Clear blue water

The institutional repository is currently a hot topic within the library and open access (OA) communities. But what is an institutional repository? What purpose does it serve? And what issues does it raise? Richard Poynder reviews its history and current status, and concludes that while OA advocates may have spawned the institutional repository, it might be time for them to walk away from it.

While the concept of the institutional repository is not new, there has over the past year been a sudden upsurge of interest in the topic. This in turn has led to considerable disagreement about the nature and scope of an institutional repository, and its role within the academic institution.

Indeed, when JISC recently created a mailing list for those wishing to discuss issues related to the topic, the initial flurry of posts suggested that there are as many definitions of an institutional repository as there are those with an opinion on it.

So where did the institutional repository come from, and why has it become the source of so much disagreement — at the very moment when it looks set to enter the mainstream? More importantly, how should the OA movement react to these developments?

Reshaping the scholarly communication process

The seminal text on the subject was a paper — *The Case for Institutional Repositories* — written by Raym Crow in 2002.

In that paper Crow defined institutional repositories as "digital collections capturing and preserving the intellectual output of a single or multiple-university community."

Their role, he suggested, should be twofold. First: to "serve as tangible indicators of an institution's quality and to demonstrate the scientific, societal, and economic relevance of its research activities, thus increasing the institution's visibility, status, and public value"; Second: to provide tools to assist universities "re-shape the scholarly communication process".

As Crow put it, "While institutional repositories centralise, preserve, and make accessible an institution's intellectual capital, at the same time they will — ideally — form part of a global system of distributed, interoperable repositories that provides the foundation for a new disaggregated model of scholarly publishing."

Essentially, Crow envisaged that institutional repositories would enable universities to exploit the new digital networked world to regain control of scholarly communication. This, he said, would mean rethinking the relative roles of authors, librarians, and publishers, and "unbundling" the traditional model of publishing.

As a consequence, access to research would expand, and the monopoly power of journal publishers broken — monopoly power that publishers acquired, he maintained, as a result of their insistence that researchers hand over copyright in their scholarly papers as a condition of publication. This then allowed publishers to sell the research back to universities in the form of ever more expensive journal subscriptions.

By unbundling the publishing process into its constituent parts ("Registration, Certification, Awareness, and Archiving"), and reasserting ownership of the raw material (the papers), Crow argued, universities could break the chokehold that publishers had acquired over scholarly publishing.

"The purpose of a disaggregated scholarly publishing model," he explained, "is not to destroy the current journal system, but to weaken the monopolistic impact of that system on academic institutions and their libraries."

His perspective was not perhaps surprising: a managing partner at Chain Bridge Group, Crow's thoughts were published as a position paper for The Scholarly Publishing and Academic Resources Coalition (SPARC) — an organisation created in 1997 by the Association of Research Libraries (ARL) "to be a constructive response to market dysfunctions in the scholarly communication system."

It was clear that something had to be done: With static or falling budgets, libraries were struggling to cope with the escalating cost of journals. Between 1986 and 2000, for instance, serial prices increased by 196%, compared to a rise in the Consumer Price Index of just 57%. The consequent "serials crisis", SPARC complained, had "reduced dissemination of scholarship and crippled libraries".

Crow's thesis was that if researchers retained copyright in their papers (merely granting publishers a non-exclusive licence), and deposited them in institutional repositories, the publisher's role could be restricted to activities like managing the peer review process, creating value-added "overlay" journals based on the content of the repositories, and providing services like "citation linking, controlled vocabularies, and the like". And this would allow universities to restore a more equitable power balance.

In other words, institutional repositories would allow universities to create a more cost-effective model, and force that model on publishers.

The evidence so far, he argued, "suggests that the resources required [to create the necessary infrastructure for the new model] would represent but a fraction of the journal costs that libraries now incur and over which they have little control."

In short, institutional repositories were viewed as a way in which librarians could address the *affordability* problem posed by the constantly rising, and eventually unsustainable, costs associated with buying serials.

Self-archiving

But while the library community may have produced the defining document on institutional repositories, the concept was originally developed by academics.

Crow's ideas, for instance, owed a great debt to the e-print service arXiv, which had been developed in 1991 by Los Alamos physicist Paul Ginsparg. A central subject-based repository where researchers could self-archive preprints of their physics papers, arXiv had by the time Crow wrote his paper in 2002 been widely embraced by the physics community and spawned a number of imitators. Ginsparg had also published a number of papers and talks in which he had developed the idea of overlay journals.

To be fair, Crow did not claim to be original. As he acknowledged, "Discipline-specific digital repositories for high-energy physics and mathematics (arXiv); economics (RePEc); cognitive science (CogPrints); astronomy, astrophysics, and geophysics (NTRS and ADS); and computer science (NCSTRL) [have all] evolved within those specific research communities as digital extensions of existing peer-to-peer research communication practice."

The problem, he added, was that while "such e-print initiatives are frequently cited as successful examples of author self-archiving, discipline-specific repositories have not enjoyed similar success in academic disciplines without established preprint traditions. Therefore, while discipline-specific repositories support some research communities, they provide only one component in the evolving structure of scholarly communication."

In short Crow concluded that unless a critical mass of papers across a broad range of disciplines was accumulated in the network of repositories he envisaged, then his model would not provide a viable alternative to the traditional system, and so would not be capable of breaking the publishers' monopoly.

He also reasoned that creating this critical mass would be more achievable if it were driven by researchers' own institutions, rather than waiting — inevitably in vain — for those disciplines without a preprint tradition to discover for themselves the benefits of self-archiving. Universities, he concluded needed to create their own repositories, and encourage, or perhaps mandate, faculty to self-archive their papers in them.

Once again, however, Crow's thinking had been derived from academics themselves. Back in 1994, for instance, cognitive scientist Stevan Harnad had publicly posted a "subversive proposal" urging all researchers to follow the lead of physicists and self-archive.

As he put it, "For centuries, it was only out of reluctant necessity that authors of esoteric publications entered into the Faustian bargain of allowing a price-tag to be erected as a barrier between their work and its (tiny) intended readership, for that was the only way

they could make their work public at all during the age when paper publication (and its substantial real expenses) was their only option. But today there is another way, and that is PUBLIC FTP."

Harnad's proposal assumed that, rather than seek out central discipline-based archives, researchers should make use of their own institutional resources. This was not yet the concept of an institutional repository, but aside from a brief flirtation with the central archive CogPrints (which he founded in 1997), Harnad has consistently argued that institutionally based archiving is the best way of breaking the "Faustian bargain."

Certainly his proposal (and subsequent activism) played a key role in the creation of the open access movement, particularly the self-archiving school of OA.

An eyeball thing

But while Crow and Harnad may have agreed on what needed to be done, their aims were very different. For the only aim of archivangelists like Harnad is to maximise the impact of published research by making it as widely available as possible. For them it is an eyeball thing: ensuring that as many fellow researchers as possible read their paper, with the aim of maximising its impact and increasing the number of times it is cited. As such they have no interest in "reshaping the scholarly communication process". Likewise, they are not concerned with library budgets.

The sole objective, Harnad maintains, should be that of "making the full-texts of all published, peer-reviewed research journal articles accessible online toll-free for all would-be users, webwide, in order to maximise their research usage and impact."

Moreover, to achieve this, he insists, all that is necessary is for "the author to publish the article in a traditional journal (with the usual copyright agreements) but also to make his own final draft freely accessible online by self-archiving it on the web, free for all (usually in his own institutional repository)."

As such, Harnad has always claimed that self-archiving "is not a substitute but a supplement [alongside publishing the article traditionally] provided for those who cannot afford the costs of the published version."

For this reason, Harnad has also been far more interested in encouraging the archiving of postprints (authors' peer-reviewed final drafts, accepted for publication), rather than unrefereed preprints.

In short, Harnad and fellow archivangelists have never been interested in the issue of *affordability*, only of *impact*.

Why is it important to improve research impact? Because improved research impact, Harnad says, not only maximises its benefits to researchers and their institution in terms of prestige, prizes, salary, and grant revenue "but it also maximises its benefits to research itself (and hence to the society that funds it) in terms of research dissemination, application and growth, hence research productivity and progress."

Interoperability

By the time Crow wrote his paper in 2002, academics had already also played a key role in developing the necessary infrastructure to enable the "global system of distributed, interoperable repositories" that he outlined.

Three years earlier, in 1999, for instance, Ginsparg had led a call to encourage author self-archiving in subjects other than physics. This was followed up several months later by a meeting in Santa Fe, New Mexico, with the aim of generating "discussion and consensus about interoperability of publicly available scholarly information archives."

The logic was that as repositories began to proliferate there was a danger that they could end up needlessly replicating each other. It would also mean that users would have to learn multiple interfaces. For these and other reasons it was felt there was a need to develop tools and protocols to allow repositories to copy content from each other, and to work in concert on a distributed basis.

Moreover, since it looked likely that many second-generation repositories would be developed around institutions, rather than disciplines, the possibility of duplication was likely to increase, since researchers would be able to choose to archive in an institutional repository, in a disciplinary repository, or in both.

At the time of the Santa Fe Convention, for instance, MIT was planning its own institutional repository, and it was subsequently announced that Hewlett-Packard would provide $1.8 million in funding to assist MIT create a digital archive "capable of holding the approximately 10,000 articles produced by MIT authors annually."

Likewise, the University of California was planning its eScholarship Repository, which would be created courtesy of a grant from SPARC.

Universal Preprint Server

When delegates gathered in Santa Fe they were presented with the concept of something called the Universal Preprint Server (UPS), intended to provide the kind of interoperability felt necessary.

Rather than proposing a model based on cross-searching multiple archives using a protocol such as Z39.50, UPS assumed a metadata harvesting approach — an approach proposed because early experience suggested that cross searching does not scale well, partly because the search service degrades to the level of the slowest and least reliable server in the cross-search set.

Persuaded by the UPS approach, delegates agreed that they should work together to develop a protocol that would allow repositories to expose metadata about the documents held in them.

Key to this approach was the concept 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 using the metadata harvesting protocol.

One obvious service third parties could provide would be a search service. By exporting metadata from repositories, such services would allow users to search on a central archive of aggregated metadata from multiple repositories, rather than conduct real-time simultaneous cross-searches. Amongst other things, this meant that users would not be searching on the documents themselves, but on what was essentially a collection of bibliographic records of those documents.

Soon after the Santa Fe Meeting it was also decided to re-brand the UPS as the Open Archives Initiative (OAI), and the "Santa Fe Convention" as the OAI Metadata Harvesting Protocol, both in order to avoid confusion with United Parcels Service (now usually shortened to UPS), and in acknowledgement of the increasing wish to self-archive postprints as well as preprints.

It is worth noting that had the Santa Fe Convention not decided to create its two-tiered system, the notion of unbundling the publishing process may not have arisen. Crow's paper, after all, assumed that data providers (universities) would provide the raw material for service providers (publishers) to create valued-added services using the content archived in institutional repositories.

But while the OAI approach provided the theoretical benefit of enabling a limitless number of value-added services to be developed on top of repositories, it was quickly apparent that once you decouple content from its metadata, then the content itself no longer has to be present in the repository either. Indeed, it could be held somewhere else all together, possibly behind a subscription firewall!

To put it another way, some of those attending Santa Fe realised that interoperability could be far more general and powerful than just self-archiving scholarly papers. It potentially applied to the metadata for all digital content (not just text), and regardless of whether it was open access or not.

As such, some of the assumptions built into OAI were at variance with the objectives of OA advocates, whose explicit aim was to expose the full-text of their papers (open access), not just metadata about them. OAI, by contrast, was based on only metadata, not open access.

This turn of events no doubt owed a great deal to the growing interest that librarians had begun to take in institutional repositories. In fact, the Santa Fe Meeting was sponsored by a number of library organisations, including the Council on Library and Information Resources, the Digital Library Federation, SPARC and ARL.

Certainly, the approach adapted at Santa Fe could be viewed as a classic library approach, where as much (if not more) of the emphasis is placed on pointing to objects as it is to making those objects directly, or indeed freely, available.

"It was clear from the outset that what would eventually become the OAI metadata harvesting protocol, intended to make OAI archives interoperable, would also have more general applicability to digital archives of any sort, whether Open Access or not," explains Harnad, who attended Santa Fe. "So 'Open Archives' partly parted ways with what later came to be called 'Open Access'.

In any case, the emphasis on open metadata, rather than open access, was to prove symptomatic of the tension that was to later arise between archivangelists and librarians as the institutional repository movement developed.

Subsequent events, indeed, suggest that the primary concern of most librarians has always been to resolve their budgetary problems, not necessarily to remove access barriers. And as we shall see, they increasingly came to see OAI and institutional repositories as a way of also addressing the growing need to build digital libraries.

Nevertheless OAI clearly provided something for everyone, and those with existing archives such as Ginsparg, Harnad, and Thomas Krichel (an economist who founded RePEc) willingly co-operated with library automation specialists like Herbert van de Sompel and Michael Nelson, librarians like Rick Luce, and researchers at Cornell and Southampton universities, as well as at Los Alamos National Labs, to develop and test what was to become the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), and v1.0 was released in 2001.

As we have seen, OAI-PMH was designed to enable the harvesting of metadata in institutional repositories. What was clearly also required was the necessary OAI-compliant repository software with which to build institutional repositories. In September 2000, therefore, Southampton University launched the EPrints software, which was later made open source and is today the most widely used repository software.

However, while specialised search engines like OAIster (launched by the University of Michigan in 2002, and which today harvests metadata from 604 institutions to create one virtual archive of over 6.5 million records) quickly developed, the anticipated flowering of value-added services (including, for instance, peer review services and overlay journals) has not materialised.

This then was the back-story to Crow's paper, which while borrowing heavily from work that had been done by archivangelists, consolidated the process of bending the objectives of OAI towards the needs of librarians.

In 2000, however, archivangelists felt that their primary objective had been achieved. OAI had created the necessary infrastructure to allow a network of interoperable institutional repositories to be created, EPrints had provided the all-important repository software, and soon specialist search engines like OAIster were aggregating the content to enable users to search multiple repositories by means of a single interface — thereby allowing researchers to maximise the impact of their papers by making them freely available on the Web.

Moreover, all that was required of them was to continue publishing in the traditional manner, and then self-archive the full text of their papers.

Open Access

It was soon apparent, however, that aside from physicists, computer scientists, and between 5% and 25% (depending on discipline) of researchers in other subject areas, authors of scholarly papers remained oblivious to the benefits of self-archiving. Clearly a major educational effort was required.

This had begun as early as 1998, when Ginsparg demonstrated arXiv to life scientists at the Banbury Center, in Cold Spring Harbor, NY. Indeed, the logic of what he had to say was sufficiently self-evident to those present that by May 1999 the then director of the US National Institutes of Health Harold Varmus had proposed a radical plan to create an arXiv-like service for all NIH-funded research. (NIH is the largest funder of medical research in the world).

Initially called E-BIOMED, Varmus' proposal sparked such virulent opposition from publishers, however, that when it was launched in February 2000 as PubMed Central the service had been drastically watered-down — evidence that this was going to be a long battle.

Fortunately, help was at hand. In February 2002, financier and philanthropist George Soros' Open Society Institute (OSI) provided $3 million in funding to launch the Budapest Open Access Initiative (BOAI).

BOAI was to be the birth of the term "open access", and of the open access movement. It also provided much-needed cohesion, and a welcome fund of money, to drive the movement forward, and publicise the virtues of open access.

As a side note, it was agreed in Budapest that there are two possible roads to OA. BOAI-1 (the so-called green road) assumed researchers would continue publishing in traditional subscription-based journals, and then self-archive their papers — in other words, the classic archivangelist approach.

BOAI-2 (the gold road), by contrast, called for the creation of new-style OA journals that would charge to publish. The logic was that by charging a one-off fee on publication, papers could then be made permanently and freely available on the Web. Moreover, argued OA publishers, this could be done at a much lower cost than traditional subscription-based publishing.

The first OA publisher — BioMed Central — had been founded in 1998 by entrepreneur Vitek Tracz. Subsequently, in 2001, the OA advocacy group Public Library of Science made the decision to launch a series of OA journals too.

While there has been some tension between advocates of these two approaches, and OA publishing has generally attracted more attention, I do not plan to discuss their relative merits here, not least because OA publishers archive the papers they publish in central discipline-based repositories, not in institutional repositories.

BOAI aside, 2002 was an important year for the development of institutional repositories. In 2002, for instance, the University of California's eScholarship Repository was finally launched. And in November 2002, MIT announced that it was making DSpace, the OAI-compliant repository software developed with Hewlett Packard (and written by Rob Tansley, who had previously developed EPrints) open source.

Given the cachet of MIT and UC, it is unsurprising that these developments stimulated other institutions to follow suit, and following MIT's decision to release DSpace as open source software the Registry of Open Access Repositories (ROAR) shows a significant second jump in the number of institutional repositories (the first was when EPrints was released in 2000).

The growing interest in institutional repositories in its turn encouraged others to jump on the repository software bandwagon. In February 2003, for instance, the open-source software Fedora was launched. In addition, The Berkeley Electronic Press (bepress) developed a commercial product which — following a subsequent joint venture with ProQuest — is now marketed as Digital Commons.

Something extraordinary

The significance of these developments was not lost on observers. In February 2003, for instance, Clifford Lynch, director of the Coalition for Networked Information (CNI), and chair of the 1999 Santa Fe Convention, published a paper called *Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age*.

"In the fall of 2002, something extraordinary occurred in the continuing networked information revolution," Lynch began grandiloquently, "shifting the dynamic among individually driven innovation, institutional progress, and the evolution of disciplinary scholarly practices.

He continued: "The development of institutional repositories emerged as a new strategy that allows universities to apply serious, systematic leverage to accelerate changes taking place in scholarship and scholarly communication, both moving beyond their historic relatively passive role of supporting established publishers in modernising scholarly publishing through the licensing of digital content, and also scaling up beyond ad-hoc alliances, partnerships, and support arrangements with a few select faculty pioneers exploring more transformative new uses of the digital medium."

Published as a bimonthly report for the Association of Research Libraries (ARL), Lynch's paper (like Crow's) was essentially a librarian's view of the world. But while he too saw a connection with scholarly publishing, the debate had apparently moved on.

Unlike Crow, for instance, Lynch was not particularly interested in portraying institutional repositories as tools for wresting back control of the scholarly publishing system. Rather, he argued (echoing Harnad), "the institutional repository is a complement and a supplement, rather than a substitute, for traditional scholarly publication venues."

Indeed, he added "it dramatically underestimates the importance of institutional repositories to characterise them as instruments for restructuring the current economics of scholarly publishing."

Rather, he suggested, they should be seen as "vehicles to advance, support, and legitimise a much broader spectrum of new scholarly communications."

He added: "I want to make the distinction between scholarly publishing as it is currently practised and the broader, much more diverse, often less formal, and certainly more rapidly evolving set of practices that comprise scholarly communication; scholarly publishing is a very specific, circumscribed example of scholarly communication."

Essentially, his argument was that institutional repositories were not catalysts for reshaping the scholarly publishing system, but vehicles for expanding, enriching and extending the scholarly communication environment.

Set of services

Moreover, while Lynch (like Crow) saw institutional repositories as an opportunity for universities to showcase their research, he saw them playing a broader and deeper role, and fulfilling a greater range of functions than Crow.

"[A] mature and fully realised institutional repository will contain the intellectual works of faculty and students — both research and teaching materials — and also documentation of the activities of the institution itself in the form of records of events and performance and of the ongoing intellectual life of the institution," he said. "It will also house experimental and observational data captured by members of the institution that support their scholarly activities."

In effect, Lynch saw institutional repositories more as a species of digital library than a publishing platform. An institutional repository, he said, is "a set of services for the management and dissemination of digital materials created by the institution and its community members. It is most essentially an organisational commitment to the stewardship of these digital materials, including long-term preservation where appropriate, as well as organisation and access or distribution."

But whatever the nuances in their approach, in retrospect we can see that between them Crow and Lynch created a compelling narrative for librarians. Increasingly concerned that the digital age was threatening to disintermediate them, many concluded that institutional repositories held out the promise of enabling them to reassert their relevance.

Certainly many librarians (upon whose shoulders the task of establishing and maintaining institutional repositories invariably falls) have jumped into the institutional repository pool. Today there are around 400 institutional repositories and, according to a survey conducted last year by Lynch and a colleague (Joan Lippincott), around 40% of US doctoral-granting institutions now have some type of institutional repository operating. Moreover, 88% of those who do not yet have one plan to introduce one, or to participate in some form of consortial repository system.

Mandating researchers

But however enthusiastically librarians have embraced institutional repositories, persuading faculty to self-archive has remained an abiding problem.

Indeed, the vast majority of repositories remain depressingly empty. Listing repositories by total number of records in ROAR, for instance, reveals that halfway down the list the average number of items in most of them is a derisory 40, and the bottom 100 or so repositories appear to have no records in them at all, although this may be a consequence

of the open archives gateway celestial struggling to interpret the interface of some repositories. (And suggests that the OAI infrastructure has still to be perfected.)

ROAR also indicates that in many cases the number of documents available in full-text is as low as 50%, or even 30%. In some cases (e.g. eScholarship@Amherst) there appear to be no full-text documents available at all — a legacy presumably of the Santa Fe decision to focus on metadata rather than documents.

In the light of the empty repository phenomenon a great deal of thought has been given to ways in which researchers can be encouraged to co-operate. What better way, OA advocates concluded, than to persuade funders and governments to mandate researchers to self-archive. After all, if public money funds this research, should not governments ensure that the money is used to best effect?

There have, therefore, been growing calls to mandate open access. Initially, much of the focus was on NIH which had, after all, already created a repository (PubMed Central)

As a consequence, last May the NIH Policy on Enhancing Public Access to Archived Publications Resulting from NIH-Funded Research was introduced. Since this came in the wake of another period of aggressive lobbying by publishers, however, what had been proposed as a mandate eventually became an emasculated statement saying that NIH "requests and strongly encourages all NIH-funded investigators to make their peer-reviewed author's final manuscripts available to other researchers and the public at the NIH National Library of Medicine's PubMed Central".

Given NIH's weakened policy, the service has achieved very little, as a recent report indicates. Moreover, the request was for researchers to archive in a central repository, not institutional repositories, and within a year of publication rather than immediately.

A matter of time?

Nevertheless, many believe it to be only a matter of time before publicly-funded researchers are compelled to self-archive, and ideally in institutional repositories. In 2004, for instance, Canada's Social Sciences and Humanities Research Council (the biggest funding body of SS&H in Canada) made a commitment to open access, and launched a nation-wide consultation process regarding implementation.

Ironically, it is a private research funder that is currently leading the way: Last year the Wellcome Trust, the largest private funder of medical research in the UK, issued the world's first mandate by a major research funder. This requires that all its funded researchers make their papers freely available on the Web. Again, however, this stipulates that these are placed in PubMed Central (or its UK equivalent), rather than in institutional repositories. It also allowed for a six-month delay.

But it was the publication of a report into scientific publications by the UK Science & Technology Select Committee in July 2004 (*Scientific Publications: Free for all?)* that has been most enthusiastically supported by archivangelists. The report called on the UK government to create a network of institutional repositories, and mandate all publicly funded researchers in the UK to deposit a copy of all their articles in the repositories, thereby making their research accessible to all "free of charge, online."

The UK government chose not to act itself, but passed the decision on to the Research Councils UK (which comprises the UK's eight research councils, the main public investors in fundamental research in the UK). RCUK later published a draft policy pretty much in line with the recommendations of the Select Committee.

While the RCUK policy was (like E-BIOMED before it) greeted with furious lobbying by scholarly publishers, and at the time of writing has missed a number of deadlines for implementation with no progress yet, many nevertheless expect it to be eventually introduced — in some shape or form.

The danger, however, is that growing confusion and controversy surrounding the nature and role of institutional repositories could see the objectives of open access lost, or seriously diluted, before all the necessary mandates are in place.

To the growing frustration of self-archiving advocates, much of this confusion flows not from the publishing lobby, but from efforts by librarians (in the light of continued intransigence by researchers to self-archive) to find other ways of justifying the existence of their institutional repositories.

Broader vision

To do them justice, librarians initially tried hard to encourage researchers to self-archive. As the JISC Digital Repositories Review pointed out last year, "Enhancing access to scholarly communications has been a main driver for establishing repositories, both institutional repositories (in particular e-print archives) and subject based archives."

Moreover, as publishers have become increasingly embarrassed at their profit levels they have learned to live with the practice of self-archiving (partly no doubt because so few researchers do it!).

But as efforts to persuade faculty to self-archive have consistently fallen on deaf ears, so librarians have sought other ways of filling their repositories. As a result, the concept of the institutional repository has both broadened and deepened, with the objective of creating an e-print archive increasingly giving way to visions of Lynch-style digital libraries, or Crow-like publishing platforms.

As [Carol Hixson](), who is responsible for the institutional repository at the University of Oregon (based on DSpace software) [puts it]() "Although the University of Oregon's institutional repository, named [Scholars' Bank](), began with the intention to provide an archive and distribution system for faculty research in digital form, it now closely resembles Lynch's broader vision."

In fact, she adds, today only around 18% of the 1,900 items archived in Scholars' Bank has been authored by faculty. The reminder, she says, consist of a hodgepodge of "campus and departmental newsletters; scholarly journals authored or edited by members of the UO faculty; student terminal projects, class papers, honours theses, or dissertations; campus administrative records; campus planning documents; Oregon city and county planning documents harvested from local government web sites; finding aids to manuscript collections owned by the library; electronic texts of Renaissance materials; and much more."

Not untypical

Oregon's story is not untypical. When, in 2003, for instance, the University of Toronto created the [T-Space]() repository (also based on DSpace software) the aim was primarily to provide an e-print archive.

The problem says researcher [Leslie Chan](), who has been a driving force behind the repository and was one of the original signatories of the BOAI, is that "uptake by faculty has been very slow." In the meantime, the vacuum has begun to be filled with other types of content, and today T-Space also includes images, datasets, learning objects, eBooks, text, audio and video.

Likewise, when in 2004 the [University of Pennsylvania]() launched its Digital Commons-based *Scholarly*Commons@Penn repository, the aim was very definitely to increase access to the university's research. To that end, librarians at Penn have actively encouraged faculty to self-archive. For instance, says engineering librarian Mary Steiner, "we encourage our subject liaison librarians to make contact with their respective constituencies and assist with recruiting content."

As part of this marketing effort, Penn library recently produced an attractive [leaflet]() for faculty. The leaflet reveals, however, that the library is no longer canvassing for e-prints alone, but for a range of other materials as well — including data sets, computer code, simulations, student scholarship, and multimedia.

In short, in the light of the disappointing self-archiving rates, librarians have looked elsewhere for content, including harvesting documents from local government web sites, loading Renaissance e-texts, and archiving videos of university events.

Consequently, these repositories are being asked to fulfil a much wider range of functions than initially envisaged. In fact, concludes the JISC Digital Repositories Review, they now vary considerably in both content and functionality, from providing access to peer-reviewed literature and theses, to archiving teaching materials, student portfolios, conference presentations, and non-research academic work to, in some cases, hosting non-academic content like locally-produced blogs.

In addition, reports JISC, they are increasingly expected to act as corporate information management tools (records management and content management systems) and data sharing platforms (e.g. for the re-use of research data and learning objects).

Just as in a physical library…

But it would be wrong to suggest that this process is being driven by librarians alone. In fact, says Ann Okerson, a librarian at Yale, rather than actively seeking material to collect, in some cases librarians are simply responding to demands that they archive and curate a widening range of digital material. What better place to put this data, they no doubt reasoned, than in their near-empty institutional repository?

Regardless of their wishes, therefore, librarians are often having their priorities set for them. "For instance," says Okerson, "the library at Yale is responsible for the university archive, and since the official documents of the university — memos from departments and the president's office, videos, and recordings of university events — are increasingly being developed electronically it is our responsibility to preserve all this. So the first gun to our heads is the need to archive university materials."

The second gun, she says, is classroom support. "We found that there are a whole lot of things being created by faculty for classroom use — syllabi, notes, outlines, lectures, images, illustrations etc. — that has to be archived somewhere. Again, we have had to respond to that."

The point, she adds, is that "While the IT department is generally responsible for setting up systems, and providing the tools for faculty to create digital material, it is not their job to provide long term curation and preservation of those materials. That is the job of the library; and there is therefore a consensus on our campus that it is our role to preserve them for the long term."

The third gun to librarians' heads, she says, is the need to deal with materials that are created digitally in the library. "If we are digitising texts or images, for instance, we have to find somewhere to put that material to ensure long term curation."

For Okerson, therefore, an institutional repository is primarily a digital version of the traditional library: "Just as in a physical library you have printed materials, an institutional repository is a place where you keep electronic materials."

But where in this picture does e-print archiving fit? "I presume that many faculty on our campus put their papers up on their web sites, and perhaps deposit them in a [central] preprint service," replies Okerson vaguely. "Certainly we haven't aggregated them in any way."

For self-archiving advocates this demonstrates the increasingly low priority now being given to OA in many institutions, and the alarming lack of interest some librarians now have in self-archiving.

Digital library

What it clearly also demonstrates, however, is that there is a growing need for institutions to create a digital library. Moreover, given the problems they have had persuading faculty to self-archive, it is unsurprising that many librarians have come to view the institutional repository as an ideal platform on which to build that digital library.

Indeed, this was essentially the conclusion Lynch reached three years ago, although he did not use the term "digital library". As he puts it today, "As an almost completely independent issue from open access, the scholarly community needs to develop systems to manage, disseminate, and preserve large amounts of complex digital content that is increasingly being produced by changes in the practice of scholarship and scholarly communication, and by changes in functioning and practices of our institutions of higher education. This is a problem that is growing in scope, scale and urgency year by year [and] I don't think that most research universities can ignore [these] developments, which drive the need for full scale institutional repositories."

For that reason, he adds, "The kinds of institutional repositories that I am focused on are services designed to address many of these needs."

For Harnad this goes to the heart of the problem, since he believes that an institutional repository should be viewed as a creature entirely different from the digital library.

As he frequently avers, for instance, an institutional repository has nothing to do with preservation. Its sole purpose is to provide access to peer-reviewed papers that have been published in traditional journals.

After all, he adds, since self-archiving is only intended to complement, not replace, the act of publishing, there is no need for an institutional repository to fulfil any preservation role for the authors' self-archived final accepted drafts (postprints) of published articles. That, he says, "is a matter for the publisher of the journal in which the self-archived article was published plus the institutional subscribers to that journal — perhaps along with national deposit libraries — to take care of, not the author's institution, which is

only making copies of papers available in order to ensure there are no access barriers to research produced by its faculty.

"The preprints and postprints that are being self-archived by their authors in their institutional e-print archives today are intended to maximise impact by providing immediate open access," he adds. "They are merely open-access *supplements* to the toll-based primary literature at this time, not *substitutes* for it.

Harnad is quick to add, however, that he is in no way opposed to the concept of the digital library. "It is not that storing and preserving every digitised object under the sun is not a good idea," he explains. "It is just a question of priorities. For universities and research institutions, the immediate priority is this: Scholarly and scientific research usage and impact have been needlessly lost, cumulatively, since paper publication first began, because paper costs and distribution necessarily meant that many would-be users could not afford to access and use most research output. This has always meant a great loss of potential research impact and hence research progress to researchers, their institutions, and to research itself.

The fear of self-archiving advocates, therefore, is that if people believe that an institutional repository is the same as a digital library, then open access will become bogged down in the difficult, time-consuming, and complex issues of preservation; issues, moreover, for which no one today has a solution.

Perhaps he has a point: even with its considerable resources, Yale today is still operating in crisis management mode. "What we have done initially is to create a 'rescue repository'," says Okerson. "This has large amounts of storage hardware and software, and we are loading into this everything that we want to save. It doesn't have a lot of functionality: the aim is simply to have a place to put things right now to keep them safe and secure while we sort out all the finer details — access issues, meta data, long term preservation etc."

Publishing platform

And as the problem of filling institutional repositories has continued to frustrate librarians, some have sought to short-circuit the process of reshaping the scholarly communication process outlined by Crow, and are attempting to reengineer their institutional repositories as publishing platforms.

Nowhere is this development more apparent than at the University of California. Like many other universities, the initial focus at UC when it set up its eScholarship Repository was to encourage scholars to self-archive their papers; and to this end an EPrints server was installed. When researchers failed to use it, however, the EPrints server was thrown out and replaced with a bespoke version of the bepress software.

As Catherine Candee, director of publishing and strategic initiatives in the Office of Scholarly Communication at the University of California, puts it "We found it so, so, so difficult to get faculty even to test the EPrints software that we abandoned the idea of providing a platform for faculty to individually publish their own works.

Instead, UC decided to experiment with new publishing models. And the attraction of bepress, explains Candee, was that it "could also support peer-reviewed publications."

Today, therefore, UC is self-consciously using the eScholarship Repository as "a kind of stepping stone, or a means to an end, to changing the [scholarly publishing] paradigm". With this in mind, it has been used to launch both new scholarly journals and new monographic series. The aim, says Candee is "to share the load" with UC Press.

The end game, she adds (echoing Crow) is to re-assert ownership of the research the University produces, and move to a position where publishers merely provide value-added services on top of papers archived in the eScholarship repository. "I really think we are heading towards a layering of services, where an awful lot of raw content will be managed more responsibly by universities, and publishers and aggregators will develop all kinds of services to add value to that content."

Once again, self-archiving advocates have been quick to criticise. "UC moved away from EPrints and postprint self-archiving because of the extremely low level of spontaneous uptake by UC faculty, assuming the low uptake was because it was "'too difficult'", says Harnad, adding: "It is far more likely that the low uptake was because UC did not adopt an institutional self-archiving mandate. Those institutions that have done so have dramatically higher self-archiving rates."

In other words, rather than give up on self-archiving, or waiting for research funders or governments to introduce mandates, universities should introduce mandates themselves.

Is this a better approach? Perhaps. The growth of deposits in the institutional repository at Australia's Queensland University of Technology (which introduced a mandate in January 2004) does suggest that an institutional mandate can make a difference.

"No Australian university with a voluntary policy collects significantly more than 15% of the Australian Department of Science Education & Technology reportable content and most much less," says Arthur Sale, Professor of Computing Research at the University of Tasmania. "This is consistent with international data for which 15% is accepted as an average limit.

Queensland University of Technology (QUT), by contrast, adds Sale, has seen self-archiving rates four times higher than any other Australian university. A difference, he attributes to the mandate.

If QUT's experience is an accurate predictor of how researches respond to a self-archiving mandate it is certainly a positive indicator. Indeed, surveys conducted by the

UK-based Key Perspectives, suggest that 95% of researchers would comply with a mandate.

The problem is that to date just four institutions in the world have introduced a mandate.

Kept separate

But UC is not alone in its publishing ambitions. When last year Purdue University decided to install Digital Commons, says bepress president Greg Tananabaum, the decision was actively co-driven by the director of Purdue University Press Tom Bacher, who "sees it as a tool to help further their publishing efforts."

For Purdue, however, these plans appear to be a little sensitive. When I contacted the library I was told my questions would have to be referred to higher management. Higher management subsequently declined to comment, and instructed my interlocutor not to answer any of the questions either.

While critical of these developments, self-archiving advocates are keen to stress that they are not opposed to universities experimenting with self-publishing — so long as it does not work against the objectives of OA.

No one is saying that libraries should not experiment with digital publishing, says Harnad, but it is not helpful to conflate this activity with open access. "They are not incompatible *as long as they are kept separate*, not conflating university OA self-archiving of their own research output, with university publishing — whether of their own output or that of other universities, like any other publisher."

Does this pose a threat to OA? Maybe, since there is a danger that a university's publishing ambitions could run counter to open access. For instance, while MIT's DSpace now contains around 10,000 theses completed between 1879 and today, these do not appear to be available on an open-access basis, but are being sold by MIT, at prices ranging from $29 to $65.

In any case, says Hixson, rather than providing an alternative to the traditional system, it seems likely that institutional repositories can — in the short-term at least — aspire to being little more than an adjunct to it.

While insisting that this is nevertheless a promising development, Hixson concedes that, "That promise, however, may not be in fomenting a revolution in scholarly publishing, as many hope, but rather in transforming scholarship by emphasising and collecting the material 'at the edge'".

In other words, making grey literature and "ephemera" (material that has historically been left sitting on dusty shelves: theses, dissertations, conference proceedings etc.) available to a much wider audience over the Web.

What those universities attempting to create new publishing platforms on top of their institutional repositories have apparently failed to appreciate, critics point out, is that before you can aspire to reshaping the scholarly communication process in the way envisaged by Crow you need first to have a critical mass of scholarly papers in your network of interoperable repositories. Until that point you are simply tinkering at the edges of publishing.

UC apparently realised this when, a year ago, it returned to the task of acquiring postprints for the eScholarship Repository. This time, however, it has taken on the task itself, tracking down faculty papers and emailing the authors to ask them to provide a copy of the paper in question.

Today, however, only 1,000 papers have been acquired in this way — a small fraction of the number of papers UC's 16,000 faculty and researchers will have authored.

Moreover, conscious that doing this raises copyright issues that self-archiving does not, UC is having to contemplate the inevitably controversial step of requiring faculty to grant rights to UC prior to assigning copyright to publishers. The recently published Scholarly Work Copyright Rights Policy, for instance, proposes that UC faculty in future "routinely grant to The Regents of the University of California a limited, irrevocable, perpetual, worldwide, non-exclusive licence to place the faculty member's scholarly work in a non-commercial open-access online repository."

If passed by UC Senate, this would undoubtedly allow the library to acquire and archive faculty postprints in a more effective way than currently. Whether it would fill institutional repositories more quickly than a mandate is a moot point.

Harnad, for one, does not think so. UC, he suggests, is doing things back to front. "A postprint mandate should not come behind a blanket copyright retention mandate. That is like making a local emission-reduction plan's adoption contingent on first getting all nations to agree to sign the Kyoto Accord!"

Mental trap

What is self-evident is that as institutional repositories grow in functionality and scope, so do the challenges they face. In fact, argue self-archiving advocates, assuming that institutional repositories can provide a solution to every technical need is nonsense.

"The mental trap that this misuse of the term 'institutional repository' leads into is thinking that taking one piece of software (one database if you like) can serve all functions," complained Sale on the American Scientist Open Access Forum

He added, "This is as stupid as thinking that a university's central administration can use one piece of software and one database for its financial records, its personnel records, its student database, its timetabling etc. No administrator would be that foolish, but some people do not seem to recognise the same is true of broad digital library functions.

Are archivangelists simply being too prescriptive? Perhaps an institutional repository can indeed be a cure-all?

"Unlike some who are rather monotheistic in their approach to the institutional repository and OA, we are rather open to multiple pathways, tools, and solutions," says Chan. "We've barely scratched the surface as far as OA and institutional repositories are concerned and no one knows what the future of scholarly communication will be like, except that it will be very different, especially in view of all the 'digital natives' who are attending universities and will become the next generation of researchers and faculty."

But what archivangelists fear is that if self-archiving becomes inextricably linked to the long-term, complex, and costly tasks of building new publishing platforms — or creating digital libraries — then what is essentially a straightforward, low-cost and immediately feasible activity will be indefinitely delayed while time-consuming experiments of uncertain outcome are undertaken.

The reality

Librarians disagree. What self-archiving advocates fail to realise, says Okerson, is that libraries have to serve a range of different interest groups, not just researchers. "If you are in a university that serves multiple purposes, and which prides itself not only on its scientific research but on undergraduate teaching you can't ignore the demands that makes."

Moreover, says the executive director of SPARC Heather Joseph, the assumption that all research can be made freely available on the Web without financial cost is naïve. "I can sympathise with the notion of wanting to have access to all the research that is out there, and if everyone searched their souls that would be the ideal goal."

She adds, however, "The reality is that we have to look at the entire system, and say: 'How do we distribute that research if we ignore the first half or two thirds of the publishing system?' Sure, we could say 'If only everyone just deposited their articles in a repository the problem would be solved.' But that would only be to look at part of the problem. It ignores the issues of cost, the issues of responsibility, the issues of scalability

of the system, and it ignores the sustainability of the system. These issues have all got to be addressed; and it will inevitably take time to work them out."

This, respond archivangelists, is just plain wrong. Since self-archiving means simply placing a raw copy of the author's final draft of a published scholarly article on the Web in order to remove all access barriers, and so increase its impact, there are no costs involved, and since today 93% of journals endorse author self-archiving, it is also an uncontroversial activity that publishers evidently support.

The reality, they counter, is that self-archiving is simple, cheap and harmless; and for that reason it should be viewed not as part of the publishing process, not as a component of the digital library, but as a separate and entirely self-contained activity.

Certainly attempts to reengineer institutional repositories as digital libraries, or as publishing platforms, is hugely challenging. It is also a costly business. For that reason, say archivangelists, it is both misleading and unfair to link self-archiving with far more expensive projects.

In a recent issue of the INASP newsletter, for instance, Okerson estimated three-year start-up costs for hardware and software alone for the kind of institutional repository Yale has in mind at over $300,000.

For those like UC who prefer to outsource the task, a Digital Commons solution will cost between $19,950 and $34,500 a year, depending on the size of the institution.

Compare this, says Arthur Sale (who set up the institutional repository at the University of Tasmania), with the costs of a simple EPrints installation. "It takes about [a one-time start-up cost of] $A3,000 to $A10,000 [$US2,2600-$US7,500] for hardware, nothing for the software since it is open source, and say two to three week's work to get the repository up and running. For the next two years, it may require say a half-time commitment by a librarian to manage and promote the repository, but thereafter the time requirement should be easily absorbable into the university's budget."

Unsurprisingly, these differences of view have led to an increasingly bitter debate over what an institutional repository is, with librarians and archivangelists taking very different positions.

"It is time we reverted to using the term 'institutional repository' for open access functions to do with deposits of research output," says Sale, adding that this should also embrace "digital theses and eResearch (research datasets) to that research output."

Librarians disagree, complaining that the open access movement is seeking to redefine what an institutional repository is to suit its own ends. "It is unfortunate that the term institutional repository has come to mean something narrower," comments Candee.

However, while librarians may have invented the term institutional repository, we should not perhaps forget that it was researchers who created the e-print archive, out of whose womb the institutional repository sprung.

"In 1999 clearly the preferred term for what the (eventual) OAI people were aiming at was *archives*," says Harnad. "Digital archives, places you could deposit digital stuff, and also where you could make it (or at least its metadata) *open*, as in 'Open Archives' and the 'Open Archives Initiative'."

For some librarians, he adds, this was not enough. "They wanted another name for what we had started calling Institutional OA Archives, and I started first hearing the word 'Institutional Repository' (sometimes 'Institutional Depository') more and more."

Largely orthogonal

In retrospect, Lynch concedes that matters are more complicated than he assumed when he wrote his paper. "As I've looked more at various institutional deployments and planned deployments, I think that the distinction between digital libraries, digital collection management systems, digital archives, and institutional repositories is less clear than I might have felt in 2003."

Consequently, he says, "it's important to recognise that there are a number of different definitions of 'institutional repository' around. In [my] paper, I gave mine. The kinds of institutional repositories I think about really are concerned with long-term preservation of digital content as a critical part of their function [and] I feel strongly that I described a class of services that are going to be essential for research universities in the digital age, whether you agree that institutional repository is the right name or not."

He agrees, however, that the needs of OA are separate and somewhat more modest. "If all you want to do is author self-archiving, I suspect that there are likely to be cheaper and more quickly deployed solutions. [After all] if we don't stress too much about long term preservation, a system to support self-archiving at an institutional level should be a pretty inexpensive service to build."

One option, therefore, he suggests, "is to do one of these fast and cheap and then get on with the hard problems, and fold the papers back in to the fully featured institutional repository later."

For this reason, he says, "I'd urge you to do whatever you can to dispel any notion that there's an opposition between open access advocates and institutional repository people [because] they are really separate, largely orthogonal, issues."

But is it perhaps too late to get that message across? Has the concept of an institutional repository now been so thoroughly contaminated by the concerns of librarians that any association with it poses more risks than benefits for the OA movement?

The jury is still out

There is another problem: there are reasons for thinking that the institutional repository may not even have a long-term future. For despite the current excitement surrounding it, says Tananabaum, "the jury is still out" on whether it will prove a fixture.

While there are currently around 400 institutional repositories, he explains, this is just a drop in the ocean. "Depending on your definition of what an institutional repository is, this number needs to be set against a potential figure of 30,000."

Moreover, he adds, most of these repositories are still in experimental mode. "They are just scratching the surface, testing it out, and seeing what they can do. But they are not really aggregating content, or pulling it together in any significant way."

Even if we assume they do eventually take off, he adds, we are left with a fundamental question: "What is an institutional repository going to be when it grows up?"

Indeed, so great is current confusion as to what an institutional repository is — even amongst librarians — that few commentators are brave enough to attempt to define it, usually restricting themselves to saying what it *is not*, rather than what *it is*; or merely listing characteristics and functions that have been associated with it.

Even the authors of the recently published book *The Institutional Repository* (two of whom are librarians at Edinburgh University) felt unable to propose a definition. Ironically, therefore, while the book defines what a digital library is, it does not attempt to do the same for the institutional repository. "We were conscious of the difficulty of pinning the concept of an institutional repository down with a single definition," explains co-author John MacColl.

And the more the institutional repository is asked to be a cure-all, the greater the confusion that surrounds it, and the greater the likelihood that it will become a white elephant instead.

Thus the danger for the self-archiving movement is not just that it has become negatively linked with a much larger, more complicated and expensive enterprise, but that it has been linked with one that is in serious danger of losing its way.

"The future of institutional repositories is uncertain," concedes Hixson. "The costs of developing and maintaining them are not well known; nor is it certain how committed individual institutions will remain to the effort in the long-term as the costs associated

with institutional repositories are better understood or increase. There is so far little consensus on the types of materials that are appropriately stored in such repositories and little practical development of federated searching across different repositories."

Walk away

If that's right, then surely the best thing the OA movement can do today is to walk away from the institutional repository, and re-focus the discussion on the more practical, quicker, and less costly objectives of self-archiving.

Perhaps it is time, for instance, to give up any claim to the term institutional repository, and revert to calling the place papers are self-archived the e-print archive. Alternatively, why not adopt the term favoured by those advocating for OA in developing countries — the Open Access Archive (OAA); a term also used by QUT. Failing that, why not use the term Open Access Repository (OAR), which would conform with the terminology of ROAR (Register of Open Access Repositories) and DOAR (The Directory of Open Access Repositories)?

Such a decision would no doubt be emotionally difficult to take (not least because it was the self-archiving movement that gave birth to institutional repositories), and it would not be without risk (turning away from the rising tide of interest in institutional repositories), but the institutional repository movement is increasingly looking like the kiss of death for self-archiving.

The danger, therefore, is that unless the self-archiving movement puts some clear blue water between itself and the mists of confusion enveloping the institutional repository it could end up shipwrecked. By the time the long-awaited self-archiving mandates are implemented, for instance, universities may have neither will nor resources to play their part in the process.

Alternatively, continual wrangling over the institutional repository may make mandates hard to secure in the first place, not least because public agonising over the problems associated with digital preservation, or the failure of a high-profile university publishing effort, could enable OA opponents to successfully portray self-archiving as too risky and impractical.

Maybe it is also time to think the unthinkable, and walk away from the library as well. As we have seen, librarians have a very different perspective, and a different set of priorities from the self-archiving movement. Would it not make sense to hand the OA task over to the IT people, and make them responsible to the university research committee?

This is not to demonise librarians, simply to accept that they see the world differently, and that they have higher-priority tasks to grapple with right now. Indeed, if anyone should be demonised it is the many researchers who have consistently and — let's say it

— irresponsibly refused to devote the estimated 40 minutes a year required to self-archive their papers.

To put this in perspective we should perhaps end with a couple of questions: Does it matter if it takes another fifteen years (or thirty even) before OA becomes a reality? Does it matter, in fact, if OA is never achieved?

For those who would like to see a cure for cancer developed as quickly as possible, or who would like to see the threat of global warming reduced before the sea rises by 20 feet, the answer to both questions is surely "Yes." For those who don't care, the answer will presumably be "No".

The point to bear in mind, says Harnad, is that the Internet has made it possible for the first time to remove all access barriers to research, and so holds out the promise of vastly accelerating research progress.

"Yet today only around 15% of the planet's annual research output is being made freely available on the Web, which means that the access to — and the impact of — 85% of it is still being lost, daily, cumulatively, and needlessly."

_____