Born-digital news preservation in perspective

Clifford Lynch, Coalition for Networked Information | Oct. 13, 2016 | Charles E. Young Research Library, UCLA

CLIFFORD LYNCH: ***[00:09]*** I'll ask a little indulgence because I'm going to start in a place that may sound strange. It's not with the news at all. It's with the world of scholarly academic journals and the challenge that we face there. Same challenge: how do you preserve this body of information. However, I would say that unlike the news where the problem has really been approached piecemeal, the community of libraries, publishers, authors and readers all came together around the scholarly journal. And started really working on that in a much more systematic way.

[00:58] This has been going on now for a good solid decade. And recently people have been starting to try and understand what's working and what's not. And I thought I would start by teasing out a few of these lessons because I think they're helpful. So, first off, there is a shared consensus among all players that preserving the record of scholarly journal publication is essential.

[01:29] You can't attract readers. You can't do stable citation. Nobody will publish in these if you can't tell people some kind of story that is convincing about how their work will be preserved. Nobody wants their scholarship to be ephemeral. Furthermore, there's a recognition that while the way delivery models have shifted, they're so that the primary responsibility for the active archive in most cases is sitting with the publisher rather than the library community. There must be some kind of fallback external system in place that will allow content to survive the failure of the publisher and the publisher's archive for whatever reason — business, technical or otherwise. So, with that a number of collaborative services were established; you know some of them if you know this area: LOCKSS and CLOCKSS, Portico, several national libraries notably the National Library of the Netherlands; the KB stepped up to this.

[02:47] The funding models on these were usually collaborative. The memory organizations put in some money, the publishers put in some money. Preserving this was recognized as a joint investment, supporting these kinds of large-scale archiving services. And I think that's an important one, especially when we think about the news. You know it used to be libraries actually paid for the print news in archival form. They actually acquired it and that was a revenue stream. Of course, with the move to broadcast in the web that got very murky. But there's an interesting set of points there.

[03:35] Next case. There's a wonderful, wonderful service that is based at the University of Edinburgh, at the Edina center there, called the Keepers Registry. This is turning out to be an essential piece of infrastructure for understanding what's going on around the systematic preservation of scholarly journals. Basically, there are a set of I believe it's currently 12 services like LOCKSS and CLOCKSS, like Portico, some national libraries, that are known as keepers. And what they do is they register their coverage into the keeper's registry. So, you can see how many keepers are preserving a given journal; is it relying on one external service or is it very, very well redundantly preserved by six. And then you can ask questions like do we really need six.

[04:47] You can ask questions like, we can make some estimate by nation of how many scholarly journals are published at a given time. And then you can say what percentage of the coverage of this publication group is covered? Are we doing better this year than last year? Can we make a case that with this kind of an increment in funding we can do a whole lot better? All of a sudden you begin to

get a lens onto the entire area of that kind of publication.

[05:15] Also, you can start doing some analysis of what's being covered. And it turns out we learn some very, very interesting and disturbing things when we start doing that analysis. If you know the world of scholarly publishing at all you know that, particularly in the sciences, it's dominated by a very small number of large players. Elsevier, Wiley Nature Springer or Springer Nature or wherever they're calling themselves this week, etc.

So, these players have a lot of resources, a lot of money, and they're very attractive to a new keeper service because if you can do a deal with Elsevier you can immediately list thousands of journals as part of your coverage, so you can get numbers really quick. So, guess what: Elsevier is really, really well covered. Well you know something — if I was thinking of things to worry about, the disappearance of the corpus of Elsevier journals wouldn't have been top on my list even with zero keepers services because they are very well resourced, they're very meticulous, they have redundant data centers all over the world. They are very much professionals at how to run a data operation.

[06:45] Well, the thing you learn is that it's the smaller ones, the ones that are really at risk that aren't covered or only covered by one service because they're expensive to cover. You have to deal with an input stream and cutting deals just to get one or two titles into your keepers service to deal with those folks. And a huge challenge right now as we try and increase the coverage in this journal area, is trying to come up with common mechanisms so that you can scale a lot less inexpensively. The small players have as much of a problem on the publisher's side as we have on the library — on the keeper side. It's expensive for them because they don't have a lot of technical capability to develop feeds into the keepers and it's expensive for the keepers to negotiate these onesie-twosie deals.

[07:46] The other thing you learn, which is also disturbing, is that there's been an enormous growth of open-access journals. And a lot of these are very small operations done on a shoestring; they're using common platforms like the open journal system; they're run by a few faculty as a labor of love. There is very little resource there. Now, there are some other open-access journals that are funded out of things like author fees; think of something like the Public Library of Science that are large enterprises and they do just fine, they are well covered. They act like major publishers but these little folks, and especially the open access little folks, nobody's looking after. They are very poorly covered and indeed the libraries tend not to reach out for them and advocate for their coverage because they're not paying for that content in the first place; it belongs to everyone and no one.

[08:51] So, we need to be very mindful of those kinds of dynamics as we think about what to do about strategies for really handling the digital news at scale. And let me turn now to the digital news with a little bit of that background in mind.

[09:12] The first thing I would say is that we need to start slicing up the universe here a little bit because I think there are some very different things that are going on. You have a small number of very large players, and everybody knows who those are. If we were talking financial marketplaces these would be called systemically important financial institutions. In the financial industry, that means too big to fail and that means that you get regulated heavily, and people are very nervous about what you're doing.

[09:56] There are a whole lot of other news outlets of various kinds. Small regional things; they use a whole lot of different business models, many of them have very limited resources and technical capabilities. And by the way, I would just like to put in a pitch to remind folks that there's been a lot of talk about rural and regional newspapers, small, small newspapers that are very geographically focused; but there's a ton of other small news platforms out there that handle everything from immigrant diasporas, you know communities of people who came from foreign nations that are scattered to the winds across the world, all the way through subject verticals. All of these things that follow specific subjects, industries, businesses; all of those are also very important news coverage sources.

They complement and only minimally overlap these regional kinds of things, and it's important not to forget those. Although it's trickier to think about, again unless you take a systemic lens, who should take responsibility for those. It's an easy spot to miss. Whereas a local public library or a local research library can sort of naturally say, "OK, we have some responsibility to our hometown or our home region or state," things like that.

[11:35] So, I think it's very important to decompose the problem a little bit and to recognize that the very high-end folks — who by the way also have monetized their archives in a pretty serious way, usually — they're operating in a very different universe than this large number of smaller news sources. And I think rather different strategies are needed for the two.

[12:07] I think we also need to be very cautious about news boundaries. And this is a phenomenon that I think has really changed radically in the past 10 or 15 years. It used to be that journalists summarized reports, hearings, events, that almost nobody else could attend or inspect without a huge amount of trouble and expense. Now in many, many cases, the journalism is built on top of and links to underlying evidence which at least in the short term is readily inspectable by anyone clicking on a link.

[12:54] The report from that commission or a public interest group, the interview, the city council the hearing, the press conference, the film from Space-X of their latest rocket launch; endless stuff.

And as we know, those links deteriorate over time, and the material on the other end of those links often goes away. It's not necessarily in scope for what we're thinking about when we talk about archiving the news sites, and we need to think very carefully about where we want to draw those lines. Recognizing this in the context of the fact that the things on the other end of those pointers often aren't archived by anything else. We have this mythology that the Internet Archive archives the web. It archives that surface web it, doesn't archive all PDFs and films and things like that that are underneath those sites. So, while it's a critical resource it's not a total solution to the problem. We need to think very carefully about these boundaries.

[14:20] At the same time, it's a huge advance just as it is in scholarly work; it is a tremendous advance in news work to link directly to the evidence, put the evidence in under common eyes, and preserving that evidence is really important.

[14:44] Now, let's talk just a little bit about some of the nature of that evidence. So, some of it is fairly static material, some of it's really complicated — it's big databases or even queries against big interactive systems. And how we deal with that is not at all clear.

We need to recognize that there's a tremendous amount of diversity linked out there. And it would be very good, by the way — there are studies that have been done on, for example, the legal record and the scholarly record of phenomenon that is sort of generically referred to as link rot. Trying to figure out how fast links to various classes of material rot over time. There's been some very good work on that. It would be superb to have somebody do some studies of this in connection with the news media. I'm not aware of any good data on that, but it would be really helpful in crafting strategies to try and step up to the challenge of preserving the news.

[16:05] We also have some really interesting modern-day problems that have a rather different flavor than what's gone on in the past. Think about the phenomenon that's been occurring so much in recent years of enormous troves of document dumps. You know, the Snowden stuff, the Panama papers, the stuff that's coming out of these various reads on people's e-mail. These become crucial evidence in the news reporting cycle now. They become really important pieces of evidence. And yet, it's totally unclear who steps up to preserving this.

[16:59] Libraries and archives aren't sure it's really their problem; news organizations aren't sure it's

their problem. And to make matters even worse, many of these databases are quite equivocated. The provenance of the data isn't entirely clear, the motives of the people who made it available aren't entirely clear and the integrity of the databases isn't entirely clear.

It may be all accurate. It may be 90 percent accurate with a few fake things thrown in, just to make life interesting for whatever reason. It may have a few things redacted. We don't know. It may indeed be almost wholesale fabrication. We're going to need to come to grips with some of these kinds of evidentiary things in a much more serious way then we've done up till now.

[18:02] Now, I want to talk about social media for a minute.

There has been a lot of discussion today about social media and the archiving of social media. And I think some kind of confused discussion about whether social media represents news or not. And I would suggest this: most social media is actually observation and testimony. Very little of it is synthesized news. It's much more of the character of a set of testimonies or photographs or things like that. And collectively it can serve to give important documentation to an event, but often it is incomplete and otherwise problematic. Furthermore, I think many of us look at social media with ethical unease in this sense: journalism, as published, as made available, as placed on network platforms, is generally a pretty deliberate act. There is very little ambiguity, usually, that that information is intended to be public and shared by the public, and that it was put up there with the recognition that it was going to be public and that it was going to be permanent or at least long lived. Whereas the whole question of how much do people using social media understand about how public and long live their various contributions to the social media are, and how they balance risks and rewards and things like that, is something we've touched upon a number of times today with profound unease.

[20:10] I would certainly argue that it is absolutely crucial that we come to some kind of social consensus as, what part of the output of social media should be part of the broad cultural record. And what are the ground rules for selecting and preserving those parts. But I think that it's very helpful to separate somewhat, the discussion of journalism from the discussion of the preservation of the cornucopia of material that's coming up on social media, while recognizing that journalism frequently will reference into social media now as part of its evidence base. And we need to be mindful of that linkage.

[21:18] So, I think that perhaps breaking down the problem like this and recognizing that none of the boundaries are as bright as you might like, but trying to structure the questions along these lines may be fruitful as we try and devise some systematic approaches to this. Especially for the large number of news organizations, journalistic organizations that really need help because I think that their archives are genuinely at risk, in many cases they're long term organizational viability is at risk. That's a harsh lesson that we've seen played out in the world of journalism over the last two or three decades.

[22:15] I want to conclude, and I hopefully will conclude, with enough time that we can field some questions and comments with a strong argument that there's a public policy issue here, too. And that this public policy issue is not getting aired sufficiently.

Particularly for journalism, which is specially recognized constitutionally, the freedom of the press, and you know we recognize it is so critical to the operation of our society. We've got to find ways to preserve this. So, this is a place where we need a much stronger public consensus. We need a recognition that responsible journalism implies a lasting public record of that work. And I don't think that exists right now in most cases.

[23:28] It's very important to recognize that our entire stewardship of our social record is about two decades now into what I'd say is a profound structural change away from ownership and objects that memory organizations could acquire, store and take care of, and towards a licensing regime that

leaves almost all of the choices in the hands of the people who own the content. We cannot, under current law, protect most of this material very effectively without the active collaboration of the content producers. And in fact, we've also seen, harking back to the journal case, that having that active cooperation is really essential to make the process work.

[24:41] It is striking to me that the only way we have right now, of assuring the preservation of news material published by an electronic news source that doesn't want to cooperate in its preservation, is the compelled copyright deposit that is part of the copyright law, which says that the Library of Congress can demand deposit. Now, as you heard in a couple of the breakout sessions, guess what? It hasn't been very interested in doing that. It is in some ways a stunning failure to meet a core public mission.

[25:36] At the same time, it is worth observing two things. One is that, I'm not wildly enthused with the notion that the Library of Congress, with all of its funding instability as a federal agency, all of its potential political shaping in various directions; I'm not wild about the idea of that being the sole and only custodian of this critical piece of our cultural memory. I think it's too much responsibility. It's too big a single point of failure. It's really also too big a job for any single organization.

[26:29] So I really think that there is a profound need now to frame a public policy discussion about how we're going to remember our cultural record going forward. It's interesting for me to note, for example in the UK, where they also have copyright deposit. In fact, that doesn't rest only with the British Library; it spread around among about six institutions, including not just national libraries but also some of the major long-lived research libraries like Oxford, and Cambridge, and Trinity and in Ireland; those sorts of places. I'm sorry not Ireland; Edinburgh in Scotland.

[27:26] We might be very well served to think about whether a model like that is more appropriate for a world where so much control has been transferred to the copyright creators and outside of the traditional market of objects, and for sale and stewardship of objects.

[27:49] So, those are the points that I wanted to make. I hope that, if nothing else, maybe it's given you a sense of the need to think about this not just in a kind of an episodic, making the world a little better working with one or two organizations, and what can my memory organization or my news organization do; but the need to complement that with some systematic thinking.

[28:22] And one of the lessons that I take very strongly away from the experience with the scholarly journal world, is the value of being able to look at what the memory community is doing collectively, as a community. I don't know of any good way to get at that right now for the news. I don't think it would be terrifically expensive to establish such a registry. It certainly wasn't a huge expense for the you know scholarly communication world, especially when put in the context of that you know market place, and the value of that marketplace per year.

[29:15] So I'd invite us to think about whether something like that might be a very useful initiative going forward. And, indeed we saw echoes of that in I believe one or two of the of the breakout group proposals. With that I am going to stop to keep us on time. I believe, Ed, I have time for a couple of questions.

[29:40] One of the things that happened was there was much more centralization of content onto a few, kind of hub sites and then walled garden kinds of social media environments. Very few people seem to really want to live in this environment where feeds were handled quite locally. It's actually been interesting. I've been looking a little bit at Apple's news on the iPhone lately and that has much the same kind of character but with the consolidation happening upstream into this sort of walled garden of the iPhone land. So, I think that the whole sort of structure of the world has changed away from the design assumptions that were baked into our SS and atom.

[30:52] Thanks again.