs, there's usenet, this is a whole cluster of things that most people kind of think of as "The Web," but I would say online resources if possible without getting into too much trouble. Then, there's the issues of do not collect certain things and robots and I think that's a different part of your question.

Dick Rudick: Just for a clarification, you mention liability which wasn't something that was on our radar screen and it may not be something that we can deal with in 108, but just for education what's that all about?

Brewster Kahle: We would like organizations that are collecting materials for some other commercial purpose that they don't own. Take the search engine, they're collecting web stuff, they've got great engineers working away on collecting this stuff, they think it's valuable but they're really hesitant to make a copy of that for an archive or library because somebody in the organization says, "let's wait," and if you wait, it's gone. So, when we've been able to get this to happen once, it's because I've created the company.

Richard Pearce-Moses: I don't have my notes in front of me, but the National Historic Publications Commission, the predecessor for the National Historical Preservations and Records Commission, is I think in some ways a model for this, for why we need this and it goes back to us, I think, at least those of us who are working in this area say, "if my peer preserves" model: If you don't put it up while it's fresh, it's gone. I think the N.H.P.C. model is really valid again because that was established to publish manuscripts that existed in unique copies. It was an attempt to preserve by making sure that those materials were distributed so in case of fire, flood, Katrina, the original was not lost.

Topic 4: New Website Preservation Exception

I think if you look at an awful lot of web content, especially a lot of things that archivists want to collect, it falls within that similar model that there is a unique copy. These are done by people who are not the traditional commercial vendors of information, do not have robots disaster recovery copies or preservation plans, and in fact archives often collect things that are not for sale so this does distinguish us from the people who go out and buy. We go for diaries of ephemeral manuscripts, snapshots, home movies, and that stuff that's on the web, and can disappear. So, I think we need to look at, there is a real difference between many people who publish on the web and many people who are really individuals who are not publishers.

And one of the things I'm very interested in and can't wait until I can play in this area are political web sites, something that is of critical importance to our history and I would say this goes beyond one fussy culture, this goes to the preservation of democracy. I want to make sure that we can start capturing political campaigns so that we can document elections. The campaign officials, the web masters, they are trying to get their candidate elected, they are not worried about record-keeping and I think that the National Archive studies of record-keeping in the federal government probably is analogous here. The only agencies in the federal government that routinely have good record keeping programs are those who get sued a lot and those who get lots of lawyer requests. Without those two pressure points, they don't have the need to do good record keeping and stuff gets lost. I think the same thing may happen here unless the individuals have some sort of particular pressure either from litigation, criminal interest, or commercial interest. They're not likely to do a particularly good job preserving their web information, so I really believe that this is important.

I would like to address a couple of things that we believe help measure this a bit. We think that in some ways our emphasis in collecting should probably be things that are freely distributed on the web. That does not mean without registration; if anybody can register at no cost, that site might be included. I'm with Patricia as I understand that we don't necessarily feel that a robots.txt is necessarily the best possible thing because libraries

and archives . . . Oh! Hold up my comment on robots.txt. I will end by saying that at one point many people described a lot of web pages as glass brochures. It is fully legal for us to collect paper brochures but we cannot clearly collect glass brochures.

Michele Kimpton: OK, I was going to address what came up before the commercial sites and just to talk a little bit about how we handle that, because obviously we do a range of archiving from very broad, to very site specific, to collection specific, to event specific, to government specific, we do all this whole range for ourselves and other partners that we work with for our broad archive, where there is no way we can go and contact fifty million web site owners; it is not practical and it is not a useful use of our time. We have found what is called the Oakland Archive Policy and I have copies here if you guys would like to take them with you. Basically it lays out for different situations for a site that is a commercial site or doesn't want to be in the Archive or accessed, what they can do prevent a crawler from archiving their site or to prevent access in the Archive, and we've been following this policy since 2002 when it was created -- actually in 2001 -- and it's been very successful. In fact very few, in fact no instances of getting sued or negative, we have very quick takedown of sites if somebody contacts us and says, "hey, we don't want our site in the archive." We quickly remove access and then no further putting on the Archive beyond that point of contact, and that's been successful for us.

Dick Rudick: How was this created?

Michele Kimpton: The Oakland Archive Policy is created by a committee of folks from Berkley Law, from Internet Archive, from . . .

Brewster Kahle: From the top lawyers from the search engine companies, Margaret Hedstrom, Abby Smith, library, archives, and search engine worlds, we sort of tried to come up with the policy that reflects kind of what it is we've all been doing, and then we've been trying to just live by it and it's been incredibly helpful. Not that you

want to cast this in law, it's a definition of what it is we've done and how we can try to balance the issues. When do you take things out of the Wayback Machine? When do you take things off the public access? How do you do it with third party requests, when somebody goes and asks us to take things out from somebody else's web site, how are they pulled apart and when do we use Chilling Effects as a mechanism of bringing light to other people's DMCA requests?

Jared Jussim: My comment was because I wanted it direct if you will defend people who don't want to make copies available and point out that we're dealing here with 17 USC, which is a copyright law, but fortunately or unfortunately depending on your perspective, global laws which are applicable, some of them at state law, such as rights of privacy, rights of publicity, so that when you put up an image, it's not just the image that may be available for that moment, it may be the use of that particular talent in that particular place so it may be at the moment, that this is news and you can copy it but two years, three years, five years down the road, it may be just you're using it for publicity, advertising and if that's your business, meaning you're an actor or an actress, that's what you sold, it's your face and you don't mind it being used but you want people to pay you for it.

To give you a case from Germany recently, came down, I think this week, there was this charming gentleman who advertised on the net that he wanted to eat somebody, I mean physically eat somebody, a cannibal, and he got a willing victim, killed him and ate him. I don't make this up, now, he's sitting in jail and they wanted to run a story about him, this is about five years it happened ago, and he got an . . .

Dick Rudick: We're focusing here on cons and pros of something to permit web preservation, not trying to solve all the problems. We're talking about copyright law and we're talking about whatever concerns people have against the need for preservation material which in many cases, not all, this is what we need to discuss because we need to know what the limits should be, if any. In many cases people want this stuff to be copied and preserved; it's

almost like there's some sort of implied license, so if this a sociably desirable goal, so that we don't accidentally do the wrong thing, should there be any limits? What limits are acceptable?

Jared Jussim: You cannot pass the law so that it overrides state law or other rules.

Dick Rudick: No, well, It's not our problem.

Jared Jussim: You know, the problem is to direct a law that does that.

Lolly Gasaway: Our problem is not to direct the law at all; we make recommendations to the Register and then we're done.

Jared Jussim: Somebody's problem is to make sure . . .

Lolly Gasaway: Not ours.

Dick Rudick: There are enough copyright issues; for our purposes we have enough problems as it is and we're going to let somebody else talk.

Patricia Cruse: I just wanted to say again the great work that the Internet Archive has done in crafting a policy, that I think it really balances the rights of us libraries who want to collect, and that of rights-holders too. And when we crafted our policy for rights protocol for web archiving, and we really looked to their work and modeled our stuff on what they did, and so far we think it holds the task and that is good and I'm glad that they shared that with you.

Lolly Gasaway: Several of you began to talk about robots.txt, but we need to make it more broader than that and say, should copyright owners of on-line content that's been captured for preservation be able to opt out? That's our real question, should we provide an opt-out mechanism? And if so, how? And then that's where we can talk about robots.txt or something. Should content providers be able to opt out?

Topic 4: New Website Preservation Exception

Mary Rasenberger: There's a two-part question, to actually be able to opt out at the moment of being crawled, and to opt out from being preserved in the archive and made available.

Brewster Kahle: There's three.

Mary Rasenberger: You're right, there's three, thank you.

Richard Pearce-Moses: Just a short thing. I would refer to someone who knows a lot more about robots.txt files as to whether or not we can say in such a file don't crawl me vs. don't preserve me. If there were such a thing that was technologically available, it would be interesting. As one of the web masters at my agency, the robots.txt is my way to do that in some ways. I don't know how to say in that file, "you could crawl me but you need to call me and ask me when, as this brought my server down once, and so it's not that I'm saying don't crawl, you can't index me. It's don't crawl me because you'll crash my server the way you are currently configured." So, I think that is one of the things we need to do. It's a great idea to include that in a robots.txt, but is it technologically feasible?

I also think there should be a discussion, and I am of a mind to say that libraries and archives have a particular social mandate to preserve information to build collections, and so I think we may want to see discussion that libraries and archives be allowed to carefully -- possibly with some hand slapping if you crash their servers, speaking from first-hand experience -- but to carefully crawl things while disrespecting robots.txt files.

Patricia Cruse: Just a little bit on robots.txt key. I think if content is posted as publicly accessible, archiving institutions should be able to capture it. Currently, robot.txt key files and the like are not used consistently. For example, some sites that are clearly in the public domain have federal robots.txt files and this is a White House robots.txt file, so what that is telling me is that I can go and capture the State of the Union Address, the Easter Egg Roll, or the T-Ball stats. So I

think that robots.txt files are also completely silent on the issue of whether content may be archived by a human being, and what I mean by that is that if I'm sitting at my browser, I can say "save as," the robot.txt file doesn't come up and say "no, you can't save as," so I think it's not consistent and it just doesn't work. It's a good idea on paper but it's not implementable. So, also often robots.txt files are put out without the understanding of the content folder. I don't think a system administrator will go to somebody and say, "I'm going to put up a content robots.txt files so your content is protected." I don't think that's how they're used so . . .

Michele Kimpton: Sorry, but a couple points on the technology. In terms of bring servers down -- and we deal with this a lot because we do crawling for all types of partners as well as taking donations from Alexa. When we go out and crawl, we identify ourselves to the server so the server says, "hi! I'm Internet Archive and I'm crawling your site." And if we have an impact on the server, the web master can look in their web logs at that time and they will have e-mail and contact details on how to get a hold of us. And we can dynamically change the speed on which our crawler captures and downloads web pages from their site, and we've done this consistently and effectively through working with partners and web site owners and it seems to work OK. So it's, you know, being there and this is a fact that's actually been followed by search engines as well. They have what is called the user agent, they identify themselves and so the web master can get in contact with them.

Regarding robots.txt I think I agree with Patricia in that we have found situations where we're misunderstood. So many times system administrators, web site producers will put robots.txt on images and pieces of content that are intensive, that are really going to take a lot of their resources to download, and these might be really important documents or images to the site; and it's not something that they try to copyright protect, it's just they don't want to have it downloaded from their server. So sometimes there's not a misuse to protect copyright material, but then there are the times that it is used to protect copyright material in terms of the implementation of the

Oakland Archive Policy. That's how we tell the site owners that we're working with, OK, if you don't want your site crawled in the future that's fine, just put a robots.txt file with a disallow our archive agent and we will not touch your site, so I think if it's well communicated we can change the implementation of that robots.txt file; we've been able to do that in the broad archive.

Kathleen Bursley: Just so it gets said I think an opt out provision is essential, some type of opt out, whether it's, "No, I didn't want you to take that, take it off," or, "no you can't crawl me at all," or whatever it is. I just think that for privacy reasons, for American way reasons, you should be able to say, "no, don't do it." And that's commercial or non-commercial I think, and in some ways maybe non-commercial even more. I have wondered while I was thinking about this whether there would be -- it sounds like from your description of what happens when something does arrive with an impact on the server from what you're doing to capture stuff, that you can then tell the crawler, "no, don't do it that way, do it this way," and then it will go and be OK. So it sounds like non-digital communications maybe an effective way to either get yourself off the archive or say, "no, don't crawl me this way because you're crashing my server." I really wonder whether the only effective opt out would be technological in nature, and it sounds like maybe it would be worth allowing for just, "Dear Sir, please don't crawl me. Sincerely yours." It might be worth allowing for a non-technological method of opting out.

Mary Rasenberger: I had a couple of follow-up questions to some ideas on the table. We talked about allowing opt out, what another proposal was: Should there be an opt in? Some kind of mark up to robots.txt that you could somehow put on the sites saying, "yes, crawl me, I want to be crawled. I want you to preserve me." So, from your reactions, it's a bad idea, in fact that's something that we won't work on.

And then, another question that we touched on a little bit but that I would like to see if we can specifically address this, is the concern that if you have an exception like this that's available to anybody that's crawling, how would you prevent multiple crawls at the same time in the

part that you were leading to Richard, which is, you know, causing interference with the operations of the web site.

Brewster Kahle: Let me see if I can meet a couple of your direct questions. I think it could work specially in particular areas. We're hoping that more applications on the Internet become archive- aware. In other words, the web came about before archives were common to the newer protocols. In terms of melting other people's web sites, the Internet Archive crawls at about a period of five seconds per hit, and it's really hard on modern computers to have that be a big problem; it's five seconds even between the last bite we got before another request is made. So, it's only when you're in early development that you just melt people, and it happens but I don't think we have that issue much any more. The Library of Congress has a robots.txt exclusion that says "go away" and this makes no sense to me given that it's a congressional record, and when I talked to the web master that originally put it on, it was because somebody had been a bad egg once. What the Internet Archive does with the robots.txt exclusion is that practically and retroactively applies to the public access materials, so somebody goes and puts up a robots.txt exclusion, we retroactively mask that from public view so it's taken out of public view and it's really effectively gone by putting up robot exclusions. Now the robot exclusion has had a difficulty when a domain name changes hands, so often a domain name changes hands, will basically expire for somebody and it will be picked up by a cyber-squatter and they'll just put a robot exclusion on it and then we get angry notes from the original site owner saying, "why isn't my site on?" So the area of robot exclusions has some particular issues.

On the preservation copies, yes, there might be some privacy issues to come up in the future and there's issues about how do we take those down, and how do we protect our archives from getting drilled too much. One unexpected consequence, a side effect of building a web collection is: We are the lawyers' friends. We get a lot of requests these days for lawyers using the stuff to sue others. Another is, we're a friend of the United States Patent and Trademark Office, and that was not our goal in life, and, you know sometimes, we'd like to make clear that we're just

a library so that we don't actually have to certify that copies came from us and things like that, it's getting smoothed out but is quite active.

Richard Pearce-Moses: In terms of something like the robots.txt file, I would always love explicit permission, there's nothing better than that. However, realistically I just don't think we'll see people doing this. And also on the issue of being crashed, I would say that systems administrators need to take some responsibility for managing that. Those of you who are technologists will tell me if we can build a gateway that will start refusing requests from the same IP after a period of time. But more importantly, as thieves got through because we didn't have a robots.txt, so I fixed that, we had just operated the server that didn't come standard and I forgot. So, I'm not really terribly worried about being hammered.

Patricia Cruse: I would just want to say I would welcome some sort of policy that will help us collect information that is very difficult, things that are for example, on the GPO Access, Government Printing Office site are almost impossible to collect because in the manner they're arranged on the site, it's not a robots.txt, but if there was something that said, "oh, robot, here you go, and here's how you get to our content and go at it so."

Kathleen Bursley: I was just wanting to confirm that I hope everybody has in mid that opt in will not be instead of opt out but will be in addition to opt out. And I think opt in sounds like a great idea if you could do it, if it could be implemented, because I'm sure there are other sites that for whatever reason, you would want to archive material off of and they'd want you to.

Michele Kimpton: Opt in would be great, just so you have that plus opt out. But the reality is, we did a Presidential web harvest for the U.S. National Archives, which you think would be a fairly authoritative archive, and they could not get compliance with the thirteen hundred government web sites that they wanted to have in this archive. They tried to make sure that they were part of the Federal Register, and just to validate that this was

the right part of the web site, and I think they ended at a couple hundred responses for their request for being archived, and then they just said, "Screw it! Here's the list." And so you know, even at a what I consider the Pentagon of archiving, they couldn't pull it off, so practicalities kind of rule the day in terms of what's achievable.

Brewster Kahle: Australia started by asking permission of everybody, and the Royal Library at Sweden went for it, just do it, opt in vs. opt out, and now even Australia has moved over to the broad web crawls. It is the way to build archives and it doesn't seem to cause any damage in the last ten years.

Lolly Gasaway: Is the problem identifying the right person within each organization?

Michele Kimpton: That's part of it, identifying the right person and then the size, the numbers of people that you have to contact. In Australia, they were doing fifteen hundred web sites, and they were going out and actively getting permissions, and they were spending a tremendous amount of money to do that. And so then just recently they said OK, we're going to just do a snapshot of all the Australian domain. Which we did for them as an agent, and they captured two hundred million documents in this capture, and they went for an opt out, not an opt in, but a robots.txt policy so if it was robots.txt blocked, we didn't crawl it, so they didn't conflict anybody.

Patricia Cruse: I just might add that a lot of times people will contract out to another host, so that they're really far removed from their content, and so to say to get in touch the web master, they have no idea on how to get in touch with the people who have provided that information.

Brewster Kahle: Here's another piece of data: When we tried to connect, contact the web master, this was early on to 1997, we sent e-mail to a bunch of web masters, we were put on black hold lists immediately. The reaction was negative.

Topic 4: New Website Preservation Exception

Lolly Gasaway: OK, let's talk about access to this stuff that we have. Should there be any restriction on the access to what we capture and if so, what to make, I mean, should access be for research purposes only, or for news purposes only?

Brewster Kahle: Open archive policy is how we regulate access in a "terms of service" on the Internet Archive site that basically says this material is for research purposes only. If people contact us we say we're a library, we're allowing them to see these materials, they do not own them in the same way that you might own them from the original site. We are just a library giving access to this material. That seems to keep people enough so they don't go forward without permission.

Cynthia Shelton: Back to the point I made in the first discussion, that the end of preservation is access, so we have to take the long view. We want to be able to capture and preserve this material because it is so ephemeral, and I don't know if this statistic was turned out, but web pages have an average life span of a hundred days. So it's already been said if there's nothing there to preserve we won't be giving the access to it at all, so we are preserving it in order to give access to it.

Lolly Gasaway: With no restrictions on that access or just in general?

Cynthia Shelton: Well, I think that since these are publicly available web sites we're talking about, we're not talking about getting into general content and commercially available material; we're talking about material that is copyable, available, and that's going to disappear, and we need to preserve it and give access for research and teaching purposes.

Michele Kimpton: Nowadays we don't really know what people are doing when they're using this material, and what we do is we simulate the experience as if they're browsing the web the way it was, and so you want to be able to go from site to site as it was in 1996 or whatever point you're starting at. And so if you only make parts of the web

available, or only make it available on certain premises you can change that experience dramatically, which then changes the whole value of preserving web content as a whole, because many of these research fields are looking at linkages and references in and out of a site and to crawl to a set of sites, and it's really that piece of it which is unique to the web environment.

Sherrie Schmidt: I just wanted to say that I couldn't say it better than Cynthia but I really think that that's why we want to preserve those publicly available web sites -- to provide access.

Kathleen Bursley: I think I'm hearing two different types of collections that are being built, if you wanted to say it that way. The Internet Archive-type collection which is, "this is our thing on the web on June 6th 1998, and this is what was there and you can go through it," and then the library-type collection, which is in effect a compiling of all kinds of material on a particular topic, and I guess indexing them or somehow making it available like the Hurricane Katrina thing in the Library of Congress or whatever it might be. Am I correct? Am I hearing that right?

Patricia Cruse: Well, I think it's in addition to that. For example, in California there's something called the California Statistical Abstract that you think would be a very basic thing that the state agency would keep up on their site, every time the new one comes up, they take the old one off and who knows where it is so we want to be able to get the 1995, 96, 97, and 98 so you can look and say, Oh! you know L. A. County is growing, or something like that, but it's not as simple as just crawling for events spaces. It's really in addition to that, it is continuing to build our collections in a digital environment so they remain meaningful to our researchers.

Kathleen Bursley: Would you store the, I understand that you keep the back versions so that you have information from fifty years ago, but would you store the current version? Or do people just go to the web site and get it there?

Topic 4: New Website Preservation Exception

Patricia Curse: We would probably have the current version to so that our users wouldn't have to go here and there, here and there, but if that version is changed we want to continue to get it, you don't know when a version is going to be changed, so it's better to get it many times.

Brewster Kahle: Some experiences on our usage: We find in practice that it does not seem to substitute for the live web, so people don't go to the Wayback Machine to try to avoid something on the live web, or to get to an access that they weren't granted or something, in general. People are using this stuff to find their own stuff so they're not trying to get New York Times stuff they don't have to pay for. We get about ten million page hits a day, which makes us about the one hundredth most popular web site of all web sites. We get about a hundred to two hundred thousand users a day.

One thing to answer Kathleen's issue, we are very responsible to remove a request that either came in electronically, robots, e-mail, etc. At the Internet Archive we want to encounter people right away because we find that we can help resolve issues. And we only, in general we make available publicly available web sites that would've been available without any fee.

Kathleen Bursley: Would you publish or make available to the general public your collection of what ever it is Katrina or the election or . . . I'm thinking more of the event base or the topic base rather than the statistical abstract.

Patricia Cruse: I think that the materials are clearly in the public domain of course, and materials that were publicly accessible at the time is something that's clearly in the content provider sector, we would seek permission from them. CNN, you know, we gathered a bunch of your sites related to Katrina, we'd like to be able to present those for educational purposes.

Kathleen Bursley: So educational and research purposes is OK by you?

Topic 4: New Website Preservation Exception

Patricia Curse: Yes, I don't know if it's OK with I.A.

Brewster Kahle: Yes, we give access to anybody coming not based on user community and it's for free, no adds. Defining non-commercial access has been one of our surviving characters. If we had ads on there, I think people would be a lot more upset, so there's a case for no commercial gain, and we're non profit.

Richard Pierce Moses: There's one other thing that I think it needs to at least be mentioned that I, I haven't thought this through but that is litigation, besides researching in scholarship, in a sense that an individual has seen something on the web site and they feel that the rights, the entitlements, the agreement they entered into on that web site, has been violated in some way, and they may want to find the copy of that web site in litigation against the company providing services, and I think there are some legal uses for this information.

And Lolly you said something that's not quite germane but I think it does very much point to the preservation project about, "just download it off the web site." Three weeks after Katrina hit, I was on the gulf coast, people kept telling people to download it off the web site. If we had paper copies and libraries, we could've helped. There is a real world need to be able to get to some of this stuff sometimes and these sometimes are the sources of last resort.

Lolly Gasaway: Well. Thank you all very much for gathering this afternoon, but for everyone who participated today, you've given us much to think about and we will take your comments, your considerate reflections, and they'll be used by the Study Group to formulate our recommendations.

Dick Rudick: Just to remind you, written comments are on record and someone will read them all. Listening to this today, the most effective comments are not the ones that say, "give me everything and don't give them anything"; they are the ones that say, "I would like to have everything but if I can't have everything, this is more important than that," Or, "If you're going to give them that, these are my concerns in publishing, you should think

about them." Because otherwise we're going to have to guess and we're going to have to make the best presumption based on the knowledge we have within the group. So I urge you to think not only about your wildest dreams but about the practical compromises that you inevitably have to make in any useful possible piece of legislation.

Mary Rasenberger: We will be having a transcript done which will be made available on the web site which is: www.loc.gov/section108. I can't tell you exactly when, but as soon as we can get the transcript done and put it out we will, and the same is true for the Washington roundtable, which is a week from today. I also want to thank all of you for participating today, I mean, this has really been informative and helpful and we really appreciate your time and look forward to receiving your written comments. I'd also like to say thank you to Dick and Lolly, they've been outstanding!