Q&A with CNI's Clifford Lynch: Time to re-think the institutional repository?

RICHARD POYNDER

22nd September 2016

To go direct to the Q&A please click here

Seventeen years ago (October 1999) 25 people gathered in Santa Fe, New Mexico, to discuss ways in which the growing number of e-print servers and digital repositories could be made interoperable.

As scholarly archives and repositories had begun to proliferate a number of issues had arisen. There was a concern, for instance, that archives would needlessly replicate each other's content, and that users would have to learn multiple interfaces in order to use them. It was therefore felt there was a need to develop tools and protocols that would allow repositories to copy content from each other, and to work in concert on a distributed basis.

With this aim in mind those attending the New Mexico event – dubbed the Santa Fe Convention for the Open Archives Initiative (OAI) – agreed to create the (somewhat wordy) Open Archives Initiative Protocol for Metadata Harvesting, or OAI-PMH for short.

Key to the OAI-PMH approach was the notion that *data providers* – the individual archives – would be given easy-to-implement mechanisms for making information about what they held in their archives externally available. This external availability would then enable *service providers* to build higher levels of functionality by using the metadata harvesting protocol.

To put it another way, the aim was to create a protocol that would enable metadata descriptions of records in compatible archives to be harvested by third-parties. These third-parties would then offer value-added services that leveraged the content in the archives – e.g. by offering peer review services and/or overlay journals. Most importantly, OAI-PMH was intended to enable third-parties to provide cross-repository search functionality, by aggregating the content in compliant archives and making it all searchable through a single interface.

Initially referred to as the Universal Preprint Server, it was later that the initiative became known as the Open Archives Initiative.

The repository model that the organisers of the Santa Fe meeting had very much in mind was the physics preprint server arXiv This had been created in 1991 by physicist Paul Ginsparg, who was one of the attendees of the New Mexico meeting. As a result, the early focus of the initiative was on increasing the speed with which research papers were shared, and it was therefore assumed that the emphasis would be on archiving papers that had yet to be published (i.e. preprints).

However, amongst the Santa Fe attendees were a number of open access advocates. They saw OAI-PMH as a way of aggregating content hosted in local – rather than central – archives. And they envisaged that the archived content would be papers that had already been

published, rather than preprints. These local archives later came to be known as institutional repositories, or IRs.

In other words, the OA advocates present were committed to the concept of author self-archiving (aka green open access). The objective for them was to encourage universities to create their own repositories and then instruct their researchers to deposit in them copies of all the papers they published in subscription journals. As these repositories would be on the open internet outside any paywall the papers would be freely available to all. And the expectation was that OAI-PMH would allow the content from all these local repositories to be aggregated into a single searchable virtual archive of (eventually) all published research.

Given these different perspectives there was inevitably some tension around the OAI from the beginning. And as the open access movement took off, and IRs proliferated, a number of other groups emerged, each with their own ideas about what the role and target content of institutional repositories should be. The resulting confusion continues to plague the IR landscape today.[1]

Disappointment

Since 1999, therefore, there have been many arguments, debates and disagreements about the institutional repository. In fact, the first dispute over the ownership and purpose of the OAI erupted within weeks of the Santa Fe meeting.

Differences aside, however, what everyone at Santa Fe shared was a desire to make distributed archives interoperable. And it was assumed that OAI-PMH would make that a reality.

Since then thousands of IRs have been created, and open access has flourished – to the point where most now assume OA is set to become the default model for scholarly publishing.

Yet for all that, the interoperability promised by OAI-PMH has never really materialised, few third-party service providers have emerged, and content duplication has not been avoided. Moreover, to the exasperation of green OA advocates, author self-archiving has remained a minority sport, with researchers reluctant to take on the task of depositing their papers in their institutional repository. Where deposit *does* take place, it is invariably hard-pressed intermediaries who do the work.

In short, neither OAI-PMH nor the IR movement has delivered on its promise. Nor has the wider objective of re-engineering scholarly communication for the networked world, which many had assumed would be the outcome of the New Mexico meeting. It is no surprise therefore that the Santa Fe attendees have become disenchanted, and some have begun to express their disappointment publicly.

Last July, for instance, one of the key architects of OAI-PMH, Herbert Van de Sompel, posted a tweet that linked to a presentation he had given in 2000 on the Universal Preprint Server. He headed his tweet: “Those were the days we thought we could change the #scholcomm system”. And in his presentations today Van de Sompel is invariably critical of the way in which OAI-PMH had been designed.

[1] More background on the early days of the IR is available in a 2006 essay I wrote.

This July another of those who attended the Santa Fe meeting – Eric Van de Velde – painted an even gloomier picture. In a blogpost entitled "Let IR RIP" he argued that the institutional repository is now at a dead end, and in fact is obsolete. "Its flawed foundation cannot be repaired. The IR must be phased out and replaced with viable alternatives."

And with researchers reluctant to self-archive, many OA advocates have become disenchanted with the progress of green OA. So while the OA movement may now appear unstoppable there is a growing sense that both the institutional repository and green OA have lost their way.

It is not hard to see why. Not only are most researchers unwilling to self-archive their papers, but they remain sceptical about open access *per se*. Consequently, despite a flood of OA mandates being introduced by funders and institutions, most IRs remain half empty. What content they do contain often consists of no more than the bibliographic details of papers rather than the full text. More strikingly, many of the papers in IRs are imprisoned behind "login walls", which makes them accessible only to members of the host institution (and this is not just because of publisher embargoes). As a result, the percentage of content in IRs that is actually open access is often pretty low. Finally, since effective interoperability remains more aspiration than reality searching repositories is difficult, time-consuming and deeply frustrating.

True OA is nevertheless growing. But this is not due to self-archiving; it is partly a consequence of mediated deposit, but increasingly a function of the fact that legacy publishers now offer pay-to-publish gold OA (notably expensive hybrid OA). And this last development is being facilitated and encouraged by research funders like Wellcome Trust and Research Councils UK, along with the many universities that have introduced gold OA funds. In addition, national organisations like the Association of Universities in the Netherlands (VSNU) have begun to negotiate new Big Deals with legacy publishers that combine payment for gold OA with traditional subscriptions.

With the EU earlier this year setting a goal of achieving 100% OA by 2020 the gold rush can be expected to accelerate going forward. After all, if most researchers are not willing to self-archive how else could such a goal realistically be achieved?

Admittedly large-scale funding of gold OA is more of a European thing today, but we should not doubt that there will be a rapid escalation of pay-to-publish gold OA, at least in the developed world. Indeed, we can expect it to become the primary means of providing open access in the global north. Even large archetypal green OA policies like those introduced by the National Institutes of Health (NIH) and other US federal agencies will eventually be primarily fulfilled by gold OA – unless something changes.

Consider, for instance, what the Deputy Director for Resource Management at the US DOE Office of Science wrote recently: "Saying that US agencies are implementing green OA models is not the same thing as saying that they *prohibit* gold OA. The payment of gold OA fees by authors or their institutions is typically an allowable cost under most federal research grants and contracts."

And as gold OA accelerates so the logic of depositing papers in IRs dissipates – a point made last September by T Scott Plutchak, Director of Digital Data Curation Strategies at the

University of Alabama at Birmingham. Pointing out that the duplication of content that OAI-PMH was meant to avoid is now a growing problem, he said: "[E]fforts to fill IRs with copies of peer reviewed papers that are already available OA somewhere else, such as the publisher's site or a repository like PubMed Central, are misguided. Such efforts, which consume a considerable amount of energy for some IR managers, have not achieved their intended benefits, and they divert resources from other activities that would have a much greater benefit. I suggest that we take a deep breath, reassess what's working and what we're trying to achieve, and, for at least some IRs, shift priorities."

What better sign that green OA has failed as a strategy than the recent decision by the most long-standing, persistent and articulate green OA advocate (and another attendee of the Santa Fe meeting) Stevan Harnad to hang up his OA boots. In March, in a Twitter exchange with PLOS co-founder Michael Eisen, the self-styled archivangelist tweeted "I fought the fight and lost and now I've left the #OA arena."

Explaining the background to the tweet, and expanding on the reasons for his "retirement" in a subsequent interview with the Polish open science site Otwarta Nauka, Harnad said: "I had long wished for all refereed research to be Green OA, and my wish has not been fulfilled. So I simply stated the fact: That he [Eisen] is right, I have lost and I have given up archivangelizing."

The question then becomes: what will happen to the thousands of IRs that have been created in order to facilitate green OA?

Conundrum

So we are confronted with a conundrum: open access has won the argument (at least so far as governments, research funders, many universities and, importantly, publishers are concerned), and the number of research papers that are open access is growing. But the institutional repository is experiencing an existential crisis. More broadly, the vision of re-engineering scholarly communication that informed the New Mexico meeting would seem to have stalled. It is therefore no surprise that the architects of that vision have become disappointed and disillusioned.

What has clearly not helped is the host of often contradictory demands and expectations that quickly latched on to the institutional repository. In the hope of focusing minds (and efforts) several high-profile individuals did in the early days try to formulate a common vision for the IR. In a widely-cited paper[2] published in 2002, for instance, SPARC's Raym Crow proposed two roles for the IR. First, wrote Crow, IRs should be viewed as "a critical component in reforming the system of scholarly communication." Second, he said, they should aim to become "tangible indicators of a university's quality … [so as to] … demonstrate the scientific, societal, and economic relevance of its research activities".

A year later (2003), CNI's Clifford Lynch published another widely-cited paper called *Institutional Repositories, Infrastructure for Scholarship*. Lynch, however, did not view the IR as a tool to help reform scholarly communication. His description of the IR was "a set of services that a university offers to the members of its community for the management and dissemination of digital materials created by the institution and its community members."

[2] *The Case for Institutional Repositories: A SPARC Position Paper*.

Be that as it may, neither Lynch nor Crow succeeded in achieving community consensus. Above all, their proposals were (both in their different ways) unsatisfactory to green OA advocates (see here and here).

In the Q&A below Lynch suggests that the IR's development was hampered because its needs and purposes were conflated with the open access agenda. The role of an IR, he says, should be to expand and diversify the existing scholarly publishing system "by providing access and stewardship for material that mainly falls *outside* of the traditional scholarly publishing system."

OA advocates fervently disagree with this. The IR, they assert, was conceived precisely in order to advance the open access agenda, by providing alternative (free) access to content locked behind the paywalls imposed by the traditional scholarly publishing system.

For Harnad, therefore, the *only* purpose of the IR is to provide a platform on which researchers can post copies of the papers they have published in subscription journals, thereby freeing them from the "subscription firewall".

Contrast this with Herbert Van de Sompel's view, which was that the goal of the OAI was to change the scholarly communication system; and with Eric Van de Velde's assumption that IRs were intended to "disrupt scholarly communication".

Meanwhile, librarians (who are invariably tasked with managing the IR) have long argued that repositories also have an important role to play in preserving the scholarly literature. And more recently, they have come to view the IR as a platform on which new OA journals can be created and managed.

Finally, in the past year or so we have begun to see calls for IRs to play a role in Research Data Management (RDM) as well.

What has surely also limited what IRs have been able to achieve is that by and large they have been seriously under resourced. This point was graphically made in 2007 by erstwhile repository manager Dorothea Salo. Her conclusion nine years ago was: there is need for a "serious reconsideration of repository missions, goals, and means."

Where next?

So where next for the institutional repository? While it is possible that some IRs may (as Van de Velde recommends) be retired, I doubt it. Rather, I see two possible scenarios. In one scenario they will be captured by commercial publishers, much as open access itself is being captured by means of pay-to-publish gold OA. In the other scenario the research community will finally come together, agree on the appropriate role and purpose of the IR, and then implement a strategic plan that will see repositories filled with the target content (whatever it is deemed to be). Vitally, they will also at last be made interoperable.

The challenge for the research community is that time is not on its side. With subscription revenues expected to decline as open access grows, and with their control of scholarly content therefore no longer a foregone conclusion, publishers feel under some pressure to

create new markets around open access, including around the institutional repository. And they are keen to do so in a way that maintains their control of the content.[3]

In this Elsevier is already having some success – as evidenced by its recent announcement of a partnership with the University of Florida in which hosting of, and access to, research papers indexed in the University's repository will be outsourced to Elsevier. Essentially this will turn the repository into a search interface and promotional tool for content hosted and controlled by Elsevier, regardless of whether or not that content is classified as paywalled or open access. And on 21st September it was reported that the University of Florida has also signed a Letter of Agreement with CHORUS in a pilot initiative that will enable the Elsevier project to be scaled up "to a multilateral, industry effort."[4]

Elsevier has also made a number of controversial acquisitions of paper sharing and central repository sites like Mendeley and SSRN.[5] Also a source of controversy is the growing realisation that Pure (the CRIS system Elsevier acquired in 2012) poses a direct threat to the IR.

Elsewhere, for-profit bepress is building a lucrative business out of its Digital Commons product, which offers a subscription-based hosted repository service.[6]

Another interesting development is Wiley's recent acquisition of the hosted platform provider Atypon – a move that industry observer David Worlock believes may be the first signs of a new generation of Super-Platforms. The acquisition is presumably Wiley's response to the competitive threat it sees Elsevier's recent purchases posing, and presumably the model Worlock has in mind is what in the pre-internet era was known as an "online host" (e.g. services like DataStar, Dialog, and QuestelOrbit.)[7]

What is interesting about the Atypon acquisition from our point of view is that it demonstrates the consequences of the research community failing to create the network of interoperable open archives envisaged by those who developed OAI-PMH. This failure has paved the way for the "academic publishing oligopoly" (as it is now known) to start colonising and building out the open access infrastructure. Certainly, there is a need for new infrastructure. (As Worlock wrote in his commentary on the acquisition. "I keep on hearing scholars complaining about how hard it is to cross search files located in many different places and governed by different access rules.") The question is: who should control the OA infrastructure, and what are the dangers if for-profit concerns come to own it?

In his commentary Worlock added: "On 24 August the Mendeley blog invited us all to try out the beta version of Elsevier's DataSearch, a cross file tool set to allow users to cross search ScienceDirect and certain other files – ArXiv, for example – in conjunction with it".

---

[3] This, we should note, was precisely what many expected IRs would prevent.

[4] Several CHORUS publisher members are participating in the pilot, including American Chemical Society, American Physical Society, Association for Computing Machinery, Elsevier, The Rockefeller University Press and Wiley.

[5] It is worth noting that most of the researchers who use these services had mistakenly assumed they were community initiatives rather than for-profit ventures.

[6] bepress' now considerable subscription list can be viewed here.

[7] Atypon's customers include many scholarly publishers, including Emerald, IEEE, McGraw-Hill, MIT Press and ACS.

It seems safe to assume that Elsevier plans at some point to start charging researchers for access to open content. Indeed, others are already doing this, or planning to. Recently, for instance, it was announced that arXiv content is to be added to the charged-for Inspec database. Elsewhere, FIZ Karlsruhe's document delivery service FIZ AutoDoc has begun to charge users for simply *linking* them to open access journal articles.

So what, some will say, open content can still be accessed for free at source. But wouldn't it have been better if the OAI-PMH dream had been realised? And what happens if the funding for OA repositories like arXiv evaporates at some point? Could open content start to find itself paywalled again, and free versions start to disappear? Importantly, by capturing and controlling OA hosting and OA search functionality legacy publishers can expect to continue earning what many believe to be their "obscene" profits from the public purse.

These developments also need to be viewed in the context of the way in which publishers have been degrading and emasculating green OA policies. (These were introduced, let's remember, to overcome publishers' resistance to open access, by mandating it). But publishers have responded to these mandates by simply imposing ever more restrictive green embargoes,[8] and pressurising researchers into choosing pay-to-publish gold OA.

Of course, as funders and universities introduce more and more gold OA funds the need to exert such pressure will lessen, but the end result will be much the same. Publishers will be able to say to researchers: If you are subject to a mandate then go to your funder or institution and ask for the money to pay for gold OA. We will then do all the hard work for you. You won't have to worry about copyright issues, you won't have to worry about embargoes, and you won't have to worry about self-archiving your paper, or agonise over which version you can archive[9]. We will take care of all that for you, including the archiving itself. Just give us the money!

Essentially, the academic publishing oligopoly has embarked on a process of capturing open access, and emasculating green OA. And now it is coming after the institutional repository.

Resistance

But as the implications of this have begun to sink in researchers are pushing back. News of Elsevier's partnership with the University of Florida, for instance, led to number of librarians and other interested parties publishing a joint statement of dissent. And the threat that Elsevier's Pure product poses for the IR has also come under fire (see here and here for instance).

In fact, some in the research community have concluded that it is time to expunge for-profit organisations from the scholarly communication process. Even the activities of start-ups like Academia.edu and ResearchGate are attracting critical scrutiny. In 2015, for instance, an entire conference was held around the question: "Why are we not boycotting Academia.edu?"

Opposition to Academia.edu was sparked by the realisation that it has accumulated 42 million users, that it can offer more and better services than IRs can, and it is now looking to

[8] This is seeing publishers prevent green OA papers from being made open access for up to four years in some cases.

[9] It turns out that there are now multiple possible versions of research papers.

monetise its users in ways not everyone is comfortable with. Amongst other things, this has led to librarians seeking to wean researchers away from the site (see here, here and here for instance).

More practically, the research community has begun to launch new non-profit initiatives to compete with publishers. We have seen non-profit OA journals emerging for some years. Now we are seeing a spate of central and/or disciplinary repositories modelled on arXiv being launched as well – e.g. bioRxiv, SocArXiv, engRxiv, and PsyArXiv. Much of this activity is being driven by the Center for Open Science (COS) whose Open Science Framework (OSF) is providing the infrastructure for many of the new repositories.

Interestingly, the OSF platform is expect to offer the interoperability that OAI-PMH failed to deliver. As COS Community Manager Matt Spitzer told me when I spoke to him recently about SocArXiv, "as other groups use the platform for a specific subset, they can brand the site however they wish. OSF Preprints can then aggregate search results across all of the sub groups."

It is important to note that the new OSF-based repositories are central disciplinary archives, not IRs. This renewed focus on central solutions makes sense: to date, the most successful repositories have tended to be disciplinary or national/regional in scope rather than institutional repositories. There is arXiv of course (which now hosts over 1 million e-prints and is seeing some 8,000 new ones added each month), but also NIH's PubMed Central (which contains 4 million articles) and repositories like the French HAL and the Latin American "electronic virtual library" SciELO.

What is also noteworthy is that many of these new repositories are (like arXiv) focused on preprints. As such, they can be viewed as part of a new preprint movement that appears to be emerging. This movement is best exemplified by the ASAPbio initiative – whose mission is "to promote the productive use of preprints in the life sciences."

Those at the more radical end of this movement believe that by focusing on the preprint it will be possible not just to share research more quickly, but to do so in ways that dispense with publishers. In this way, it is hoped, the research community will finally be able to eject publishers from the nest.

With this thought in mind, for instance, last year Sir Timothy Gowers launched a new mathematics journal called *Discrete Analysis*. This uses arXiv as its content platform, and the publishing process involves no traditional publisher whatsoever.

Elsewhere, the Episciences initiative has also developed a new infrastructure for overlay journals. These journals can use arXiv as their content layer, or alternatively they can use a number of other preprint platforms such as HAL or the Dutch repository Centrum Wiskunde & Informatica (CWI). An example of such a journal is the recently launched *Journal of Interdisciplinary Methodologies and Issues in Science*.

If the new preprint movement takes off it will of course further complicate the picture for IRs. While the OSF platform can act as an IR, says COS, it is more likely that institutions would want to connect their existing repository to OSF. But what implications this does or does not have for the future of IRs is not immediately clear to me.

We should also mention that a number of central services have been developed in recent years that aggregate open access content by indexing papers both from repositories and from OA journal sites. This includes Paperity, 1Science and BASE. However, many of these services appear (like SSRN) to be operated by for-profit concerns (here and here for instance). And given the continuing interoperability problems that bedevil IRs they are often not as effective as users might assume. It does not help that they often seem to rely on pulling the text down at the time of access (i.e. in real time), which can be extremely slow, and fails completely when the source document is not available (e.g. here).

BASE – operated by Bielefeld University Library – *is* a non-profit. It collects and indexes the metadata of web documents provided via the OAI-PMH protocol (which I assume means content from IRs). Clearly, however, it cannot retrieve full text if the repository itself hosts only the bibliographic details, or has put the paper behind a login wall. Consequently only around 60% of the records in BASE are full text. Moreover, many of the records do not appear to be peer-reviewed documents. BASE includes, for instance, blog posts – some as brief as the first one listed here (which is effectively a link to a link).

Finally, there is the EU-funded OpenAire, which aggregates content hosted in European repositories. Again, no central harvester can provide OA to documents if the underlying paper is absent in the IR, or behind a login wall. Like BASE, OpenAire also seems to harvest blog posts and other non-peer-reviewed content. Unfortunately, it does not signal in its records whether a document has been peer reviewed (see here and here for example). And since it is funded through the Horizon 2020 programme for 42 months from January 2015 there must be some uncertainty about its permanence.

What we learn is that while these harvesters are keen to boast about the number of records they hold, when you look under the hood you discover a number of issues, including all the difficulties that have haunted the IR movement since the beginning.

Third-time lucky

In light of the challenging, volatile, but inherently interesting situation that IRs now find themselves in I decided recently to contact a few of the Santa Fe attendees and put some questions to them. My first two approaches were unsuccessful, but I struck third-time lucky when Clifford Lynch agreed to answer the questions in the Q&A below.

Lynch is long-time director of the Washington-based Coalition for Networked Information (CNI), an organisation jointly sponsored by the Association of Research Libraries (ARL) and EDUCAUSE, and whose agenda includes work in digital preservation, data intensive scholarship, teaching, learning and technology, and infrastructure and standards development.

As noted earlier, in 2003 Lynch wrote a paper outlining what he saw to be the role and purpose of the IR – a vision somewhat different to that articulated by Crow a year earlier, and at a variance with that promoted by green OA advocates. Amongst other things, therefore, I was interested to know whether, and how, Lynch's views have changed over the past 13 years.[10]

[10] Not much, appears to be the answer.

I think it fair to say that Lynch's answers present a somewhat more sanguine view of the current situation vis-à-vis IRs than presented in this introduction, or in Van de Velde's blogpost. What I did find interesting, however, is that Lynch too believes it is time to re-think "the real prospects and best approaches and roles for IRs in this much-changed world". He cautions, however, that this would be no easy task "given the investment (not just in terms of time and expense, but in many cases in ideology) related to IRs."

The last 17 years suggests that such caution is warranted. We might also wonder who exactly would initiate and manage any such process of reevaluation. As with all issues concerning scholarly communication and open access, no one appears to have the necessary authority (or even perhaps the capability) to oversee strategic decision-making at this level effectively.

And that is why it seems to me most likely that the academic publishing oligopoly will succeed in appropriating both OA and the institutional repository.

Brace yourselves taxpayers!

The interview begins …

Photo courtesy Susan van Hengstum

RP: Seventeen years ago you were one of those who attended the Santa Fe Convention of the Open Archives Initiative (OAI). The mission of the OAI was to promote and encourage "the development of author self-archiving solutions (also commonly called e-print systems) through the development of technical mechanisms and organizational structures to support interoperability of e-print archives." OAI came to be viewed as a key part of the open access infrastructure. As I understand it, the aim was to create what was initially called a "universal preprint server" in which multiple e-print archives (now more commonly referred to as institutional repositories) would be linked together in an interoperable way so that a single search could be conducted across multiple repositories as though the user were searching on one large repository – by, for instance, the use of harvesters like OAIster (now part of OCLC). Would that be an accurate reading? What are your recollections of that meeting, and to what extent do you feel that the OAI goal has been realised to date?

CL: This was a long time ago, and my memory may be somewhat inaccurate, but I remember this a bit differently. What had happened as context was the emergence of pre-print/e-print

systems as a way to greatly accelerate the speed of scholarly communication and to free some of this process from the intermediation of journals – ArXiv (then at Los Alamos) being the prime example. This was really exciting. The next step, and the main focus of the Santa Fe meeting, was to figure out how to federate these archives for searching or harvesting.

The open access agenda (author self-archiving) was not much in evidence at Santa Fe as I recall; nor were institutional repositories and their implications. It was more about pre-prints (including the final pre-print) and my recollection is that it was more focused on speeding up communication among scholars through a parallel system than trying to mount a direct challenge to the existing journal systems. Indeed, as I recall, at least in Physics many of the key publishers quickly made their peace and found co-existence with the ArXiv.

The OAI protocol for Metadata Harvesting was, I think, pretty successful and found fairly wide adoption within the digital library community. There were of course problems and oversights (easy enough to identify with the benefit of 15 years' experience and hindsight). The choices for formats to transfer metadata were problematic and we reluctantly went with Dublin Core as a lowest common denominator at the Santa Fe meeting.

Remarkably, even now, this remains a real problem (and really isn't an OAI or PMH problem, it's much more extensive. OAI shouldn't have to solve it): we don't have a well-accepted, good format for exchanging extended bibliographic citations for articles, preprints, etc. It would have been good to include some security provisions in the PMH; we didn't, perhaps because our perception of the environmental threat levels was much lower fifteen or twenty years ago.

I should say two other things about the success and limitations of the PMH. Google, as I understand it, went in a different direction for indexing site content for the preprint archives (and more generally), and this was a big problem and a big barrier. I don't really know the story behind this.

Finally, I have seen Herbert [Van de Sompel] in some recent talks be very critical of the design of PMH (which he of course played a key role in) as in some sense inconsistent with a full embrace of the web and web-related data models, and he has done some fascinating work in designing a successor approach (ORE), which is much more sophisticated and complex, and deeply connected to and embedded in the whole W3C semantic web thinking. So far, I think that implementation has been rather limited; I don't think it's in most of the popular repository software yet.

While Herbert's critique of PHM is insightful and accurate, I think that he's being far too harsh on himself and the work that was done at the time. I am also more cautious than he is about the benefits of modularity of components in the information eco-system; he is very sensitized to the negatives, perhaps because of the critiques that PMH received from some of the W3C world, and its disappointingly limited uptake outside of the repositories and digital libraries world.

Qualms

RP: Subsequently, I think you personally developed a somewhat broader view of what an institutional repository should be, a view you articulated in 2003 in a document called "Institutional Repositories, Infrastructure for Scholarship". You said, for instance,

"Institutional repositories can encourage the exploration and adoption of new forms of scholarly communication that exploit the digital medium in fundamental ways. This, to me, is perhaps the most important and exciting payoff: facilitating change not so much in the existing system of scholarly publishing but by opening up entire new forms of scholarly communication that will need to be legitimized and nurtured with guarantees of both short- and long-term accessibility". Some OA advocates took issue with you on this (not least Stevan Harnad), insisting that the repository should not aspire to be anything more than a tool to allow researchers to self-archive the papers they published in traditional journals, and certainly not a vehicle for creating new forms of scholarly communication. As such, they argued, the IR should be no more than a supplement to traditional methods, not substitutive. Looking back, do you feel the critics had a point? Have your views on the function and role of IRs changed in the intervening years?

CL: Yes, though I am not sure that I'd say "subsequently"; I can't remember exactly when I started thinking seriously about IRs, but I can't recall ever thinking of them narrowly in the context of article pre-prints/e-prints particularly.

To give you another timeline that's very US-centric: I know that EPrints at Southampton was very early [2000], but it didn't get much attention or uptake in the US that I recall, particularly at the institutional level; MIT DSpace (which was as far as I know, for instance, *never* designed early on primarily as a store for journal articles) really only began to be known in 2001, I think. There were earlier solutions (1990s) that were very document oriented focused on things like departmental tech reports, notably Dienst and other endeavours coming out of work like the DARPA Computer Science Technical Report effort.

Note that I'm just giving you this from my doubtless faulty memory and without going back to verify dates; it would be wonderful if someone someday wrote a really good history of digital libraries that covered all this material!

By the way, that 2003 piece was actually an article that was published in a couple of places. Here are the citations:

Clifford A. Lynch, "Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age," *ARL Bimonthly Report* 226 (February 2003), 1-7. Online at www.arl.org. Reprinted in *Portal: Libraries and the Academy* 3:2 (2003), pp. 327-336. Translated to French by Simone Jerome as "Les dépôts de documents institutionnels: une infrastructure éventuelle pour l'enseignement à l'ère numérique," *Cahiers de la documentation Bladen voor de documentatie* 57:4 (December 2003), pp. 135-143.

I still respectfully disagree with the position that the primary (or exclusive) purpose of IRs should be self-archiving of the traditional journal literature. I think it's a rather cumbersome and perhaps expensive way to do this (though I also readily recognize that institutions would want to host and control a copy of this part of their faculty's scholarly output and that there are very substantial merits to this – perhaps the way to do this would be an IR that is primarily populated by copying material from other places?); disciplinary archives with cross-repository replication seems more promising.

It's very impressive what's been accomplished with PubMed Central, for example, and the way NCBI has integrated this with other data resources in the biomedical and life sciences.

It's also interesting to me that when you look at the recent provisions coming out of the US federal funding agencies in support of the mandate for public access to journal articles arising from their grants, most of them don't look to me like they recognize the IR structure; they favor more centralized solutions.

Having said that, I think that many of the institutional IRs have fallen far short of their potential. The software isn't where it needs to be; barriers to submission are too high; we don't yet have smooth cross-repository replication in place, which would allow the IR to act as the author point of interface into the various funder requirements for deposit, etc.

I also have some qualms about the results of promoting IRs to faculty as a central part of a green open access agenda, particularly in light of the US Funder requirements that aren't consistent with this message that's been shared with the faculty by open access advocates over the past decade or more. And there's the whole question of where IRs fit with the growing challenge of research data management.

It's definitely time for a re-think about the real prospects and best approaches and roles for IRs in this much-changed world. This is going to be difficult given the investment (not just in terms of time and expense, but in many cases in ideology) related to IRs.

And any change is going to have to be a careful transition and evolution that will probably take a substantial amount of time, not least because of the need to communicate with a large population of faculty and other researchers, which is a slow process. And the need to ensure that promises of stewardship are honored; there's nothing more perilous to continuity of stewardship than a major underlying system and infrastructure shift.

Not a good support structure for green OA

RP: Recently one of the other attendees of the Santa Fe meeting – Eric Van de Velde – published a blog post in which he asserted that the Institutional Repository is now at a dead end, and indeed is obsolete. "Its flawed foundation cannot be repaired," he said. "The IR must be phased out and replaced with viable alternatives." He then went on to list a number of reasons for his conclusion. Do you agree or disagree with what Eric says? How do you respond to his rather pessimistic view of the future of the IR?

CL: As I think you can see from my response to the previous question I agree with much of Eric's critique about IRs, particularly as now implemented. They are not a good support structure for green OA. As I say, that should not be their main function (and I never argued that it should, I'm pretty sure.).

The point of IRs, in my view, isn't to disrupt the existing scholarly publishing system, but to allow it to be expanded and diversified by providing access and stewardship for material that mainly falls *outside* of the traditional scholarly publishing system as it exists today – both material created by faculty and material created by the institution, or departments or other groups within it.

Technology has moved on quite a bit in the last fifteen years, and it may be that it makes more sense to think about how to do this in a way that involves more shared or collective platforms and services rather than highly distributed approaches.

But I would note that a lot of the problem here isn't technical: to the extent that individual institutions hold responsibility for lasting stewardship of content, it's been very difficult to financially and administratively structure and implement stable, robust, and resilient collective mechanisms for institutions to pool infrastructure to reduce costs and improve quality of services.

This has also been a challenge with what are clearly community stewardship responsibilities like the traditional journal literature (consider, LOCKSS and CLOCKSS, Portico, etc.) or major disciplinary archives (consider the saga of how to fund and govern the Cornell ArXiv system).

So I absolutely agree with Eric: the sooner we totally disentangle the discussion of Green OA and how (and if) to move it forward from the discussion of IRs, the better off everyone will be.

RP: IRs were nevertheless viewed as key to the success of so-called green open access. Of course, in recent years legacy publishers have increasingly bought into open access, and most, if not all, now offer gold open access solutions. Might it be that green OA and the IR were important mechanisms for helping to persuade publishers to embrace OA, and that now they have been persuaded the research community should be putting more effort into advocating for and promoting gold OA rather than green OA (e.g. by means of journal flipping strategies such as the one proposed by the Max Planck Society), and so perhaps think of retiring the IR in the way Eric suggests? However, I guess you are saying it just that our understanding of what an institutional repository is, and the role it should play, has been too narrowly defined and is in any case in need of an upgrade?

CL: Gold OA has great benefits, not the least of which is simplicity from the faculty perspective. I don't think it's sufficiently recognized as an approach that really does address the goals of the public access policy mandates within the various US Federal Funder requirements (assuming that the gold journals have robust continuity plans through various Keepers services).

Some of the article processing fees look very high, and the transition of a journal from subscription to fees is very messy; not all gold journals call for author fees but I suspect that the ongoing survival of a good number of today's well-known journals as gold open access will call for processing fees, if it happens. The UK experiences here have been very sobering, I think.

But I also think that this question really underscores the problems of conflating the needs and purposes of IRs – which I assert go, or should go, far beyond just tracking the traditional journal article system – with how to deal with the very real issues around open access and public access to the existing system of scholarly journal publishing.

Alternatives

RP: If we assume for a moment that Eric is right to call for IRs to be retired what, in your view, are viable alternatives? Could it be that, as we see commercial providers like ResearchGate and Academia.edu offering more and more flexible and appealing solutions for researchers wanting to self-archive their papers, and as we see libraries starting to outsource key functionality of the IR to traditional publishers like Elsevier, IRs will

become at best no more than search interfaces pointing users to commercial sites? Would it matter if they did?

CL: The right question here is alternatives for the IR with regard to what functions and purposes? If, for example, we are using them as data archives, there's one set of considerations; if we want vehicles for faculty to gain greater visibility for their work, certainly things like ResearchGate may play a role (though I think that the emergence of these systems is an interesting and confusing development that needs a lot more attention and analysis than it's received, and I am very deeply concerned about any proposal to exclusively rely upon or mandate the use of these systems.)

No question that there are going to be commercial software suppliers and also cloud hosted solutions for many, perhaps all, of the IR functions, and some of these are or will be excellent choices for many institutions.

There's nothing wrong with running these on a commercial provider in the cloud – with the caveat that to the extent that the institution wants or needs to have long-term control and responsibility for the data, that there's an appropriate contractual agreement and a good contingency and business failure/service continuity plan in place and maintained.

A key thing to have in mind is where we want long term responsibility for stewardship to be situated.

RP: Thank you very much for taking the time to answer my questions.