Collections and Collecting in the 21st Century

by Clifford Lynch (Faculty symposium transcript)

We are here to talk about the future of research libraries and how they are evolving. I think that one very clear theme that has been woven through all of today's talks, though not explicitly stated, is that this evolution is going to take place in a way that is integrally tied to the changing practices of scholarly work, be it in the humanities, the social sciences, the physical and life sciences, or in professions such as engineering or medicine. It is also going to be integrally tied to the ways in which scholarly communication is evolving in response to these changing scholarly practices. In the first segment of my talk today I'm going to talk about the changing scholarly record, which has and will continue to be a core concern in research library collections. Later I'll focus on sources and evidence – the broader cultural record – which supports the scholarly enterprise, and the role of research libraries in managing this much broader and diffuse wealth of content.

I think it is helpful to make a distinction between what we have seen happen during the last 20 years in "mainstream" scholarly publishing (if you will forgive me the rather awkward term), and the collection of things that are going under the banner of digital scholarship or digital scholarly communication; sometimes these are out at the periphery, but sometimes they now stand very near the core of scholarly practice (consider something like GenBank). If you look at the sciences in particular, but increasingly the social sciences and some of the humanities, the print journal has essentially passed into history. It's gone. Yes, it is still hanging on here and there for a variety of reasons (appropriately cautious attitudes about archiving digital material, the economics of scholarly societies, the special needs of a few fields like art history), but usually it's just a vestigial hindrance to progress. (Note, for example, that the version of record now is almost always the electronic edition of a journal article.) The other great pillar of traditional mainstream scholarly communication, the monograph, has had a slower and more complex evolution that now includes a complex range of digital artifacts, and will probably continue to include traditional codex "books" both in print and in electronic versions for the foreseeable future; we'll come back to this later.

We are growing more confident in our ability to preserve at least some simple classes of electronic files (such as page images) for the long term. One of the reasons print held on for as long as it did was because we had an exquisitely well-developed and resilient (and expensive) system for preserving print. Until we were able to gain a reasonable level of confidence about preserving electronic files through a number of collective systems,

organizational and technological developments, stakeholders such as libraries, authors, and readers, quite responsibly, were not prepared to get rid of print.
To the extent that print journals are still produced and distributed, it's for idiosyncratic and probably transient reasons: a good example being when you join a scholarly society, you often receive that journal in print and/or in electronic form as part of your membership. Typically, if you're affiliated with a university, the university will, of course, license access to electronic versions for the entire university community, setting up a problem of economic viability of scholarly societies: why join when you get the journal through your library? But, at least for some readers, having your own personal print copy of a journal is still very convenient for skimming the current issues and helps justify paying the membership dues. Oddly, the scholarly society gets paid better for print advertising than eyeballs on their digital journal.

And for monographs, even given very significant advances in book reading platforms, many readers still prefer print on paper for intensive reading (ideally they would like both print and electronic versions without paying much incremental cost).

The really striking thing about this, all of the migration and investment in electronic form, is this: if you could dredge up a scientist from the 1950s and show them a printout of a journal article from one of those journals, they would know exactly what they were looking at. They might notice that some elements are different, but fundamentally the discourse style, the template, the way ideas are developed, has not changed very much, save for, perhaps more colour illustration, maybe some (fairly subtle) typesetting changes. The kinds of innovations that Harriette Hemmasi showed us in her presentation finally start creeping towards genuine conceptual change rather than the storage and transmission of "virtual print" articles or monographs by digital means. Indeed, scholars see these new digital works and it is hard to know what to make of them. Consider something as simple and trivial as this: with a traditional printed monograph, you know when you have finished it. If I were to give you a large complex database that is much like a monograph, but offers many non-linear pathways and connections to transverse the argument and the underlying evidence, and particularly if I update it continuously, how do you know when have you have finished with it? When have you looked at it enough to perform a review you may be comfortable with? How do you test whether your students have absorbed it as part of their learning?

We are really dealing with two classes of things here in scholarly communication. One has to do with things that we have migrated electronically but not really revised much conceptually. Then there is this whole new and fascinating emergence of things that genuinely represent conceptual innovation in scholarly communication. A lot of interesting opportunities come with this innovation, by the way. The ability to include interactivity, simulation models, nonlinear navigation; the ability to dynamically situate a scholarly work in a broader context

of evolving knowledge; conversational and social media tools (and their relationships to peer review) – all of these are being explored in various aspects as part of the new scholarly communication. There is another key issue: because it is easy and helpful to integrate evidentiary and source material very extensively with many of these digital projects, you get a much greater intertwining of presentation of evidence along with actual construction of argument. In a traditional print monograph you may see short quotations but you will see lots of references to (often effectively inaccessible) archival material of various kinds, to other works; in the digital world the evidence is right there for engagement.

There's an evaluative challenge here, sorting out the genuinely new, analytic scholarship from the very important work of the marshaling and organizing of evidence, which at least today receives considerably less scholarly recognition in many fields. Let me simply underscore the argument here: there was a time in scholarship, in many humanistic fields, where people got a lot more credit for the preparation of critical editions and related apparatus than they do today. If you go over to the sciences you still encounter quiet conversations about colleagues over drinks: "He or she could have been a really good biologist if they hadn't gotten sidetracked in building and maintaining that database," the database in question being something like GenBank or the Protein Data Bank, which serve as essential core resources for the scholarly work of multiple disciplines and for thousands of researchers. There is this whole question of the extent to which informatics is legitimate disciplinary scholarship or whether it is scholarship of a parallel sort or something else. This is not a completely new phenomenon; if you were to ask engineers who do work on detectors on high-energy physics, there is a debate on whether it is engineering or whether it is physics and how much credit the work should get within physics. We are seeing analogues of those arguments reemerging in many ways. I think that distinction is really critical.

I think that it's important for libraries to remain somewhat agnostic to these controversies and debates, which are an integral part of the evolution in scholarly communications that is taking place today. Over a period of decades, the various disciplines will sort out these issues. But at the same time research libraries must not fall into the trap of only collecting the most traditional (and perhaps, sometimes, dysfunctional) scholarly communication because it's only around such materials that there is a firm consensus about relevance and value. A reasonable confidence in preservation is a necessary but certainly not sufficient condition for accepting innovation in scholarly communications, and the research library community can provide such confidence, thus serving as a key enabler of innovation.

So with this prelude about the evolution of the corpus of scholarly communication, which obviously constitutes a central part of the collections that research libraries must gather, organize and preserve, let us talk systematically about collections and how we deal with them.

I want to be clear that collections aren't the only aspect or role of research libraries, as talks earlier this morning have reminded us. But they will be my focus today.

Let me also note that library is a very slippery term that leads to all sorts of conversations that go past each other because it refers to a physical place, it refers to a collection and it also refers to a living organization that sits within a matrix of larger social compacts and other sorts of things. It is profoundly connected to the societal strategy to maintain our memory and our records of what we did. We still struggle with the legacy of inspiring but intellectually sloppy terms like "digital library" from the latter years of the last century that further confuse our understanding of collections and collecting. Rhetoric about research libraries in the digital age is going to have to navigate these reefs, among others.

Historically, research libraries have worried about collecting two primary corpora. The first is the scholarly record, focusing particularly on the contributions that their faculty and their academic community have made to that scholarly record, and, also, reflecting the parts of that record that are of greatest interest to the ongoing scholarly work of the local community. That is a relatively constrained body of knowledge; I've already talked about the ways in which the scholarly record and our conception of this record are evolving.

The other part of the collecting goals of research libraries–and here they begin to share common cause (and indeed prospects for translating common cause to genuine deep collaboration within their own community and with other forms of memory organizations like museums, archives, national libraries and the like)–is to build and maintain a collection of evidence that can support both future scholarship and scholarship of today. This is something that is much broader than the scholarly record, and that we sometimes allude to as the "cultural record," within which the scholarly record comprises a fairly small subset. No institution can collect the cultural record by itself; there is a certain sense that all of the memory institutions operating collectively are fighting an eternally losing battle here, but they are trying to make wise and good choices within severe resource constraints to select materials and hold them in trust for future scholars and (importantly) all other members of the society who may need that material.

The rules are changing drastically in all areas here, but nowhere so much as those surrounding the broad cultural record. If you look at the collection of broader cultural material, you can roughly subdivide it into two pieces. There is a segment that is commercially produced; it is intended for broad distribution to society through marketplace mechanisms. This would include published books, movies, television broadcast, music, and commercial computer games. Today this would also include social media. Then there are materials that might be more appropriate for library special collections, the one-of-a-kind

things: datasets of research observations or analysis, or proprietary commercial transactions, hand built digital objects, personal unpublished papers, records of digital lives, corporate and other organizational records; all of these are terrifically important for research and have traditionally formed the distinctive treasures characterizing our great research libraries (see Clifford A. Lynch, "The Future of Personal Digital Archiving: Defining the Research Agendas," Personal Archiving: Preserving Our Digital Heritage, edited by Donald T. Hawkins. Information Today, 2013. http://goo.gl/vweC7n).

It is worth noting here one of the many strange and wonderful things occurring in our society as the move to digital content continues. It's a shift that many people don't like to acknowledge (and, indeed, there are many people in denial, and even who are heavily invested in actively arguing this isn't true), but the fact is that our ability to digitally capture and reproduce (through 2D and 3D printing) physical artifacts is getting very good. We can capture 2D objects pretty much perfectly; the issues are about cost and workflow, and rights clearance; 2.5D (paintings with surface texture, numismatics, etc.) are already good and improving steadily, and our ability to capture 3D objects is getting very good as well (sculptures, dinosaur bones, the built environment [cathedrals, etc.]), although industrial scale, high throughput, inexpensive 3D capture of collections of diverse objects is still in its infancy. There is a whole school of philosophical argument (cf. Walter Benjamin) that says these are no substitute for the original, but actually we are getting steadily better at making extraordinarily high quality substitutes. I doubt it will be very long before we are at a point where we can produce copies that are really hard to tell apart from the originals. In fact, if we are honest, these digital surrogates, copies, representations, whatever you want to call them, are a lot better than the original in the sense that you can use various forms of image enhancement, you can get face to face with the details on a nine foot tall sculpture, say Michelangelo's David, whereas they probably wouldn't leave you in peace in a museum to build a scaffold to help you study the statue. You can use multi-spectral enhancement, and that's been used to great effect in various types of manuscripts.

Another part of what we are learning is that physical and digital things have very different vulnerability profiles. They are both vulnerable, but as a matter of responsible stewardship, I would be more uneasy by the day, responsible for a treasure house of physical objects that did not have high quality digital surrogates spread widely around the world, because the testimony of history is that bad things happen to physical collections. There's an enormous Pandora's box of issues about things like art markets and valuations, authenticity, provenance, the future of museums and national cultural heritage and the economics that underpin them that unfold from near-perfect digital capture and reproduction of artifacts which we do not have time to explore here. But it's essential to recognize that the world has fundamentally changed in managing this part of the cultural record. (See Clifford A. Lynch,

"Special Collections at the Cusp of the Digital Age: A Credo," Research Library Issues, no. 267 [December 2009], http://publications.arl.org/rli267/4.)

Let us look at what happens as we move into the 21st century. I think that one theme that we see emerging very clearly (and this was echoed in the other talks) is the need to deal with the overwhelmingly large and complex cultural record collectively. For example, ensuring the digital preservation of publications of various kinds, particularly those where the scholarly sector is the primary consumer, like scholarly journals, seems to be handled best collectively. This doesn't mean that you necessarily engage in engineering hubris that says, "we are so smart that we will optimize everything and build one perfect giant bunker somewhere and that will be the preservation point for the scholarly and cultural record going forward." Conveniently, this model can also serve as a choke point for all sorts of intellectual property and selection issues (a.k.a. censorship and registry) related to preserving and providing access to the cultural record. Those approaches don't have a good record. Monoculture is a very poisonous development when you are trying to preserve things. So you do want to spread responsibility around, across national borders, across technologies, across different organizational and governmental accountabilities and strategies and funding sources, but it is still an approach that looks to more collaborative action than we have seen in the past.

There are other areas where we are just starting to recognize that only by collaborative negotiation and action can we have any hope of dealing with certain critical materials; note that in many cases (in part thanks to failures in copyright law and public policy), for the memory institutions, it's a negotiation from a position of weakness, and success calls for a significant reliance on noblesse oblige and moral responsibility in the commercial sector. For example: newspapers/journalism/the news (language about how to characterize this mainly commercial sector, but also a sector with a clear public interest and special legal and public policy protections, is getting really awkward), which is an essential part of the cultural record, and an essential part of scholarship going forward. It is no accident that, for a long time, research libraries have collected newspapers, ranging from national newspapers of record all the way to local specialized papers. What is happening now, other than a general economic implosion in the journalism marketplace, is that the large successful journalistic enterprises are no longer just newspapers anymore, or TV stations; they are databases with very complicated interfaces. Yes, some of them may still provide printed copies, or evening news broadcasts, but they are merely a poor shadow of what is in these databases. I cannot imagine how we are going to manage the preservation of journalism on a continuing basis without some kind of collective agreements with the key players here.

Actually, I think that we still need to see the long-term responsibility for preservation of the cultural record as falling to memory organizations: to libraries, museums, and archives. It

doesn't sit with the producers of that material, and particularly not with producers that are commercial entities (especially publically traded ones). The notion that publishers are going to preserve their output for hundreds of years seems rather improbable. We are in a short term state right now where they are being asked to, where content distributers are sitting on their archives, though this is rather unstable and more unstable the farther you get from businesses whose primary revenue stream is the world of research, teaching and learning (the university and scholarly community). If you go to a mass market publisher, the attitude often is that they could not care less; if it stops generating revenue, they will stop preserving it, except if someone else finds a way to generate revenue in which case the publisher will license it to them or sue them. It is not one of their priorities in most cases. This is a very large challenge – going beyond what we think of as core cultural goods like published materials into journalism, social media, and many other areas.

I also want to say a few words on what fully articulated and elaborated research collections hold in the 21st century. It is very clear that they hold a lot of data. In all fields of scholarship, reliance on data is becoming more and more important; it is not the only kind of scholarship being done, but it is a significant amount that relies on large amounts of data that needs to be reused, combined and revisited in order to replicate research and build upon research that has been done. Those data collections come from and are an integral part of the work of our faculty at our research institutions. They come, by the way, with a messy intertwining with software that's used to interpret them, so research libraries will be dealing with software as well, from multiple perspectives. Research data is not going to be managed solely by universities; science agencies are going to build repositories and we are going to see all kinds of disciplinary activity. Museums will play a role. It is going to be a very complicated patchwork, but at some level, I cannot help but feel that, especially given the fickle government nature of funding to science agencies, the ultimate responsibility for a lot of this material in the long term is going to fall to the cultural memory institutions and particularly the universities (primarily through their research libraries). Evidentiary collections, observations of all kinds, also things that we don't think about: every major urban center is generating and capturing vast amounts of sensor data that tell you about how people are living, moving and conducting their lives. New scholarly centers focused on "urban informatics" are beginning to explore what we can learn from this data. Who is going to preserve that? How much do we need to preserve, and for how long?

There has been an explosion of non-mass market consumer goods forming an interesting and important part of the cultural record, materials produced for reasons other than direct sale or licensing revenues. The memory sector needs to engage the collecting of these materials more effectively. There are sectors of mass-market material that shut out direct collection by cultural memory organizations for various reasons; other sectors remain unrecognized by

these cultural memory organizations as important and in need of stewardship. This is a very complex landscape that we desperately need to survey in detail, and that I wish I could spend the rest of the morning beginning to explore with you in detail. But time doesn't permit.

One of the very interesting things that we have had going for us for the last few centuries is that, when the official cultural memory sector messes up and does not recognize something that is happening, that a new genre or format is emerging as an important part of the cultural record, often there is a cadre or community of passionate individual collectors that do recognize it. This is a great safety net for cultural memory. Over the course of subsequent generations, these materials roll into the institutions as gifts or purchases.

We've seen this with sound recordings, with film, with various kinds of modern art, with graphic novels, various kinds of genre fiction, etc. At present we have computer games that have grown into a huge economic sector (and absorb a great deal of time by the public) but we are still uncomfortable about whether they belong in libraries, or how much of these materials to collect, not to mention the technical and conceptual challenges of actually collecting them.

But we are building technical, legal and marketplace barriers that are making it difficult to be a collector, various kinds of digital rights management, licensing rather than ownership market frameworks, rental-only business models, the notion that perhaps you cannot pass on your digital music collection when you are done with it. These are developments that are putting the traditional collector's safety net into some jeopardy. And non-commercial stuff is really hard to deal with: you don't just buy and own an object anymore, or obtain and keep one. One of the wonderful traditional roles of the publishing industry is that it packages up materials in ways that make it easy for libraries to acquire them, and then offers nice list of materials ready for acquisition. Now we face a time where libraries need to turn everything around: go out, look for what's important and attempt to clear the rights to build collections. This is going to need to be done individually and collectively and going to need to be done with a lot of poring over what is going out on the 'net, such as self-published material of all artistic and creative genres, commercial and business records, etc. I think that these things are going to characterize a very different kind of collective collection building that is going to be essential to libraries in the 21st century.

I could go on a lot longer about this – I genuinely believe this is perhaps the core challenge for the future of our stewardship institutions, and it is absolutely not getting enough discussion and scrutiny – but I would like to allow some time for questions, so I will just close with two comments.

One ties back to what Harriette Hemmasi, in particular, was highlighting in her talk about the new forms of scholarly discourse, our discomfort about how to evaluate and deal with it, to legitimize it, in light of its fragility from a preservation point of view. It is very hard to take seriously scholarship that is going to evaporate in 10 years without leaving a trace. It doesn't feel like scholarship if it doesn't have some hope of permanence. If we cannot create and deploy a stewardship strategy for new forms of scholarly discourse, we undermine and marginalize them without regard to their merits. Here, I think we are going to increasingly see libraries moving into archival roles that are going to involve working closely with the creators of this material. Whether you put some sort of a press or press-like organization in the middle or not, it doesn't make much of a difference in the end. One of the great barriers to innovation in scholarly communication is that we haven't made sufficiently firm commitments to, or been clear enough about, what we are and are not able to archive in the long run. And we have failed to fully appreciate the connections between these commitments and legitimacy of new forms of scholarly work and communication.

Finally, I have to revisit Chris Bourg's reference to Vint Cerf and the "digital Dark Age." I was actually present when Cerf gave that talk a couple of months ago (February 2015) at the American Association for the Advancement of Science meeting in San Jose, California. He had clearly worked up a full court press release and interviews, priming some journalists on this (see http://www.npr.org/sections/thetwo-way/2015/02/13/386000092/internet-pio...or http://www.bbc.com/news/science-environment-31450389, for example), and yes, it is absolutely true that, to those of us who have spent the last 20-30 years deep in the specific technical and policy issues, that there was a certain slightly annoying naïveté to his comments. These professionals and scholars would, I think, correctly argue that some parts of the picture were much grimmer than Cerf indicated and that, in other places, a lot of progress had been made in the past couple of decades, and the prospects were more promising that he suggested.

And, in some of the same communities, there was legitimate frustration about the failure to recognize that people have been both sounding the alarm and working hard on the problem for a long time now; Chris used, as an example, a tweet by Dorothea Salo (@LibSkrat) from February 15, 2015, "I just wanna pat Vint Cerf onna head and give him some hot cocoa and tell him it's gonna be okay, we got this" (https://chrisbourg.wordpress.com/2015/03/18/the-once-and-future-librarian/) which expresses this beautifully.

I also need to say that I've known and admired Vint Cerf for more decades than I suspect either of us want to definitively document, and view him as a valued colleague but not a close collaborator. He has been very kind to me over the years with his insights and encouragement.

When he spoke of the digital Dark Ages, I don't think that Vint Cerf was talking to the memory professionals, at least not primarily. He was speaking to a considerably broader public.

Personally, I have to say that I am only moderately worried about the technological issues for a very large class of core material, while at the same time I feel that we are in unbelievably deep trouble in some more specialized software-intensive areas, and that there are parts of the cultural record that we have no idea how to handle either technically or intellectually, and that may prove to be preservable only in rather limited ways. Overall, though, my sense is that the legal and public policy issues are going to be far more of a threat to our ability to preserve.

There are at least three huge challenges in preserving enough of the cultural record in the digital age. The first is technologies of capture, organization and preservation. The second is funding. The third, and most critical, is social consensus and social will. Historically, this agreement on the need and indeed the necessity to preserve the cultural record was woven into both our infrastructure of memory institutions and the various legal regimes (such as intellectual property law) and broad social agreements about the roles and acceptable activities of our cultural memory organizations. Politics played a role, particularly in ensuring funding for these critical activities. During the end of the 20th century and the beginning of the 21st I fear we have lost this national consensus; marketplace forces and intellectual property legislation passed on behalf of aggressive commercial interests have eviscerated many, perhaps most, of the agreements under which our memory organizations function without regard to balancing public interests. If research libraries are going to succeed in their collecting missions in the 21st century and beyond, we must re-establish this social consensus. We must make the case about why social will and societal commitment is essential.

This is a challenge that is too big for libraries, or cultural memory organizations more broadly, to undertake alone. It's too big for the research and education community alone. It will take a broad movement among the intellectual land cultural leadership in our society. This is where I rejoice to see the calls for action on the part of widely known and recognized leaders like Vint Cerf. We need similar calls from all sectors of our society – literary figures, university presidents, artists, business and political leaders, etc. We must join together to support these testimonies and to make them visible, to move them into the center of public discourse and discussion whenever we can. The advocacy strategy around this challenge is, in my view, going to be a central issue and point of focus for research libraries in the 21st century, and intimately tied to collecting strategies.

To conclude, then, that's a look at collection building in the 21st century and how it fits into a

broader context. I think it's going to be vital for research libraries to recognize, as we see all of this digital convergence, that they are part of a broader context. They are going to need to form alliances and collaborations, not just with other colleagues in the research library community, but substantially beyond that, in order to be able to ensure that, particularly, the evidentiary material is there to do scholarship, not just later in the 21st century, but maybe even in the 22nd century.

This is a lightly edited transcript of a talk that I gave at the McGill University Faculty Symposium on the Future of Academic Research Libraries on March 18, 2015. My thanks to Jenn Riley of McGill, Joan Lippincott and particularly Diane Goldenberg-Hart of CNI for help in preparing the transcript for publication. In a few places, I have taken the liberty of adding citations to other publications where I have explored issues discussed here in more depth.