Preserving the Scholarly Record
Interview with digital preservation specialist Neil Beagrie

One of the many challenges of our increasingly digital world is that of establishing effective ways of preserving digital information — which is far more fragile than printed material. What are the implications of this for the scholarly record, and where does Open Access (OA) fit into the picture?

Richard Poynder 12th August 2010

In a 1999 report for the Council on Library and Information Resources Jeff Rothenberg , a senior research scientist at the RAND Corporation, pointed out that while we were generating more and more digital content each year no one really knew how to preserve it effectively. If we didn't find a way of doing it soon, he warned, "our increasingly digital heritage is in grave risk of being lost."

In launching the UK Web Archive earlier this year British Library chief executive Dame Lynne Brindley estimated that the Library would only be able to archive about one per cent of the 8.8 million .co.uk domains expected to exist by 2011. The remaining 99 per cent, she said, was in danger of falling into a "digital black hole".

In the context of Rothenberg's earlier warning Brindley's comment might seem to suggest that very little has changed in the past eleven years so far as digital preservation is concerned. But that would be the wrong conclusion to reach. Rather, it draws attention to the fact that digital preservation is not just a technical issue.

As it happens, many of the technical issues associated with digital preservation have now been resolved. In their place, however, a bunch of other issues have emerged — including legal, organisational, social, and financial issues.

What concerns Brindley, for instance, are not the technical issues associated with archiving the Web, but the undesirable barrier that today's copyright laws imposes on anyone trying to do so. Since copyright requires obtaining permission from the owner of every web site before archiving it the task is time consuming, expensive, and quite often impossible.

Clearly there are implications here for the research community.

State of play

So what is the current state of play so far as preserving the scholarly record is concerned?

First we need to distinguish between two different categories of digital information. There is retro-digitised material, which in the research context consists mainly of data created as a result of research libraries digitising their print holdings — journals, books, theses, special collections etc. Then there is born-digital material — which includes ejournals, eBooks and raw data produced during the research process.

It is worth noting that the quantities of raw data generated by Big Science can be mind-boggling. In the case of the Large Hadron Collider, for instance, CERN expects that it will generate 27 terabytes of raw data *every day* when it is running at full throttle — plus 10 terabytes of "event summary data".

To cater for this deluge CERN has created a bespoke computing grid called the WLCG. While the costs associated with the WLCG will be shared amongst 130 computing centres around the world, the personnel and materials costs to CERN alone reached 100 million Euros in 2008, and CERN's budget for the grid going forward is 14 million Euros per annum.

Of course, these figures by no means represent preservation costs alone, and they are not typical — but they provide some perspective on the kind of challenges the science community faces.

So how is the research community coping with the challenges? With the aim of finding out the Alliance of German Science Organisations recently commissioned a report (which was published in February).

What were the main findings?

So far as retro-digitisation is concerned, the Report points out that funding is limited and "the quantity of non-digitised material is huge". Even so, it adds, there is general concern about "the sustainability of hosting" the data that has been generated from digitisation. This is a particular concern for small and medium-sized institutions.

With regard to born-digital material the Report found that the largest gaps are currently in the "provision for perpetual access for e-journals".

The situation with regard to eBooks and databases is less clear since, as the Report points out, "experience in digital preservation with these content types is currently more limited."

While the Report focused on the situation in Germany the international nature of today's research environment means that the situation will be similar in all developed nations (Although Germany does have two unique mass digitisation centres).

We should not be surprised that the German Report found the largest gap to be in the preservation of journal content. As we shall see, the migration from a print to a digital environment has disrupted traditional practices and responsibilities, and led to some uncertainty about who is ultimately responsible for preserving the scholarly record.

We should also point out that one important area the German Report did not look at is the growing trend for scholars to make use of blogs, wikis, open notebooks and other Web 2.0 applications. Should this data not be preserved? If so, whose responsibility is it to do it, and what particular challenges does it raise? As we have seen, for instance, preserving web content is not a technical issue alone. Amongst other things there are copyright issues. (Although as the research community starts to adopt more liberal copyright licences these difficulties will ease somewhat).

Another recently published report did look at the issue of web-created scholarly content, but reached no firm conclusion. Produced by the Blue Ribbon Task Force, this Report concluded: "[I]n scholarly discourse there is a clear community consensus about the value of e-journals over time. There is much less clarity about the long-term value of emerging forms of scholarly communication such as blogs, products of collaborative workspaces, digital lab books, and grey literature (at least in those fields that do not use preprints). Demand may be hypothesised — *social networking sites should be preserved for future generations* — but that does not tell us what to do or why."

Open Access

One issue likely to be of interest to OA advocates is whether institutional repositories should be expected to play a part in preserving research output.

Evidence cited by the German Report suggests that repositories are not generally viewed as preservation tools. It pointed out, for instance, that the Dutch National Library's KB e-Depot currently archives the content hosted in 13 institutional repositories in the Netherlands.

The Blue Ribbon Report, by contrast, appears to believe that repositories *do* have a long-term archiving role. It suggests, for instance, that self-archiving mandates should always be accompanied by a "preservation mandate".

The Report goes on to suggest that the inevitable additional costs associated with repository preservation should be taken out of the institution's Gold OA fund (where such a fund exists).

I emailed Paul Ayris — one of the authors of the Blue Ribbon report — to check my understanding of this. Ayris confirmed that the Report envisages repositories having a preservation function. "Yes, the Report does advocate that this change is necessary," he replied. "It is irresponsible to create and make available digital assets and not to curate them for the long-term. The Report acknowledges this and makes recommendations based on this realisation."

Many OA advocates would disagree, likely arguing that such a proposal is based on a misunderstanding. Since the papers deposited in repositories are only ever meant to be supplements to the official (publisher's) version (not the originating documents themselves) the responsibility for preservation should lie not with repository managers, but with publishers and/or the library community.

As self-styled archivangelist Stevan Harnad puts it in his self-arching FAQ: "If you are worried about the preservation of the online version, it is to its publishers and subscribing/licensing librarians that your worry needs to be addressed", not repository managers.

It is difficult enough, he adds, to get mandates passed without the added complication of having to persuade university administrators to fund long-term preservation as well. Harnad is concerned that this would serve to further slow the growth of OA.

Neil Beagrie, the lead author and editor of the German Report, disagrees. "I honestly don't think that preservation represents the major barrier to self- archiving mandates or researchers' compliance with them," he told me.

Nevertheless it is hard not to conclude that there is a potential conflict between OA and preservation.

What further complicates the picture is that many universities have come to view repositories as the logical place to store not just self-archived research papers, but a range of other material too. Since much of this additional content will probably not be replicated elsewhere, the argument that repositories should not play a preservation role as well as providing access becomes more difficult to make.

One further complication is that repositories are highly unlikely to ever be able to provide a universal preservation solution for the research literature. As Beagrie concedes, "even with 100% compliance with an open-access mandate in Germany perhaps less than 10% of the e-journal content licensed by German academic libraries would be available in German institutional repositories."

Again, we can assume that this this will be similar in most countries.

Disrupted practices

Unfortunately, therefore, we are likely to see a certain degree of conflict between OA advocates and preservationists, particularly if the cause of preservation not only muddies the OA message but also begins to devour funds originally intended to support OA.

The purpose of Gold OA funds, after all, is to assist researchers pay to publish their papers in OA journals; if some of this money is sequestered to meet the costs of preservation this will surely be viewed as a further impediment to the progress of OA.

(It does not help that there is pressure on these funds from a number of directions right now: In a recent interview with me, for instance, Northwestern University librarian and dean Sarah Pritchard said she believes Gold OA funds should be seen as a pot for a variety of different purposes, including publishing traditional print books. As she put it, "[T]hey are being called OA funds, but you could equally call them subvention funds").

As suggested earlier, the nub of the problem for the research community is that the digital revolution has disrupted historical practices and responsibilities. As Beagrie explains, "For commercial scholarly e-journals the major challenge is that libraries no longer hold the content themselves but license access to it on publishers' servers. This radically changes the model for preservation and access compared to print, when libraries held copies and preserved them to ensure on-going access."

In the print world, therefore, preserving the research literature was never seen as the responsibility of publishers. Since they routinely acquire ownership of the papers they publish (by insisting that researchers assign copyright to them as a condition of publication), and since they do not permit anyone else to copy, distribute or even store those papers, however, it would seem natural that in a digital environment the responsibility should pass to publishers. After all, if they now only buy access to a digital file, not ownership of a physical product, librarians are not able to fulfil their traditional preservation role.

But publishers have made it clear that they do not want responsibility. Some have therefore sought to offload the task to national libraries (as we saw with the 2002 deal between Elsevier and the National Library of the Netherlands). Such deals are done, we should note, on the clear understanding that it will be done on an exclusively "current basis" arrangement, and that only those with "permitted access to the library's collections" will be able to access the content.

There is logic to this kind of outsourcing: like any company, a publisher will sometimes go out of business, or be acquired. This is not something that generally happens to libraries — certainly not national libraries. Consequently libraries can provide a more secure preservation environment.

However, not all publishers seem interested in doing such deals.

Publishers' disinterest in preservation is not unique to the digital environment; it was an issue in the print world too. For that reason most countries introduced legal deposit laws requiring publishers to provide national deposit libraries with copies of every new publication.

Again, however, the digital revolution has disrupted the status quo. Legal deposit laws apply only to print publications, and most countries have not adjusted the deposit requirement for the electronic environment. Even in the UK — which passed the Legal Deposit Libraries Act 2003 in order to give British legal deposit libraries the right to claim online publications — the law has yet to be implemented. It was this problem that Brindley drew attention to earlier this year.

So what should the scholarly community be doing to ensure the survival of the scholarly record? Clearly it would help if research libraries and publishers were to mutually agree arrangements that would guarantee the content of all scholarly journals survived into the future.

What's needed, suggests Beagrie, is for libraries to negotiate "the right clauses in licences with publishers and, even more critically, that both libraries and publishers have mechanisms in place which can ensure the terms of these clauses can be delivered well into the future."

This could be achieved, for instance, if librarians tied content licensing agreements to a commitment from the publisher to deposit all the licensed content with a national library or third party service provider like LOCKKS, CLOCKSS and Portico.

But it is not clear that publishers are happy to allow libraries to set the agenda in this way. And as can be seen from the recent dispute between the University of California and *Nature*, publishers are difficult people to negotiate with.

Fall through the cracks?

We suggested earlier that there is a potential conflict between the equally desirable goals of OA and preservation. From another perspective, however, it could be argued that OA will make preservation much easier — at least as regards Gold OA (OA journals) as distinct from Green OA (self-archiving).

Unlike subscription publishers, for instance, OA journals use creative commons licences. These allow anyone to download, distribute and archive their papers. Moreover, not only do OA publishers make copies of all their papers freely available on their own web server, but most also deposit copies in independent central repositories like arXiv and PubMed Central.

OA journals, however, are still very much the minority and thus, like repositories, they are far from able to provide a universal solution.

In short, despite the progress that has been made on the technical aspects of preservation — there remains a real risk that some of the scholarly record will fall through the cracks, and disappear down Brindley's digital black hole.

It would seem that the biggest problem today is that there is no universal solution, and no one person or organisation with overall responsibility for putting one in place.

When I put this point to Ayris, he replied: "The landscape is complex and the Report acknowledges this by not making concrete recommendations as to *who* is responsible. Rather the Report wishes to

stimulate this debate amongst relevant stakeholders so that *they* can reach agreement about roles and responsibilities in this area."

Where does this leave us? As we've said, responsibility for digital preservation is currently dispersed amongst a variety of different organisations and service providers — including publishers, institutional repository managers, and third party service providers, as well as national libraries and research libraries; moreover, what preservation does currently take place is being done in an ad hoc manner — because there is no overall authority that has been given responsibility for preserving the scholarly record.

In his Report to the Alliance of German Science Organisations Beagrie argues that the solution lies in the development of internationally co-ordinated national hosting strategies. In the case of Germany, he suggests, this could be done by means of a "federated organisational and funding model" supported by the Verbünde and regional library services.

The Blue Ribbon report, by contrast, recommends that "public-private partnerships" be encouraged. As the Report puts it, the best solution would be to "create mechanisms for public-private partnerships to align or reconcile benefits that accrue to commercial and cultural entities. These agencies can play a critical role in convening stakeholders, sponsoring cooperation and collaboration, and ensuring representation of all stakeholders."

German preservation specialist Eberhard Hilf, however, is sceptical about the practicality of relying on public-private partnerships in Europe. "While that might work in the USA, we do not see it working in Germany," he emailed me.

Hilf believes that long term archiving should remain the task of national libraries and/or subject specific libraries. In any case, he adds, "This is exclusively a task for government-funded institutions."

Hilf has given a great deal of thought to preservation: In 2005 he too authored a report on the topic for nestor, an organisation funded by the German government. "Our report gave all the arguments as to what should be done and by whom and how," he told me. And yet, he adds disappointingly, nothing came of his report.

Thorny issues remain to be resolved

All in all, it would seem there is likely to be a great deal more debate, and some disagreement, before a secure preservation environment is established for the scholarly record. While many of the technical issues have been addressed, a number of thorny issues remain to be resolved.

Beagrie agrees there is still a danger that some scholarly content could fall into Brindley's digital black hole. "We have made huge progress in the last 5-10 years," he says. But he adds: "There are real risks going forward. I think this is going to be even more acute as public funding across many developed countries is squeezed hard."

One type of scholarly content that is surely not being given sufficient attention today is the growing amount of material created in blogs, wikis, open notebooks etc. Currently most discussion is focused on how we preserve electronic versions of traditional forms of scholarly communication like journals and books. And yet in the future these traditional forms will likely be only a small component of the overall scholarly record.

Be that as it may, Beagrie suggests that we treat the current challenges as opportunities to be more creative. For instance, he says, there is great scope for automating many of the essential preservation tasks, and for exploiting new technologies like cloud computing.

More radically, he believes it is time to end the practice of producing both print and electronic versions of scholarly publications. If the print versions were discontinued, he points out, the research community would free up valuable time and money to be able to redirect more of its energies to tackling the obstacles that continue to impede the creation of a secure preservation environment for scholarly digital content. Above all, he adds, a great deal more collaboration is needed between the various stakeholders.

Below I attach a detailed Q&A interview with Beagrie in which he discusses these and other issues.

Neil Beagrie, director Charles Beagrie Limited

The interview begins …

RP: Can you start by saying something about yourself, and your interest and experience in digital preservation?

NB: My interest in digital preservation started some 15 years ago when I was Head of Archaeological Archives in the National Monuments Record of the Royal Commission on the Historical Monuments of England.

Archaeology was one of the first disciplines to shift to large-scale use of computing and digital records. The information from archaeological excavations is unique and irreplaceable and needs to be kept accessible digitally in perpetuity. It was a new and very challenging problem and the whole area of long-term management, preservation and access to digital scholarship, and making sure we can achieve it, has been part of my career ever since.

I have always worked across a wide range of disciplines not just the arts and humanities but also the social and physical sciences. Amongst other things, I was responsible for establishing the Digital

Preservation Coalition which now has a membership of some 35 major organisations; and the JISC digital preservation programme which set-up the Digital Curation Centre.

RP: Tell me about Charles Beagrie Limited?

NB: Charles Beagrie Limited was established in 2002. We are a specialist research and consultancy firm. Our staff and associates have senior experience in research and business, academic libraries, university computing services and data archives.

RP: What sort of clients do you have?

NB: We work with clients in the UK and internationally. Recent examples of our work include the Keeping Research Data Safe projects for JISC, which have investigated the costs and benefits for long-term preservation of research data; in the USA we recently did a sustainability plan for Dryad — which is an innovative new open-access repository for supplementary data associated with journal articles being established by a collaboration of scholarly societies and journals, and funded in its pilot phase by the National Science Foundation; and in Europe, and in March we produced a report on a Federated strategy on perpetual access and hosting of electronic resources for Germany for the Alliance of German Science Organisations.

RP: I want to discuss your German report in a minute. First, can we define what we mean when we talk about digital preservation. Some people use the term long-term archiving (LTA), others talk of achieving "persistence over time", often without specifying a time period. What exactly do we mean when we use the term digital preservation?

NB: I would define digital preservation as a series of managed activities necessary to ensure continued access to digital materials for as long as necessary. It is really important to recognise that digital preservation is "a means to an end": the benefit and goal of digital preservation is access for as long as we require it — for some materials potentially in perpetuity.

The issues

RP: In other words, it's not just a matter of making sure digital information is stored securely but — given that formats change all the time and older ones constantly become redundant — there is a need for continuous curation. The nub of the problem is that digital media are uniquely vulnerable to "fading away" over time?

NB: Correct. And as the volume and complexity of digital information has grown, there has been growing realisation of the complexity of the activities needed to ensure long-term access to digital materials, and the extent to which this differs radically from preservation activities in the paper environment.

In the right conditions papyrus or paper can survive by accident, or through benign neglect, for centuries or, in the case of the Dead Sea Scrolls, for thousands of years. It takes hundreds of years for languages and handwriting to change to the point where only a few specialists can read them.

In contrast, digital information will not survive and remain accessible by accident: it requires on-going active management. The information and the ability to read it can be lost in a few years. Storage media such as punched paper tape, floppy disks, CD-ROM, DVD evolve and fall out of use. Digital storage media have relatively short archival life-spans compared to other media.

As the volumes, heterogeneity, and complexity of digital information grows the requirement for active management becomes more challenging and more critical to a wider range of organisations.

RP: You mentioned your German report on digital preservation: As you said, this was commissioned by a group of German research organisations. You were the lead author and editor of the report, which made 30 recommendations. And in the conclusion you suggested the Report be viewed as "the starting point from which to arrive at concrete ideas and activities related to a coordinated national hosting strategy." Digital preservation has been viewed as a serious problem for a good many years now. Eleven years ago, in a report for the Council on Library and Information Resources (CLIR), Jeff Rothenberg, a senior research scientist at the RAND Corporation, pointed out that unless the matter was addressed urgently "our increasingly digital heritage is in grave risk of being lost." There has subsequently been a great deal of talk about the topic, and a great many reports written, but I have the feeling that progress is proving very slow. When I spoke to German preservation specialist Eberhard Hilf recently he pointed out that he had authored a report on digital preservation in 2005. "This was written for nestor, a German-funded project of the government," he told me. "Our report gave all the arguments as to what should be done and by whom and how." And yet, he points out, little came of the report. Why is it taking so long to make progress?

NB: Actually I think we have made huge progress in the last 5-10 years! We have seen major new archives such as Portico, CLOCKSS, and the KB e-Depot established; national coalitions to promote digital preservation have emerged in the UK, USA, the Netherlands and Germany; and we have seen the completion of major research projects such as Planets.

There is still much to do of course but I recognise that a real mass of digital content needs to be created before there is widespread concern for many people over the long-term.

We have been in a period of moving from analogue to digital. The first priority in the early years has been major investments in retro-digitisation of existing print (JSTOR, Google Books, etc.) and creating/purchasing new "born digital" content for immediate access.

RP: We should not be surprised that it is proving a slow process then?

NB: Although the challenges of digital preservation were recognised early on, I think some institutions have needed to acquire or create a significant amount of digital content before feeling that they can or should concentrate more resources on digital preservation.

Since the threat is not solely technological, progress is a slower process: it can also involve social factors and organisational risks particularly over extended periods of time.

RP: I guess the true extent of the social issues has only become apparent over time. Certainly people seem more aware today that this is not just a technical problem. There are funding issues, there are political issues, there are policy issues and, as you say, there are organisational issues.

NB: Yes. We have learned that digital preservation is a complex, inter-related set of challenges: not just technical but often involving new business models and legislation, new or changed organisations and collaborations. These are time-consuming and difficult changes that have needed dialogue and "action research" to move them forward. Hence any major advances take a lot of groundwork and time to reach fruition.

RP: Would it be fair to say that the technical issues are more or less resolved, and it is now primarily a question of how you allow, encourage and/or make people preserve digital data effectively?

NB: Yes and no. Technologies and digital content continue to evolve rapidly so there is a continuous need to address the new problems and opportunities they create. New technical issues of preservation (and sometimes new solutions) are therefore always emerging.

RP: Another report on digital preservation was published in February by the Blue Ribbon Task Force. Called "Sustainable Economics for a Digital Planet", this — as the name implies — focused primarily on the economic issues of preservation. How well do you think it characterised the situation, and what particular contribution do you think it has made to the on-going debate about digital preservation?

NB: I think it has characterised the situation well. Its particular contribution has been to provide an economist's view of the challenges and potential solutions. I think though the report will need a more detailed action plan for carrying through the recommendations for its impact to be maximised.

Retro vs. born digital

RP: As you said, there are two different types of digital content: That created by digitising analogue material (retro-digitised to use your term) and born-digital content. What are the respective challenges and solutions when it comes to preserving these two different types of digital content?

NB: Our report concentrated on commercial scholarly e-journals and retro-digitised content for researchers and research support organisations in Germany, but most of the generic issues apply in other countries too.

For commercial scholarly e-journals the major challenge is that libraries no longer hold the content themselves but license access to it on publishers' servers. This radically changes the model for preservation and access compared to print, when libraries held copies and preserved them to ensure on-going access.

Libraries now need assurance that they can continue to have access to the digital content they have paid for well into the future regardless of any changes to the journal, publisher, or future subscriptions.

This means continuing (or "perpetual") access (and the digital preservation which underpins this) is dependent on negotiating the right clauses in licences with publishers and, even more critically, that both libraries and publishers have mechanisms in place which can ensure the terms of these clauses can be delivered well into the future. Immediate post-cancellation access is often available via the publisher's own server but access in the medium and longer-term is much more uncertain.

RP: So independent third-party organisations are key?

NB: This is far from fully resolved but efforts like Project Transfer have sought to address best practice when e-journals are transferred between publishers, and e-journal archive solutions such as

Portico, CLOCKSS, LOCKSS, and the KB e-Depot have emerged that offer differing degrees of preservation and access guarantee or insurance.

Most of the large publishers are now involved in these initiatives but there is a long-tail of small scholarly publishers, often with only one or just a couple of journals, who by and large are not.

Another challenge is that the publishing industry has become increasingly global rather than national. National institutions such as national libraries and national legal deposit legislation continue to play a vital role in preserving the cultural patrimony of different nations.

However it is much more difficult to achieve continuing/perpetual access or preservation for global publishing within that framework. New international partnerships involving national libraries such as the International Internet Preservation Consortium, or new organisations such as Portico, have emerged as a result.

RP: about retro-digitised material, which I guess is locally-held print material that libraries have digitised themselves?

NB: As you say, this content is largely held by the institutions themselves. For retro-digitised materials in Germany the issues are somewhat different to other countries however — since two major mass digitisation and national centres of expertise have been in place for 10 years and shared standards and procedures have evolved as a result.

RP: What are these two institutions?

NB: They are the Munich Digitisation Center (MDZ) at the Bavarian State Library, and the Centre for Retrospective Digitisation at Göttingen University Library.

There is very large-scale storage and IT expertise available to these centres through partnerships with major regional IT centres. The challenges here are mainly around how to support other institutions, particularly smaller institutions, and securing the sustainability and on-going access services for the content once it has been digitised.

RP: Are there are not equivalent initiatives/institutions in other countries?

NB: Not identical initiatives but de facto centres of expertise and equipment have been built up around national libraries and state and university libraries through major digitisation projects (for example the library partners in the Google books project).

RP: The Blue Ribbon report focused on four different areas: research data, web content, commercially owned cultural content and scholarly discourse. As you say, your main focus was on scholarly resources — primarily journals, although you do also discuss eBooks and databases. What specific challenges then does digital preservation raise for scholarly communication and the research community?

NB: I think some of the specific challenges are around the increasingly global nature of research and scholarly communication. There is a need to support this and align it with our frameworks for governance, funding, and interests which are largely national.

There is also the need to align the interests of the various stakeholders — who range from individual researchers, their employing institutions, the scholarly societies, libraries, research funders, and publishers.

Scholarly communication is also becoming more complex. Alongside journals articles there is growing demand across a wider range of disciplines for access to research data and for more access to the evidence base both for inter-disciplinary research and the public.

These all have preservation implications for scholarly content and for access to preserved material.

Initiatives

RP: What would you say were the main initiatives focused on the preservation of scholarly content today? And what is still missing?

NB: There are the national, state, and university libraries, as well as archives focused on national records or publications, and defined "special collections". I think we still need to do more on digital special collections but there is some exciting work emerging in personal digital collections funded in the US by the Mellon Foundation, and in the UK through projects such as Paradigm at Oxford or Digital Lives at the British Library.

Once you look beyond the national archives there are also big gaps in archive provision for digital records, and more work is needed to address regional and local records.

Then there are new international organisations like Portico, LOCKSS, and JSTOR (which we have already mentioned) and there is the Internet Archive, which is focused on preserving global resources in different ways. Their coverage is extensive but still a long way from complete.

We also still lack enough established tools and the means to sustain them. Creating tools or open-source alone is not enough: there need to be sustainable business models for them. I'm encouraged to see more commercial companies providing solutions as well as not-for-profits like the Open Planets Foundation, DuraSpace, and the LOCKSS Alliance.

Then there are national and international research data centres in some disciplines such as genetics, crystallography, astronomy and social sciences. However there are many gaps — some reflect the fact that data plays a less significant role, but others are genuinely missing and needed. I would include the "small sciences" in this.

RP: What is the issue for the small sciences?

NB: The most established and well-populated research data archives or repositories largely concentrate upon the requirements of "large science" disciplines, such as astronomy and particle physics. They have naturally arisen out of the systematic collection of the primary data by a few data centres and major research projects.

So far, however, there have been relatively few initiatives that focus upon enabling researchers in "small science" disciplines, or "small grant" projects, to share and preserve research data. Options for this may involve developing better repositories in universities, perhaps through federated partnerships between departments and the university library and/or more cross-institutional repositories organised on a subject basis.

Many researchers need better data management infrastructure and support for sharing data with collaborators globally than is currently available. This is very apparent for example in the survey of researchers undertaken as part of the UK Research Data Service (UKRDS) feasibility study.

I think one of the most interesting recent experimental developments in this area for the small sciences has been Dryad .

RP: Dryad is the open-access repository service you did some work for?

NB: Yes. It is a collaboration between a consortium of scholarly societies, journals and libraries to establish an open-access repository for supplementary data associated with published journal articles in ecology and evolutionary biology. It is a unique partnership, which may be an interesting model for other small sciences.

RP: And of course there is big science. As you said, in some disciplines it is necessary to create and preserve huge silos of raw data. Do you have any views on the specific challenges this raises? Presumably there are funding issues?

NB: That is right. At this scale, storage and performance are still major issues because the exponential increases in data are pushing the boundaries of what is technically possible within available budgets. This is forcing selection issues and a curation process for the data to the fore as it is not a viable strategy — or cheap — to keep everything.

RP: We are also seeing a huge growth in the use of Web 2.0 technologies for disseminating scholarly content (blogs, wikis, open notebooks etc.). This is not something that you were asked to look at in you report, but what particular challenges do you think these technologies raise so far as preservation is concerned? Or is there perhaps no need to preserve this kind of data?

NB: No they were outside the scope of our report, although the national hosting strategy is seen by the German partners as being incremental and capable of extension once the core materials covered in our report have been addressed.

I think there is a need to preserve this kind of data but the challenges here are around building relationships with new Web 2.0 companies (such as has recently happened between Twitter and the Library of Congress), developing new tools to acquire or preserve these new types of content, and last but by no means least how to select what has or will have scholarly value.

Copyright

RP: What about copyright? The British Library is very keen to archive the UK portion of the Web. It recently warned, however, that copyright law will prevent its Web archive from collecting a lot of material. The BL's CEO Dame Lynne Brindley pointed out that unless the Department for Culture, Media and Sport (DCMS) steps in, the British Library will be able to collect only one per cent of the 8.8 million .co.uk domain address that will exist by 2011. Brindley wants the DCMS to extend the provisions of legal deposit in order to resolve this issue. Copyright, it seems, is another significant issue when it comes to digital preservation?

NB: Yes. And I hope the British Library gets the regulations it needs to enact the provisions of legal deposit for Web archiving. The current legal framework is fine for print where at most there are a

few thousand publishers and the bulk of content is concentrated in the hands of the few largest. However it is wildly out of date and impractical for a digital environment with over a million personal creators and owners of content on the UK web.

This means that around a quarter to a third of the cost of web-archiving projects currently is devoted to clearing rights laboriously with individual owners — and the success rate in terms of replies is inevitably low. So there is an overwhelming public interest case for change.

If there are a relatively small number of commercial sites within this that need different treatment I hope that could be accommodated: it should be an opt-out system rather than an opt-in one in my view.

RP: Legal deposit aside, I wonder if there are other copyright issues here? You said earlier that since the bulk of scholarly content is in the hands of a few publishers "The current legal framework is fine for print". But is it that simple? The Blue Ribbon report argued that a change in copyright laws is required. Did you not reach a similar conclusion in your report? We have seen, for instance, that scholarly publishers are extremely protective of the current maximalist approach to IP regime and the constantly growing powers that copyright gives to rights holders. We have also seen publishers citing copyright in their efforts to prevent Open Access mandates. Does copyright really offer no threat to the digital preservation of scholarly papers?

NB: There is a separate working group of the Alliance of German Science Organisations looking at proposed changes to copyright law in Germany, so we did not specifically look at this in our report.

There are some important potential threats to digital preservation from new copyright regimes. In particular digital preservation often needs to make copies of a work in order to preserve it, or to transform versions of a work or software in order to address obsolescence.

Similarly, encryption used for digital rights management can be a major threat as it adds complexity and cost to preservation processes.

It is important therefore that the role and legitimate preservation actions of preservation agencies are recognised and protected in copyright or digital economy legislation.

It is also essential to note that in many states contract law can often over-ride copyright law, so safeguards are also needed to protect these rights in general use.

Open Access

RP: Open Access is the hot topic in scholarly communication today. You discuss Open Access in your report, particularly I think as it concerns content placed in institutional repositories. The Blue Ribbon report also looked at this issue, and seemed to be suggesting that Open Access mandates should always be accompanied by a preservation mandate, and funding to enable that preservation. As Paul Ayris, one of the authors of the Blue Ribbon report, put it to me, "It is irresponsible to create and make available digital assets and not to curate them for the long-term. The Report acknowledges this and makes recommendations based on this realisation." Do you agree?

NB: Yes and no. This is important but the Blue Ribbon report also mentions the role of collective licensing of e-journals and getting the right terms and conditions in them with publishers for continuing access and preservation.

That is the major focus of our report too as we were asked to focus on commercial e-journals. I think the Blue Ribbon report recommendation on open access mandates and preservation was addressing open-access journals and the community's responsibility to ensure they are preserved and available over the long-term.

As you say, we do touch on open access in our report, but it is important to recognise for our recommendations that even with 100% compliance with an open-access mandate in Germany perhaps less than 10% of the e-journal content licensed by German academic libraries would be available in German institutional repositories. So in some ways it is a parallel issue.

RP: Are you saying that if libraries can get the right licensing terms with publishers they do not need to worry about preservation?

NB: Not exactly. I am saying that different materials can have different preservation issues and approaches needed to address them. For commercially licensed e-journals I think that is both getting the licences right and ensuring there are organisations and mechanisms in place to ensure that preservation happens.

RP: Do you think it might be a mistake to focus so much on the journal rather than the article when looking at the preservation of scholarly papers? Some would argue that the idea of the journal (a collection of packaged and branded papers) will soon be anachronistic. In the future, they say, the individual article will be the primary unit of scholarly communication, not the journal. Indeed we can see this becoming a reality in terms of the impact factor — e.g. the Public Library of Science has introduced article-level metrics. Perhaps this sounds the death knell for the journal?

NB: Possibly but I think journals will remain a major part of the scholarly landscape for the foreseeable future. The reputation and prestige of specific journals are very important to authors and our systems of academic tenure and recognition. It would be premature to anticipate a major change any time soon.

Scholarly communication though is continuing to evolve and changes at article level are increasingly significant. For example, the way authors can choose between an open-access option and a traditional subscription option when publishing their article in a journal is now beginning to change the granularity of rights and rights metadata within a journal to the article level.

RP: It seems to me that our discussion raises a fundamental issue about who has responsibility for the preservation of scholarly journals. Should responsibility lie with publishers, or with the research community? For instance, there are those who question whether it is necessary to preserve the content placed in institutional repositories. After all, they argue, the purpose of OA mandates is to require authors to self-archive papers that they have published in regular scholarly journals, and they should therefore assume that the official version of these papers will be preserved by the publisher and/or by third party organisations like national libraries (e.g. the Royal Dutch Library), or services like PORTICO (which you mentioned). Why, they ask, introduce additional obstacles to getting self-archiving mandates passed, which is a difficult enough task as it is?

NB: I think the responsibility for the preservation of scholarly journals must lie with libraries but they need the active partnership and support of publishers to do this.

I agree with Paul Ayris that it is irresponsible to create and make available scholarly digital assets and not to curate them for the long-term. Universities and their researchers create a wide variety of content which can be deposited in their institutional repositories and that should be preserved. Preservation mandates within an institution help to deliver that.

RP: In making its proposals for preservation mandates The Blue Ribbon report appears to have had the Harvard model very much in mind. It also sees the Harvard approach to copyright as the model to clone. As the Report puts it, "Academic policy makers have begun to use defaults to promote preservation within their own institutions. Many universities have created digital repositories, such as the Digital Access to Scholarship at Harvard (DASH) repository, for works created by members of their faculty and staff. Traditionally, the contract between a faculty member and his or her university includes a copyright policy that gives all rights to the faculty author, including the right to decide whether or not to provide open access to the work. To encourage preservation, however, the Harvard University Faculty of Arts and Sciences in 2008 unanimously approved a motion to require each faculty member to give an electronic copy of every scholarly article he or she publishes to the Harvard Provost's Office. The provost then deposits the article in DASH, which is freely open on the Web. While a faculty member can request a waiver, the default process requires preservation." I'm not so sure that Harvard's OA policy places as much stress on preservation as this implies, but do you think that such a model is practical for many research institutions? While it might be possible for a wealthy and prestigious institution like Harvard to adopt such an approach, many research institutions will surely lack both the necessary money (or at least the willingness to spend it) to fund effective long-term preservation, or sufficient academic prestige to be able to force publishers to agree to publish papers under such conditions?

NB: I honestly don't think that preservation represents the major barrier to self-archiving mandates or researchers' compliance with them.

I think the Blue Ribbon Task Force had a number of preservation mandates in mind in the report, including things like legal deposit legislation; national record acts, or grant terms and conditions imposed by research funders so it covers a range of different materials. I don't think the Harvard mandate specifically mentions preservation: its motivation is access. The Blue Ribbon perspective is that a preservation obligation should be recognised by universities as part of this.

RP: You don't think then that given the difficulties in obtaining self-archiving mandates that insisting they should be accompanied by a preservation mandate could hold back the OA movement (and so perhaps of web-based scholarly communication)?

NB: I don't. Preservation is an essential part of scholarship and access and what major research universities do. Some university libraries have actually been preserving the outputs of their universities and other scholarly collections for centuries.

For others, recognising a need for preservation might be a shared obligation. In a digital environment there are also many new potential opportunities to develop shared preservation services between institutional repositories or between them and others such as national libraries.

RP: Do you think perhaps we are taking too scattergun an approach to preservation today? As you said, national libraries are involved; so too are library coalitions, along with various for-profit and

non-profit organisations. You have also suggested that individual university libraries have a role to play too, via their institutional repositories. Then there are the publishers, and doubtless a number of other government-funded organisations. Is there not a danger that we could find some content falling between the cracks? Hilf believes a more co-ordinated approach is needed, and that this requires national policies. If so, does it not imply an important role for governments?

NB: I think in practice the stakeholders need to organise from the ground up and then look for government and others to support their efforts. Strategies need to be drawn up and owned by the major participants, and you can see this beginning to happen in some countries with their national digital preservation coalitions.

There is a huge level of collaboration evident in digital preservation initiatives today. We have advocated a federated and collaborative approach in our report: it's a process which needs to build trust and gain leverage from mutually advantageous partnerships.

RP: Hilf argues that Australia is the only country in the world to have an effective digital preservation policy in place today. As he puts it, "The leading country is definitely Australia, particularly through the Australian National Library (ANB). The government has been the driving force since 1998, investing money, introducing legislation and actively realising a process of long-term archiving. For instance, all public Information in Australia anywhere must be LTA." Do you agree? Why Australia?

NB: No I don't agree, although I admire hugely the work of colleagues in Australia, who were early pioneers in the field of digital preservation. One factor may have been that they were also early innovators in digital culture and communication. That combined with a relatively small heritage of print may have encouraged that early interest.

However many other European countries (particularly the UK and the Netherlands) and the USA have been equally active now for a long period and I would give them equal recognition internationally.

RP: Recommendation #1 in your report is to "Maintain an international dimension to the Strategy, evaluate potential international partnerships and service providers, and maintain an oversight of emerging best practice and trends internationally." That clearly makes sense, and indeed — contrary to what I was just saying — it seems illogical to talk about "national" preservation initiatives in a global networked environment. But I guess the debate needs to be framed in national terms in to order to get funding, political buy-in, and a co-ordinated policy?

NB: Yes I agree. Much of research (particularly in the sciences, technology and medicine) is global and so is scholarly communication. However most of our funding, legislation, and other interests are national. We need to find ways to mediate and finesse between them.

RP: Of all the issues we have discussed I suspect that funding will prove the most intractable problem so far as digital preservation is concerned. Whose responsibility is it to fund digital preservation? Hilf argues that it is "exclusively a task for government-funded institutions". The Blue Ribbon report talks of the need for "developing public-private partnerships". What's your view? Where can we expect the money to come from, and how likely is it that sufficient funding will ever be made available, given the exponential growth of digital data and the current economic climate?

NB: I think over the short to medium term there will always be the need for public-private partnerships. This is because ownership is private for much of the content we are interested in and as copyright terms have been extended so owners' rights persist for a very long time.

Service and product providers such as Google, Twitter or Microsoft also have big roles and need to be part of the dialogue and solutions. Over longer time periods it almost certainly requires public funding (or a majority of public funding). This may be via government funded organisations or some of the international organisations and consortia funded from a variety of mainly public sources.

During the transition from print to electronic we have seen a long period of experimentation, with both print and digital products running in parallel. This is becoming increasingly unaffordable.

RP: So the funding issues are likely to spur an increasing migration to digital-only content?

NB: I think the current financial crisis is likely to accelerate the transition to e-only and the transfer of funding from print activities to digital access and preservation. It is unlikely that there will be new money for core services.

We are also moving from digital preservation being a "cottage industry" based on research projects towards services and tools delivering economies of scale and lower per unit costs. Understanding our costs and achieving greater efficiencies to make available funding go further will also be critical.

Finally we have never sought to preserve everything. Going forward as digital data grows exponentially, we are also going to have to select even more carefully what we choose to preserve.

Going forward

RP: Commenting on the Blue Ribbon report, Victoria McCargar, a consulting digital archivist, said: "The problem is that when it comes to digital preservation, it's often the same people talking to each other. You get that problem in any insular research space. But the larger taxpaying public that you need to build the infrastructure has problems understanding the problem. Preservation is expensive and difficult in an era where people are drowning in data and seemingly limitless access. It's very counter intuitive. It's a tough sell to the public." Do you agree? Could this partly explain why progress is proving so slow — the echo chamber effect of many specialists talking to one another, and the consequent failure to convince taxpayers and/or politicians of the need to provide sufficient money for preservation?

NB: I agree with Vicky that the field is quite small and it would benefit from greater research input, for example preservation hardly features in the work of many information or library schools in universities.

However I think that digital preservation and long-term access are part of the wider public agenda and on the political radar, although retaining and increasing that attention needs constant advocacy and effort.

The work of the national coalitions and partnerships such as the Digital Preservation Coalition in the UK has been very successful in raising public perception of the issues.

RP: Is there a danger that we could see a "digital black hole" (as Lynne Brindley calls it), either with web-based content, or scholarly content, or both?

NB: Yes as we have discussed, there are real risks going forward. I think this is going to be even more acute as public funding across many developed countries is squeezed hard.

However, this crisis is also an opportunity and I hope we can make some decisive changes and build on some of the major successes achieved so far.

RP: Finally then, what should the priorities be going forward in your view?

NB: In my view this is what needs to be done:

1. Accelerate the move to e-only and increase the balance of funding devoted to access and preservation of digital content;

2. Collaborate and automate to achieve the lowest possible costs and greatest economies of scale for digital preservation;

3. Look at variants of cloud computing for digital preservation storage and services – establishing more secure, persistent, and defined clouds and services than available currently;

4. Give more attention to preservation research data in the small sciences and to supplementary data underpinning scholarly articles;

5. Develop and encourage personal archiving, private digital collection and philanthropy;

6. Get more digital preservation and digital curation into university curricula and professional training programmes and support career paths for appropriate specialists.

RP: Thank you for your time.

I have in mind a figure of $8, but whatever anyone felt inspired to contribute would be fine. This can be done quite simply by sending a payment to my PayPal account quoting the email address richard.poynder@btinternet.com. It is not necessary to have a PayPal account to make a payment.