

Climbing the Scholarly Publishing Mountain with Project SHERPA

John MacColl
Sub-Librarian (Online Services) & SELLIC Director
University of Edinburgh
Darwin Library
The King's Buildings
Mayfield Road
Edinburgh EH9 3JU
john.maccoll@ed.ac.uk

- JISC launched its **FAIR programme** (Focus on Access to Institutional Resources) in January of this year. The central objective of the Programme is to test ways of releasing institutionally-produced content onto the web.
- FAIR describes its scope as ‘to support access to and sharing of institutional content within Higher Education (HE) and Further Education (FE) and to allow intelligence to be gathered about the technical, organisational and cultural challenges of these processes ... This programme is part of a broader area of development to build an Information Environment for the UK's Distributed National Electronic Resource.’
- It specifically sought projects in the following areas:
 - Support for disclosure of institutional assets including institutional e-print archives and other types of collections through the use of the OAI protocol.
 - Support for the harvesting of the metadata disclosed through this protocol into services which can be provided to the community on a national basis. These services may be based around subject areas or other groupings of relevance for learning and research.
 - Support for disclosure of institutional assets through the use of other relevant protocols, for example Z39.50 and RSS
 - Exploration of the deposit of institutional collections with a community archive or

to augment existing collections which have proven learning, teaching or research value

- Experiments with the embedding of JISC collections and services in local institutional portals and how well they can be presented in conjunction with institutionally managed assets.
- Studies into the related issues and challenges of institutional asset disclosure and deposit, including, collections management, IPR, technical, organisational, educational, cultural and digital preservation challenges.

- FAIR awarded funding to 14 projects in five clusters:

- Museums and Images Cluster (4 projects)

Petrie Museum, University College London - *Accessing the Virtual Museum*

Fitzwilliam Museum, University of Cambridge/Archaeology Data Service, University of York - *Harvesting the Fitzwilliam*

AHDS Executive, King's College London/Theatre Museum, V&A/Courtald Institute of Art, University of London/Visual Arts Data Service, University of Surrey/Performing Arts Data Service, University of Glasgow - *Partial Deposit*

ILRT, University of Bristol/University of Cambridge - *BioBank*

- E-Prints Cluster (4 projects)

CURL (University of Nottingham/ University of Edinburgh/University of Glasgow/Universities of Leeds, Sheffield and York ('White Rose' partnership)/University of Oxford/British Library) - *SHERPA (Securing a Hybrid Environment for Research Preservation and Access)*

RDN, King's College London/ University of Southampton/UKOLN, University of Bath/UMIST/University of Bath/University of Strathclyde/University of Leeds/ILRT, University of Bristol/Heriot Watt University/University of Birmingham/Manchester Metropolitan University/University of Oxford/University of Nottingham/OCLC - *e-prints UK*

University of Strathclyde/University of St. Andrews/Napier University/Glasgow Colleges Group - *Harvesting Institutional Resources in Scotland Testbed*

University of Southampton - *Targeting Academic Research for Deposit and dISclosure*

- E-Theses Cluster (3 projects)

Robert Gordon University/University of Aberdeen/Cranfield
University/University of London/British Library - ***Electronic Theses***

University of Edinburgh - ***Theses Alive!***

University of Glasgow - ***DAEDALUS***

- Intellectual Property Rights Cluster (1 project)

Loughborough University/Birkbeck College, University of London/University
of Greenwich/University of Southampton - ***Machine-readable rights metadata***

- Institutional Portals Cluster (2 projects)

University of Hull/RDN, King's College London/UKOLN, University of Bath
- ***Presenting natiOnal Resources To Audiences Locally***

Norton Radstock College, Bristol/City of Bath College/City of Bristol
College/Filton College, Bristol/Weston College, Weston-super-Mare/Western
College Consortium, Bristol - ***FAIR Enough***

- **The Open Archives Initiative** lay very firmly behind FAIR, as the call document says: 'This programme is inspired by the vision of the Open Archives Initiative (OAI) (<http://www.openarchives.org>), that digital resources can be shared between organisations based on a simple mechanism allowing metadata about those resources to be harvested into services. ... The model can clearly be extended to include other kinds of objects, for example learning objects, images, video clips, finding aids, etc. The vision here is of a complex web of resources built by groups with a long term stake in the future of those resources, but made available through service providers to the whole community of learning.'
- The libraries of the major research libraries of the UK (known collectively as **CURL**, the Consortium of University Research Libraries) decided to bid within the FAIR Programme. CURL's Task Force on Scholarly Communication was remitted with the preparation of the bid. Coordination was by Paul Ayris, Director of Library Services at UCL, and the bid was authored by Stephen Pinfield, Assistant Director of Information Services at the University of Nottingham, and me.

- Why was CURL interested in this?
- CURL libraries, which spend relatively more on journals than other academic libraries, have suffered most from the price rises in journals which have seen the average costs rise by 10% for the past 15 years. Exasperated by this journal stranglehold on their budgets, libraries have become very acutely aware of the irony of a situation in which they are effectively ‘buying back’, in a collective sense, the research collectively produced – without any intention of profit-making - by the research universities they serve. What is in reality an economically absurd way to present publishers with profits becomes, because of this price crisis, an urgent problem which demands a solution.
- The nub of the issue is copyright. If academic authors, or universities themselves, can hold on to copyright in this research, then they can post it on their own web sites.
- Until OAI-PMH came along, this was not a very attractive prospect. It required authors to go individually to the web sites of research groups in academic departments in potentially hundreds of universities worldwide, and therefore there was no facility for comprehensive access.
- Of course, the ‘invisible college’ has always operated in this selective way in any case. The fact is that researchers do make free copies of their research available to their peers – via conferences, and on web sites. An interesting variant on this is the culture of working papers produced by academic staff belonging to particular institutions.
- But this is exclusive. Senior researchers in any discipline will know which institutions across the world have the strongest departments, or those with research interests which match their own – but what about junior researchers, or researchers in interdisciplinary areas? There is no question that this selective access reduces the

impact of the research which is produced, and is not in the best interest either of the authors of the research, or of scholarship generally.

- The key to changing this situation so that exclusive research becomes comprehensively available lay, of course, in metadata, and therefore the arrival of the Open Archives Initiative Protocol for Metadata Harvesting has the potential to make an enormous difference, creating an effective ‘union catalogue’ from departmental collections.
- How will this help CURL? In two ways.
- First, CURL is interested in developing the culture of research publication and communication generally. Academic libraries are much more than mere custodians of literature: they see themselves as infrastructure providers, aiming to make the work of their researcher colleagues more productive and more effective. Librarians used to talk about the desire for ‘universal bibliographic control’. In fact, control-freakishness has often hampered library efforts to engage with the researchers in their institutions – but the aim is still central to what our profession is for: ‘universal knowledge resource management’ may be a better – if blander – way of putting it.
- Second, inasmuch as these open archives can replicate content only otherwise available commercially, so they have the potential to save libraries money. If all the articles our researchers wish to consult are available via a free corpus based upon a myriad of institutional open archives all complying with OAI-PMH and therefore inter-searchable, then we no longer need to hand over large sums of money to journal publishers.
- But of course, it is likely to take a long time before the efforts of researchers worldwide can be sufficiently concerted to allow journal subscriptions to be dropped. This is truly a massive mountain to be climbed. In some disciplines the potential is

more apparent than in others. The case of the high-energy physicists who have been using **arXiv.org** and its predecessors for more than a decade is well-known, but few other disciplines have yet shown an interest in organising themselves around a disciplinary server in this way.

- One suggested means of redressing this is to put the emphasis on servers at the institutional level instead of the disciplinary model, and that is what Project SHERPA will seek to test in the UK. If the impetus comes from within the university, with an institutional framework in place, coordinated by the Library or another central service, to permit the growth of an institutional database of research articles, then the current unevenness in disciplinary spread in the free corpus may be reduced. Over time, the argument goes, a snowball effect will operate within institutions, and at a national – and international – level, assuming effective advocacy, so that a wide-ranging free corpus can be built, eventually rivalling the commercial corpus. It will be a steep and difficult climb! The technology is relatively straightforward, and the legal barriers may be more imaginary than real. The hardest challenge of all will be that of convincing academics – in sufficient numbers - that they must join the expedition.
- Project SHERPA will aim to encourage the institutional archive agenda. **Six open archives will be funded from the project, at the Universities of Edinburgh, Glasgow, Oxford and Nottingham, together with a shared archive somewhere within the ‘White Rose’ partnership of York, Leeds and Sheffield, and one at the British Library** for the research outputs of ‘non-aligned’ researchers.
- It will use the open source eprints.org software produced by the University of Southampton.
- The leadership of the project will be provided by the University of Nottingham, and a serious advocacy campaign will be mounted across the UK to encourage the population of these archives with content. The project will not confine itself to these

six archives, but will encourage the development of others, either on an individual institutional or an institutional cluster basis.

- But SHERPA has another objective. The CURL Directors, in considering the potential of the ‘Open Archives Initiative’, were very interested in the archiving dimension. They wanted a project which would ‘put the archiving into Open Archives’.
- The reason for this is that, as we move into an electronic journal-dominated future for research, there are real concerns emerging about the preservation of digital material. Who should take responsibility for the preservation of the academic record? This has traditionally been a research library activity.
- The emphasis throughout SHERPA as a whole will be on refereed content. We wish to demonstrate the value of the open archive approach in generating an alternative free location for research articles authored by academics. The idea that servers on the web are filled with poor quality content and not up to the task of publishing research is of course one which commercial research journal publishers are keen to peddle themselves, and a lot of people are suspicious of the use of free web sites for research for just this reason.
- While SHERPA therefore will not require only refereed articles, it will seek these as its first priority, since a good proportion of refereed articles searchable within the SHERPA corpus will quickly help to demonstrate the viability of the approach to ‘respectable’ research publishing.
- Another reason to focus on refereed material is that it is likely that this will define what proportion of the SHERPA content is selected for digital preservation. While a preprint which an author never intends to submit for peer review would still be worth preserving, generally the approach will be to preserve articles once they have finished

changing – and this is most easily witnessed by their appearance in the journal literature.

- For these reasons, SHERPA is keen to engage publisher support for the project. The very choice of the name, indeed, is designed to convey this. ‘Securing a Hybrid Environment for Research Preservation and Access’. This particular ‘hybrid environment’ is one in which a free corpus of research literature subsists with a commercial one, and is not necessarily – or wholly -in conflict with it. Our expectation would be that as the free corpus grows, publishers will seek to diversify their markets in order to retain their income sources.
- This idea fits with the aims of the **Budapest Open Access Initiative**¹, funded by the Soros Foundation’s Open Society Institute. The BOAI aims to promote the growth of a free corpus particularly for the benefit of researchers in developing countries, by twinning self-archiving with ‘open journals’.
- It defines ‘open journals’ as ‘... journals committed to open access ... these new journals will no longer invoke copyright to restrict access to and use of the material they publish. Instead they will use copyright and other tools to ensure permanent open access to all the articles they publish. Because price is a barrier to access, these new journals will not charge subscription or access fees, and will turn to other methods for covering their expenses. There are many alternative sources of funds for this purpose, including the foundations and governments that fund research, the universities and laboratories that employ researchers, endowments set up by discipline or institution, friends of the cause of open access, profits from the sale of add-ons to the basic texts, funds freed up by the demise or cancellation of journals charging traditional subscription or access fees, or even contributions from the researchers themselves.’

¹ www.soros.org/openaccess/

- Of course, it is possible to be cynical about the likelihood of ‘alternative sources of funds’. Why would Elsevier stop charging for its journals, and begin to seek to cover its costs another way? The threat of widespread journal cancellations might be one reason. However, monopoly publishers don’t become dominant by missing tricks. The marketing of Elsevier’s aggregated offering, *ScienceDirect*, seems to be setting out to ensure that libraries will be paying the same amount for the electronic-only content as they were for the print-only content, by the time they are ready to dispense with the latter. This is an interesting development. Does it imply a recognition by Elsevier that they will not be able to wield copyright as a weapon against the academic community for ever?
- One of the most difficult problems we face in advocacy is the grip *ScienceDirect* and other ejournal bundles already have on our academics. Asking them to give us content is one thing. Asking them to boycott *ScienceDirect* to help us grow the free corpus more rapidly would be quite another. The fact is that libraries have shielded academic departments from a true understanding of the cost of research publication. John Houghton, the Australian economist, recently likened this relationship to the purchase – on an increasingly limited budget - of luxury dog-food: those who purchase it are not those who consume it. If the dogs had to pay for the food themselves, and were given limited budgets to do so, that might finally change behaviour. But of course that raises all sorts of worrying questions about the future role of libraries.
- The market advantage provided by the bundling of publisher content via databases like *ScienceDirect* lends urgency to the need to create an alternative, but free, database, of as much of the same content as can possibly be provided by it.
- Another reason why publishers would give authors this permission, is because they had no legal right not to. There are already legal ways round publisher copyright – albeit somewhat clumsy. Further legal challenges may result if the academic author

community becomes sufficiently engaged in the effort of projects like SHERPA.

- Another reason still may be that, faced with the prospect of losing its editors to rival publications – perhaps low-cost ‘open journal’ or cheap journal start-ups such as those supported by the **Scholarly Publishing and Academic Resources Coalition** (SPARC)² – publishers will decide their only course is to comply.
- SHERPA wishes to set the ball rolling properly within the UK in examining the creation and consequences of a free corpus. We do want to engage publishers in testing this new model of scholarly publishing, and hope we might persuade some publishers to grant the project a blanket waiver to permit copyright retention in refereed articles for the duration of the project.
- This may not be as difficult as all that. The editor-in-chief of an Elsevier journal in informatics, one of the professors of informatics at the University of Edinburgh, recently pointed out to me that he had pursued Elsevier over its policy regarding eprints. He received a reply in the *Bulletin of the European Association for Theoretical Computer Science* for October 2001, in an article entitled ‘Recent Elsevier Science Publishing Policies’, which stated ‘... the exclusive distribution rights obtained by Elsevier Science refer to the article *as published*, bearing our logo and having exactly the same appearance as it has in the journal. Authors retain the right to keep preprints of their articles on their homepages (and/or relevant preprint servers) and to update their content, for example to take account of errors discovered during the editorial process, provided these do not mimic the appearance of the published version. They are encouraged to include a link to Elsevier Science’s online version of the paper to give readers easy access to the definitive version.’
- Another objective of Project SHERPA will be to encourage its partner sites to deploy

² www.arl.org/sparc/

open linking technology, based on the OpenURL protocol, in order to allow the partner sites to control the links provided within their own information environments, in order to maximise the use of the SHERPA archives. Thus, for sites which already have OpenURL resolvers in place, whether through a commercial tool like **Ex Libris SFX**, Endeavor LinkFinderPlus (also now available through *ScienceDirect*) or whether through an open source resolver, these libraries should be able to control the destination of the links so that they can resolve only to the SHERPA archives, or to either, depending on the wishes of the user. In this way, the provision of an alternative free corpus can be made very real. This would also obviate the need for the link to *ScienceDirect* which Elsevier ‘encourage’ as described in the policy I mentioned a moment ago.

- Clearly, the number of SHERPA papers duplicating papers also available through *ScienceDirect* or via some other publisher will be a very tiny proportion to begin with. For open linking to work, it will also be necessary for the OpenURL compliant software, eprints 2.0, to support the protocol. We hope to explore these possibilities in SHERPA – though have no direct funding with which to do this, so we will be looking for some volunteer effort from within the partner group.
- Another area for volunteers is in increasing the size of the Project. The JISC funding ceiling for FAIR found us having to work frantically on the bid to keep the project within JISC budget (we have the maximum funding - £300k, but we need this over three years). But Paul Ayris, in inviting CURL members to join the bid, found it encouragingly over-subscribed within CURL, and we are keen to admit other CURL libraries to join us in setting up their own institutional servers. They will only have to fund the cost of their own servers: technical support and advocacy materials will be provided from the project team.
- We are aiming at a minimum of 4,500 deposited eprints by the end of the three years.

- Let's now return, finally, to the other – critically important – element of SHERPA, digital archiving. Peter Hirtle, writing in *D-Lib* in April 2001, said 'an OAI system that complied with the OAIS reference model, and which offered assurances of long-term accessibility, reliability, and integrity, would be a real benefit to scholarship.'³
- CURL had a strong interest in implementing an OAIS-based digital preservation project, having funded the successful work in OAIS model development undertaken by the CEDARS project over the past several years. Unfortunately, the JISC FAIR programme does not have digital preservation within its scope. We as yet therefore do not know whether the digital preservation workpackage will be funded: there are possibilities under a forthcoming digital preservation call from JISC, which will emerge in July. Short of that, there may be other possibilities involving existing JISC infrastructure providers, but these have still to be clarified.
- To us, the funding of the digital preservation element is essential to the Project as we conceived it. The 'hybrid environment for research preservation and access' depends upon a mean of preservation, and our original bid had included a partner to provide a 'dark vault', based on OAIS, into which preservation copies of all refereed articles would have been deposited. We are therefore still in discussion with JISC over this question, and will be unlikely to be able to start it properly until the digital preservation question had been resolved.
- **Thank You!**

June 2002.

© John MacColl. Non-exclusive right of publication granted.

³ Peter Hirtle, 'Editorial: OAI and OAIS: What's in a Name?' *D-Lib Magazine* 7, 4, April 2001
<<http://www.dlib.org/dlib/april01/04editorial.html>>