Learned Publishing (2005)**18**, 25–40

Developing a model for e-prints and open access journal content in UK further and higher education

Alma Swan
Key Perspectives Ltd

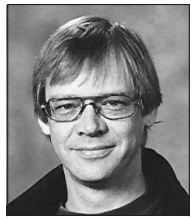
Paul Needham
Cranfield University

Steve Probets, Adrienne Muir, Charles Oppenheim, Ann O'Brien, Rachel Hardy and Fytton Rowland
Loughborough University

Sheridan Brown
Key Perspectives Ltd

ABSTRACT: A study carried out for the UK Joint Information Systems Committee examined models for the provision of access to material in institutional and subject-based archives and in open access journals. Their relative merits were considered, addressing not only technical concerns but also how e-print provision (by authors) can be achieved – an essential factor for an effective e-print delivery service (for users). A 'harvesting' model is recommended, where the metadata of articles deposited in distributed archives are harvested, stored and enhanced by a national service. This model has major advantages over the alternatives of a national centralized service or a completely decentralized one. Options for the implementation of a service based on the harvesting model are presented.

Alma Swan



Paul Needham

Introduction

In this article we describe a delivery, management and access model for e-prints and open access journal content for UK further and higher education commissioned by the Joint Information Systems Committee (JISC). The target content is (i) e-prints – digital copies of academic research articles published in subscription-based journals that are made available online to permit increased access; and (ii) articles published in open access journals. The proposed service would provide immediate and maximal access to scholarly research, supplementing the more limited access provided by subscription-based journals, in turn maximizing the impact of research.

Other benefits accrue from such a system too. It would enable the generation of standardized online CVs for each institution's researchers and these could be used for evaluation purposes – internally within the institution or for external purposes such as the UK's national Research Assessment Exercise. A nationally organized service in the UK for the delivery of e-prints and open access journal content to the scholarly community would therefore be an important development.

There are two ways for researchers to provide open access for their work – by publishing their articles in open access journals (or in hybrid journals that will provide open access to individual articles for a publication fee) or by depositing ('self-archiving') copies ('e-prints') of their subscription-journal articles in open archives (known variously, depending on circumstances, as e-print archives, institutional archives or institutional repositories). Although often used interchangeably, we use the term 'institutional archive' here in preference to 'institutional repository'. This is in part because the term 'archive' is used

in many official names (e.g. Institutional Archives Registry, Open Archives Initiative) and in part because it reflects an activity (authors 'self-archive' their work – they do not 'self-reposit'). Most importantly, though, the use of the term repository is now generally coming to denote something more than an e-print archive; rather, an institutional collection of material that contains far more than e-prints, such as grey literature, institutional-specific digital collections and so on. Since the remit of our study was to develop a model for the delivery and management of e-print and open access journal content only, the term *institutional archive* is the most accurate and appropriate.

repository is now generally coming to denote something more than an e-print archive

The JISC commissioned the study as part of an overall programme on open access in the UK and beyond. The brief was to forecast a delivery, access and management model for e-prints and open access journal content within the UK further and higher education communities. The study took place during the same period as the investigation into scholarly scientific publication by the UK House of Commons Select Committee on Science & Technology[1] and the two reported almost simultaneously. They were followed very shortly thereafter by the recommendations of the National Institutes of Health in the USA.[2]

An article that summarizes all the findings of our study, including preservation issues, legal issues and some aspects of the costs involved in setting up and running e-print archives, is in press[3] and the full report has been published by the JISC.[4] This present article specifically focuses on the model we devised and the reasons why we chose this one above the other possible options.

The open access material to be delivered

For the purposes of the study (and for this article) an e-print was defined by the JISC as

> . . . a digital duplicate of an academic research paper that is made available online as a way of improving access to the paper. E-prints are divided into *preprints* (papers that are circulated before they have been formally approved for publication), and *postprints* (papers that have been approved for publication).

It was specified that the model should encompass both e-prints and the content of open access journals (where this can be harvested). Harvestable articles – both e-prints and open access journal articles – are those that are OAI-compliant, that is, their metadata (bibliographic records) are exposed in the form laid down in the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH).[5] So long as their exposed metadata are OAI-compliant, they can be harvested by OAI *service providers* whose databases are then searchable by users who are pointed at articles of interest, wherever those articles reside. There are a number of OAI service providers in existence. Perhaps the best known examples are OAIster[6] (University of Michigan) and Citebase[7] (University of Southampton). Users type in a search term (author name, keyword, etc.); the software searches metadata already harvested from all available open access OAI-compliant archives, and returns a list of appropriate articles with links to their full text. Users then access the full text at its original location.

With respect to *open access journals*, the current situation (Oct 2004) is that the Directory of Open Access Journals (DOAJ[8]) lists 1,277 titles, of which 324 are harvestable at the article level. Most, but not all, of these titles are published by BioMed Central[9] or the SciElo[10] project.

Existing *e-print* material resides in open archives which can take two forms – centralized, subject-based archives, or distributed archives located at research-based institutions around the world. Two well-known and long-established examples of subject-based archives are arXiv,[11] set up in 1991 and covering physics, mathematics and related disciplines; and Cogprints[12], set up in 1997 and covering cognitive sciences (psychology, neuroscience, linguistics and related areas).

Some figures may help to illustrate activity in these two subject-based archives. arXiv currently houses some 300,000 digital items and is accessed approximately 1.5 million times per month by users. Cogprints currently contains around 2,000 items. Almost

half of these (985) are postprints (journal articles that are either published or in press); there are also 377 preprints, 283 conference papers, 44 conference posters and 234 book chapters.

As well as subject-based e-print archives, distributed broad-based archives have been set up in universities and research institutions around the world. There are now several hundred of these. OAIster is currently harvesting from 351 archives (institutional, subject-based and open access journal archives) and has over 3.5 million records in its database. Not all of these records will be e-prints, however, as OAIster harvests other types of digital object as well (see below) and from archives that do not contain e-prints.

The Institutional Archives Registry[13] currently lists 227 archives housing e-prints (along with other types of digital object in many cases); 33 of these are in the UK. Of the total, 121 are archives that are based at an individual institution and contain research output material (i.e. e-prints) and their content generally reflects the broad scope of scholarly activity at those institutions; that is, it covers many subject areas. Some are sub-institutional or departmental archives, however, which usually cover a single subject or discipline. Of the total number, 32 are archives with e-print content but which are cross-institutional and in general these tend to be subject-based, though not exclusively, since there are some broad-scope archives formed by collaborating institutions. An example of this type is the White Rose Consortium e-Prints Repository, a collaborative project by the universities of York, Sheffield and Leeds.

In May 2004 there were just short of 25,000 articles in the 20 e-print archives harvested by the RDN/e-Prints UK[14] project. One had no articles at all while at the other extreme the open access publisher BioMed Central's archive contained over 12,000 from the *circa* 120 journals it publishes. By far the best-populated university-based archive is the University of Southampton's ECS EPrints service with 8143 articles. Two other Southampton-based archives also had reasonably high numbers of articles: e-Prints Soton had 758, and Psycprints (a journal archive) had 720.

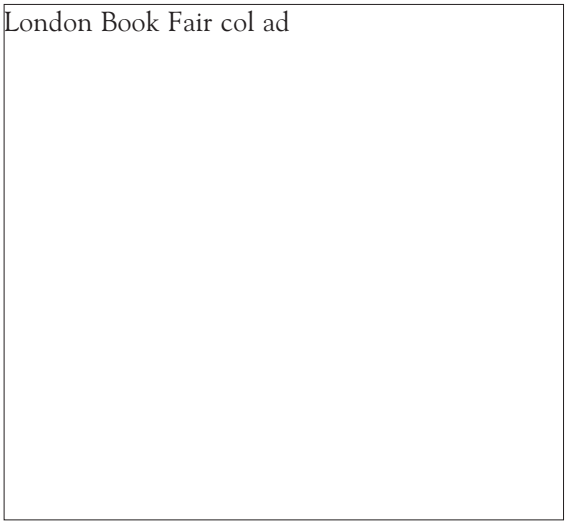
The types of digital object collected and stored in archives

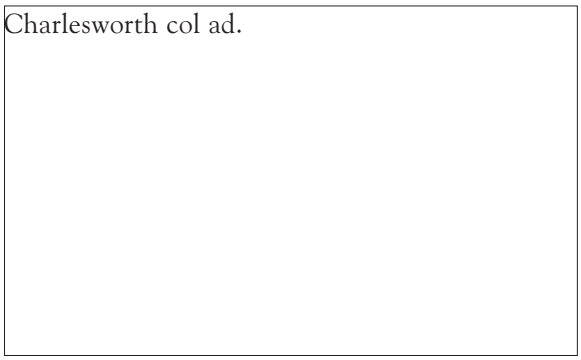
While some archives concentrate only on e-prints, others may house a considerably varied selection of digital objects, some of which may be very specific to local requirements. For our study, the JISC had specified that it required models for the access and delivery of just two types of digital object – e-prints (preprints and postprints) and open access journal articles. These are shown in bold type in the list below. Nevertheless, we kept in mind that other types of object are archived and that in time the inclusion of these other types of object might be deemed desirable. Some examples of the types of digital item that might be stored in archives set up by UK educational institutions are:

- **Preprints**
- **Postprints**
- All drafts and working papers plus corrigenda (i.e. a trail from first draft to the postprint: this is sometimes referred to as the 'low threshold' model)
- Ancillary data from research, e.g. video, audio, large datasets. Some archives accommodate these types of data, which cannot be published in a traditional peer-reviewed print journal, because there is merit in them being made available to other researchers, and in being preserved digitally in a formal way
- Books and monographs
- Non-published digital objects:
  - Teaching materials
  - Collections (music, images, etc.)
  - Research output from specialised subject fields, such as performing arts, where output is usually in the form of performance, video or audio
  - Dissertations and theses
  - Multimedia items
  - Local institutional 'events', e.g. performances, lectures, exhibitions
- **Open access journal articles**

Although the scope of the study did not include 'grey literature', we were certain that it should not be ignored, not least because

broad-based archives have been set up in universities and research institutions around the world


London Book Fair col ad


Charlesworth col ad.

preprints submitted to journals and then subsequently rejected may end up forming a permanent piece of grey literature (in fact, the DAEDALUS project formally groups preprints with grey literature[15]). Conversely, many e-print archives contain much content in the form of reports (conventionally classed as grey literature) despite many definitions limiting e-prints to preprints and postprints.

Although grey literature was outside the scope of the study, we addressed it for the reasons above. Readers of this article may find a useful introduction to the grey literature landscape in the UK in the MAGiC (Managing Access to Grey Literature Collections) Final Report.[16] In short, the MAGiC project, which was sponsored by the British Library and the Research Support Libraries Programme, was proposed to deal with the paucity of grey literature cataloguing and proposed the creation of a national grey literature service built around the OAI harvesting model, which would incorporate both electronic and hardcopy (legacy) documents.

the creation of a national grey literature service

The form and format of digital objects collected and stored

For the purpose of our study an e-print was defined as an article published in the scholarly literature, given away free by its author. In other words the objects archived are going to be, in the main, standard journal articles, perhaps accompanied in the archive by additional supporting material such as the large datasets generated in some branches of the sciences, or video or audio clips.

In some disciplines, however, research output may frequently take other forms. For instance, in the performing arts output is often in the form of a performance. In this context this presents additional issues for the archiving of research results: video records of performance use large amounts of digital storage space, for example. In the arts and humanities, too, although scholars do publish work in traditional journals, there is also a large volume of output in the form of monographs. These may be archived in the same way as journal articles, but there may be certain differences. Monographs may have multiple authors, each contributing a chapter and possibly from different institutions, perhaps requiring separate deposition and submission policies. In general, too, monographs tend to be much larger documents than journal articles, so there is again a space implication. Finally, it is fairly common for monograph authors to be paid royalties by the publisher and though these are usually small, they nonetheless represent payment, so in these cases this is not 'giveaway' literature.

The range of data formats permitted by existing e-print archives varies from one archive to another. Many e-print archives, for example the University of Oxford's Oxford E-Prints, upload text-based documents only in PDF format, whereas another example, the University of Glasgow's eprints@Glasgow, accepts (and stores) digital objects in a much wider range of formats.

The global picture

There is something of a global race taking place with respect to achieving open access via national policies on self-archiving. Initiatives are in place in India, Norway, The Netherlands, Germany, Canada, Scotland and France, among others, of which Australia is an exemplar: the government gave funds of AU$12 m. in Oct 2003 to make 'Australia's research information . . . more easily accessible and better managed'.[17] The country's major research universities all have institutional archives, and the Department for Education, Science and Training (DEST) has decided that a national linked-up approach is the best way forward. It now supports four projects covering 15 Australian universities, Australian and international libraries, representatives from industry and various international organizations. The Australian Partnership for Sustainable Repositories (APSR) has been set up and is working through the Australian National University's Centre for Sustainable Digital Collections to develop a national research infrastructure through broad, archive-based architecture which will ensure access continuity and the sustainability of digital collections, and facilitate national co-ordination and international linkages.

The e-print situation in the UK

There are some projects and programmes already in operation related to national delivery of e-prints in the UK. The situation is promising but lacks co-ordination, consisting of a series of linked pilot projects and a number of already established institutional e-print archives and, as yet, no involvement in e-print archiving by the British Library. There are three particularly significant national initiatives operating.

ePrints UK

The ePrints UK[18] project is developing a series of national, discipline-focused services for e-prints from compliant open archive repositories, particularly those provided by UK universities and colleges. The interface will be provided through OCLC and will use 'name authority' and 'citation analysis' Web services (offered by OCLC and the University of Southampton, respectively) to enhance the metadata harvested from available archives. With respect to the study we are reporting here, it is significant that ePrints UK already also harvests metadata from a number of e-journal repositories, demonstrating that integration of metadata from journal articles and e-prints is a practical and achievable proposition.

SHERPA

The SHERPA[19] project is initiating the development of openly accessible institutional digital collections of research output in a number of universities. Among the issues the project is investigating are intellectual property rights, quality control, other key management issues associated with making the research literature freely available to the research community and technical aspects of such a system, including interoperability between repositories and the digital preservation of e-prints.

FAIR

The Focus on Access to Institutional Resources[20] programme, funded by the JISC, has a mission 'to evaluate and explore different mechanisms for the disclosure and sharing of content (and the related challenges) to fulfil the vision of a web of resources built by groups with a long term stake in the future of those resources, but made available to the whole community of learning.' The JISC Information Environment is a virtual location where authors can deposit and share useful content (e.g. research outputs) and it is envisaged that this will join the current collection of JISC-funded content, which has the potential to include externally generated content from publishers and aggregators as well.

Other issues pertaining to a model for a new service in the UK

There were other issues to investigate and consider as we deliberated on the development of a suitable model. These fell into two categories – technical and 'cultural' – and are summarized below.

Technical issues

The technical aspects we considered were:

Software

Should a new system run on one of the available software packages or should a new, bespoke package be developed? There are several open-source (free) software packages available for running open archives. The best known are DSpace, developed by MIT, and EPrints, developed at Southampton University. Both Eprints and DSpace offer interoperability via the OAI-PMH. DSpace uses persistent identifiers that, unlike ordinary URLs, do not change when the physical location of the digital item alters. Other OAI-compliant software systems of note are CDSware, developed by CERN; and Fedora, developed jointly by the University of Virginia and Cornell University, with funding from the Andrew W. Mellon Foundation

Preservation policies

Among the various issues that we identified here as having implications for a national e-prints service were: what should happen if an author wishes to withdraw an article; how to handle and track repeated revisions of an article after it is first deposited; and

what should happen if an author wishes to withdraw an article

Portland press col ad

what constitutes the final version and how this can be indicated.

The technical costs and resources involved in establishing archives

From this technical viewpoint, the main costs will arise from the initial outlay on IT equipment and staffing, and from ongoing costs for these. At this stage it was difficult to assess potential costs for a new agency since we had no idea of the structure and operational requirements of such a body. We did, however, determine the real costs of setting up and maintaining archives at four universities, which would provide the JISC with approximate-figure data on the nationwide cost of establishing e-print archives at all higher and further education institutions. These figures are shown in Table 1. All costs are approximate, due to currency conversions, and are rounded.

Cultural issues

'Cultural' includes political and business-related issues in this context. There has been some useful discussion already in the literature suggesting that cultural change will be necessary before self-archiving becomes the norm.[21] Some of the more concrete issues that we needed to take into account were the following.

Institutional attitudes to archives

Although there are a number of institutions in the UK that have set up archives, most research-led establishments have not yet done this. The likelihood of them doing so, and within a reasonable period of time, was pertinent to which model we finally decided upon. There are several advantages to institutions in establishing open access archives. First, open access accelerates and enhances the impact of scholarly research.[22] Second, it enables improved methods of impact measurement and analysis which in turn can generate better scientometric performance indicators for research productivity, usage and impact. Third, it also enables the generation of standardized online CVs for each institution's researchers and these can be used for internal as well as external (such as the UK's national Research Assessment Exercise[23]) evaluation purposes. Fourth, it helps to monitor and enable the fulfilment of any research-council funding requirements. Also, at the same time as we reported our study in full, the recommendations of the House of Commons Select Committee on Science & Technology were published[1] and these include the following points:

cultural change will be necessary before self-archiving becomes the norm

> **43.** Institutions need an incentive to set up repositories. We recommend that the requirement for universities to disseminate their research as widely as possible be written into their charters. In addition, SHERPA should be funded by DfES to allow it to make grants available to all research institutions for the establishment and maintenance of repositories.
> **44.** Academic authors currently lack sufficient motivation to self-archive in institutional repositories. We recommend that the Research Councils and other Government funders mandate their funded researchers to deposit a copy of all their articles in their institution's repository within one month of publication or a

Table 1 Examples of actual costs incurred for setting up and maintaining institutional archives

University	Set-up costs (£)	Annual running costs (£)
MIT (using DSpace software)	1.3 m.	160,000
Queen's University (Canada) QSpace (using DSpace software)	22,750	22,250
National University of Ireland, Maynooth (using Eprints software)	17,500	26,250
Nottingham University (using Eprints software)	3,900	31,250 (includes provision for a triennial upgrade of hardware and software)

reasonable period to be agreed following publication, as a condition of their research grant.

If these recommendations are followed up by the government, archives will be set up by every university and research-led institution in the UK.

The populating of archives

Where archives exist today – and with notable exceptions – in the main they are rather sparsely populated with e-prints. Administrators or champions of existing archives have tackled the problem in a number of ways – sustained advocacy by campaigns, demonstrations, presentations and seminars being the main route. Support – tacit or actual – from the pro-vice chancellor (PVC) or provost responsible for research policy is crucial.

a real and non-negligible cost element here

Author inertia is the main enemy of an e-print archive once it is established. The alternative to the 'author chooses to comply' model is to mandate self-archiving. To date, there are a few educational institutions that have gone so far as to mandate that their authors deposit copies of all their research articles in the institutional e-print archive.[24] There are also examples of departmental mandates, one such being the School of Electronics and Computer Science at the University of Southampton, which has produced a policy that could be used by other departments travelling the same route.[25] To allay fears about the process of self-archiving and its legality, Eprints.org has produced a FAQ[26] and a handbook[27] on the subject.

The agreements that authors have with publishers and with archives

Are there any restrictive licensing arrangements with publishers or exclusive agreements with archives that hamper or deter author self-archiving activity? Many publishers have now officially endorsed the practice of authors self-archiving the articles published in their journals in the author's own institutional archive. At the time of writing, over 70% of the (103) publishers surveyed by Eprints.org have adopted this policy,[28] and 92% of the (8853) journals surveyed are 'green' (i.e. they endorse author self-archiving of either the preprints or postprints of articles).[29]

The management costs and resources involved in establishing archives

These include the staff resources required for planning, promoting and training. Hard data were difficult to come by, largely because in institutions where archives have been established these sorts of costs have been absorbed into existing activities and provisions, but it was clear from our discussions that there is a real and non-negligible cost element here.

With this background information in place, we set about determining the candidate models for e-print delivery, management and access.

The models we considered

There are three basic models that can support access to metadata and the associated scholarly digital resources:

- *Centralized model* both metadata and the resources themselves are deposited *directly* in a central archive.
- *Distributed model* – all metadata and resources remain in their source archives, and metadata are cross-searched 'on the fly'.
- *Harvesting model* (a hybrid model) – metadata are harvested into a central searchable archive *but* also remain distributed among the original archives.

The centralized model

In this model, users (authors) deposit their e-prints in a central archive. This would have a service provision component of its own, providing the interface through which users (readers) search, browse and retrieve the articles they require. The metadata of articles in the archive would also be exposed via OIA-PMH (and any other protocols that play the same role; specifically, in our study, we also included RSS and SRW/SSU as possibilities in this respect) for use by other service providers. The configuration of this model is shown in Figure 1.

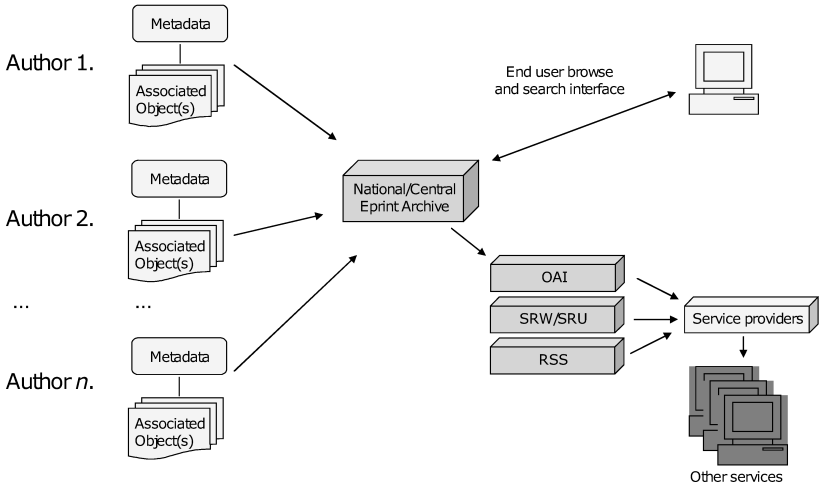


Figure 1 The centralized model.

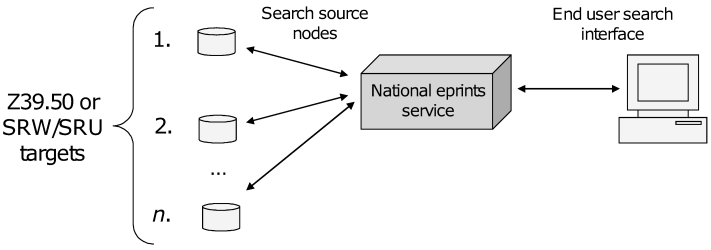


Figure 2 The distributed model.

The distributed model

Under this model, the service would search all available archives, metadata would be obtained in real time as the user made his/her request, and the user would be pointed at the digital resource (the article) located in its distributed archive. The model is configured as in Figure 2.

The 'harvesting' model

Under this model, the proposed new UK service (the *service provider*) would harvest and store metadata from available e-print archives and open access journals (the *data providers*), using the OAI-PMH. It would have a service provision component of its own, which would provide the interface through which readers would search, browse and retrieve articles. We included an additional element here – the metadata from these articles would also be exposed via OAI-PMH, SRW/SRU and RSS for use by other service providers. Figure 3 illustrates the harvesting model diagrammatically.

A detailed account of the technical requirements of each of the models appears in an appendix to the full report.[4]

Each of these models had arguments for and against it. We needed to weigh these up before we settled on which model would be best for a new UK e-prints service to adopt. In considering the relative merits of the models we addressed not only technical concerns but also the cultural issues discussed earlier, and especially how e-print provision (by authors) can be achieved, since without this content provision there can be no effective e-print delivery service (for users). The points for and against each of the candidate models are summarized in Table 2.

each of these models had arguments for and against it

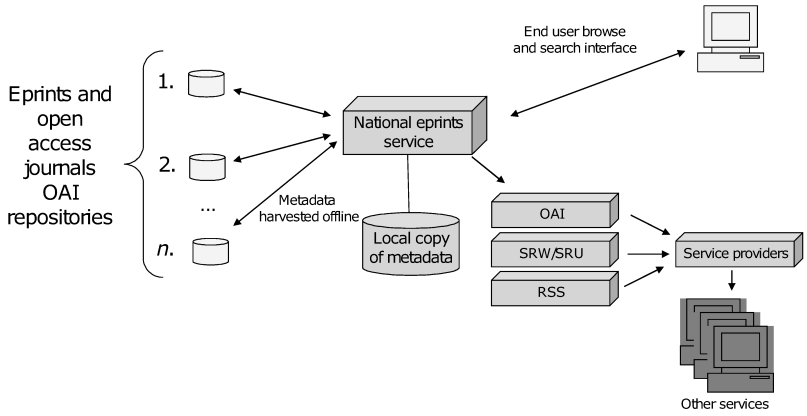


Figure 3 The harvesting model.

The model we recommended

it was clear to us that the harvesting model had the greatest promise

For technical and cultural reasons, the study recommended that the centralized model should not be adopted for the proposed UK service. First, this would have been the costliest option. Second, it would have omitted the growing body of content in distributed institutional, subject-based and open access journal archives. Third, the central archiving approach is the 'wrong way round' with respect to e-print provision (see below) and would therefore not provide so effective a route to a critical mass of e-print material as the other models.

The distributed model had some distinct advantages over the centralized model in respect of the points above, but operationally it did not allow the kind of quality of service that we believed was possible to attain. In particular, the consistency and quality of metadata was out of the control of the agency that would run the new service, yet the quality of metadata is profoundly important to the overall usefulness and effectiveness of such a service. This model, we felt, was adequate but not optimal.

It was clear to us that the harvesting model had the greatest promise. Not only is it compatible with the technical requirements and capabilities we had identified, and provides the means to standardize, improve and enhance the metadata and, hence, service level, but it also provides the most effective means of overcoming one of the major 'cultural' obstacles, which is the *provision* of e-print content in the first place. One of the critical aspects of our decision was that any model for delivering e-prints must operate in, and help to create, the arena most likely to generate the maximum amount of e-print content-provision by authors. Since this issue is so important, it is worth developing the argument further here.

Two things have a bearing on the level of author self-archiving – archives being available for authors to use and authors actually archiving their articles. From the evidence we looked at – existing archives – it was clear to us that even when archives are available there is still precious little peer-reviewed material yet being deposited, so it is author behaviour that is at the very root of this matter. How may authors be 'encouraged' to self-archive? The evidence shows that while the carrot of increased visibility and impact does prompt a proportion of authors to archive their work, 'encouragement' would best also take the form of a stick by self-archiving being made mandatory as a condition of funding or employment.

There are few examples of such mandates in operation as yet (though where they exist, they are working[25]), but plenty of promise for those to come. A recent study on open access publishing produced clear evidence that authors have, in general and in principle, no objection to self-archiving and will comply with a mandate to do so from their employer or research funder: in total 77% of authors would comply with such a mandate (69% would do so *willingly*), while only 3% said they would not comply.[30,31]

Table 2 Advantages and disadvantages of alternative models

	Advantages	Disadvantages
Centralized model	The agency running the service would: – have overall administration of the whole process, from article deposition through to the user interface – be able to standardize the protocols used – be able to select the archive software that provided the most appropriate set of storage and output capabilities – be able to manage preservation issues – be able impose requirements for the format in which articles are deposited – be able to develop facilities that maximize search capabilities (categorization of the data, subject classification, etc.) – be able to establish an overall programme of continuing development and improvement	With all administrative and maintenance functions centralized, it is an expensive option It ignores the existence of, and renders useless, already-established institutional and subject-based archives Creating a scheme for nationwide author deposition of articles within or across disciplines in one central pan-disciplinary archive, or multiple central disciplinary archives, would be extremely difficult if not impossible, for political or cultural reasons It is not reasonable or practical to expect open access journal publishers to submit articles they publish directly into a central archive
Distributed model	No replication of metadata is required The metadata retrieved are always current It provides a consistent look and feel for searching and retrieving metadata from heterogeneous sources It is relatively cheap to implement compared to a centralized solution	The model does not permit any improvements to be made in the management of e-prints and open access journals It does not permit enhancements to the metadata, because these are only grabbed at the time of need (when the user searches) As the number of sources to be searched increases, performance decreases it can only work as fast as the slowest server in the group of archives it is searching Query syntax varies across source nodes, and syntax changes over time If results are to be returned using relevance ranking, it is difficult to merge results from multiple sets in a meaningful manner The institutional and subject-based archives employ software that supports the OAI-PMH. At the time of writing the vast majority of archives do not support Z39.50 or SRW/SRU.
Harvesting model	The OAI-PMH is a standard protocol that is easy to implement It is flexible although the use of unqualified DC is mandated to be OAI-compliant, additionally other richer, more complex, metadata schemes may be employed The OAI-PMH is designed to allow metadata exchange and the sharing of scholarly knowledge The institutional and subject-based archives employ software that supports the OAI-PMH. Much of the harvesting can be carried out by automatic scheduled tasks, minimizing the need for human intervention Once stored in a local database, the metadata can be processed, enhanced and re-exposed both to the original data providers and to other service providers It is possible to develop facilities that maximize search capabilities (categorization of the data, subject classification, etc.) It can form the basis for an overall programme of continuing development and improvement It is a low-cost option which can work equally well for journal articles, e-prints, journal descriptions and collection level descriptions.	Unqualified DC, which is mandated as the minimum metadata standard for use by the OAI, is the only metadata scheme in common use as yet. It is a lowest common denominator which lacks semantic richness and limits the possibilities of providing enhancements The metadata exposed by the service may not always be the very latest version of that metadata. Changes made to metadata at institutional archives, subject-based archives and open access journals will not be reflected until a subsequent re-harvest

mandating self-archiving in institutional *archives is the optimal way forward*

The recent parliamentary recommendations in the UK[1] on mandating self-archiving in institutional archives – published as our study was concluding – are therefore perfectly on target to address the issue most critical to open access provision. Scholars will self-archive if told to do so. Employers and research funders have the authority to do the telling, but tell authors to do what, and which authors? Funders can only tell their grant-holders, but they do have the choice of telling them to deposit their articles in the funder's own archive (if there is one), in some other centralized archive, or in the researcher's own institutional archive, or all of these.

Employers can do all these too, but since they not only have shared goals with their researchers in respect of dissemination of research findings, but also see additional value in, and uses for, the content of an institutional archive, they are very likely to be eager to see it maximally populated and will insist on authors depositing there, at the very least. Moreover, they can mandate and monitor self-archiving across the board, including by researchers who are not supported by external funding (a large number in many subject areas), and in *every* scholarly discipline. This is far more effective a route to comprehensive e-print provision than relying on funder mandates alone, and is much more likely to provide e-prints in *all* disciplines relatively quickly than relying on the eventual establishment of centralized archives in all subject areas. A two-pronged attack from funders and employers (universities and research institutions) mandating self-archiving in *institutional* archives is the optimal way forward.

Finally, many publishers have formally endorsed the practice of authors of articles published in their journals self-archiving them locally in institutional archives or departmental or personal websites, but will *not* permit them to self-archive in 'third party' archives. A centralized model for the new service would presumably be viewed as belonging in the 'third party' category and would therefore suffer from publisher prohibition policies.

Our conclusion was, then, that this scenario is the one most likely to provide the maximum level of self-archived content, a major plank of any model for the provision of e-prints nationwide in the UK.

Implementation of the harvesting model and services based upon it

Harvesting methods

There are several fundamental ways in which metadata might be harvested from OAI-compliant archives:

1. Harvesting from institutional archives, subject-based archives and open access journals is carried out at a national, central level. Then subject-based and other service providers harvest or cross-search subsets from the central national service
2. Harvesting within subject disciplines is carried out by subject-based service providers. These then act as data providers to national services.
3. Harvesting by resource types – e-prints/OAJs, e-theses, reports literature – is carried out by agencies dedicated to those types. These agencies act as data providers to national services

The first option is the most realistic and workable at this time, though the second and third options do have some advantages, not least the modular management of services. Neither the subject portals required for option 2 nor the hypothetical agencies required for option 3 are in place, however, so at this time harvesting at national level (option 1) offers the best route to consistency and avoidance of duplication of effort. In fact, option 1 already exists in prototype in the Eprints UK project.[32]

Alternative services based upon the broad-level harvesting model (option 1 above)

We see three basic ways in which the harvesting model might be used as a basis for e-print and open access journal content service provision in the UK.

An ePrints UK-type service

ePrints UK, one prototype, harvests metadata from e-print archives and enhances them through web services. Data providers (the universities and other research institutions) can then re-harvest their enhanced records to improve their own local services. Meanwhile, ePrints UK provides access to the whole dataset via its own interface.[33]

A portal-in-a-browser service

Our view is that this model is simple, elegant and in keeping with the JISC Information Environment architecture. It has been adopted by the European Library project.[34] This model differs from the ePrints UK model in that we have stripped out the web services (though these may be added in again at a future date when they have matured) and added a central archive that 'mops up' articles deposited by authors who do not have an institutional archive to use. The model is shown in Figure 4. All the protocols used in this model are standard protocols, which are straightforward and inexpensive to implement.

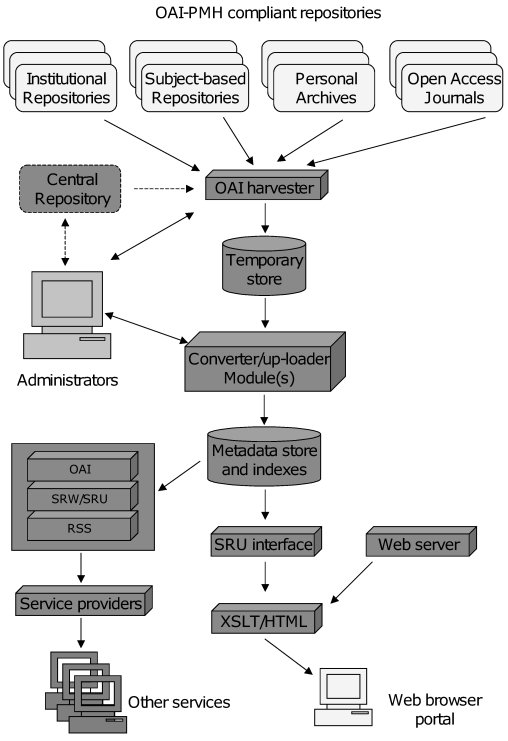


Figure 4 Portal-in-a-browser model.

The Google service model

The use of Google to search university archives using DSpace via a search system set up by OCLC is now being piloted.[35] If trials prove successful, and if the pilot can be extended to search archives powered by software other than just DSpace, this may provide a complementary strand to the other two models.

Whichever of these detailed service models is ultimately adopted for a national UK open access service, it is clear that the parent-level harvesting model is the one which should form the basis of a UK service. Its advantages heavily outweigh its disadvantages, and overall it presents a superior option to the centralized and distributed models that form the alternatives. The Open Archives Initiative employs a philosophy whose time has come, and the harvesting model has gained worldwide acceptance. It makes it easy to share information about scholarly resources and to offer enhanced resource discovery tools, and its use is becoming widespread elsewhere. In view of this, we recommended that the harvesting model should form the basis of open access service provision in the United Kingdom.

the parent-level harvesting model is the one which should form the basis of a UK service

References

1. UK House of Commons Science and Technology Select Committee Recommendations: http://www.publications.parliament.uk/pa/cm200304/cmselect/cmsctech/399/39903.htm
2. National Institutes of Health recommendations: http://thomas.loc.gov/cgi-bin/cpquery/?&db_id=cp108&r_n=hr636.108&sel=TOC_338641&
3. Rowland, F., Swan, A., Needham, P., Probets, S., Muir, A., Oppenheim, C., O'Brien, A. and Hardy, R. Delivery, management and access model for E-prints and open access journals. *Serials Review* 2004, in press.
4. Swan, A., Needham, P., Probets, S., Muir, A., O'Brien, A., Oppenheim, C., Hardy, R. and Rowland, F. Delivery, management and access model for e-prints and open access journals within further and higher education. 2004. http://www.jisc.ac.uk/journals_work.html
5. Open Archives Initiative Protocol for Metadata Harvesting. http://www.openarchives.org/OAI/openarchivesprotocol.html.
6. OAIster. http://oaister.umdl.umich.edu/o/oaister/
7. Citebase Search. http://citebase.eprints.org/cgi-bin/search
8. Directory of Open Access Journals (Lund University Library). http://www.doaj.org

9. BioMed Central. http://www.biomedcentral.com/
10. Scientific Electronic Library Online. http://www.scielo.br/scielo.php/lng_en
11. arXiv.org e-print archive. http://arxiv.org/
12. Cogprints: Cognitive Sciences Eprints Archive. http://cogprints.ecs.soton.ac.uk/
13. Institutional Archives Registry. http://archives.eprints.org/
14. Eprints-UK statistics. http://eprints-uk.rdn.ac.uk/stats/
15. Daedalus project at Glasgow University. http://www.lib.gla.ac.uk/daedalus/
16. Needham, P., Sidwell, K., Bevan, S. and Harrington, J. The MAGiC Project. managing access to grey literature collections. Final report – October 2002. http://www.bl.uk/concord/docs/magic-final.doc
17. McGauran, P. $12million for managing university information. 2003. http://www.dest.gov.au/Ministers/Media/McGauran/2003/10/mcg002221003.asp
18. Eprints UK. http://www.rdn.ac.uk/projects/eprints-uk/. 2004.
19. SHERPA project. http://www.sherpa.ac.uk/
20. Focus on Access to Institutional Resources (FAIR) Programme. http://www.jisc.ac.uk/index.cfm?name=programme_fair
21. Pinfield, S. Open archives and UK institutions: an overview. *D-Lib Magazine* 2003:9 (3). Available at http://www.dlib.org/dlib/march03/pinfield/03pinfield.html
22. Harnad, S. and Brody, T. Comparing the impact of open access (OA) vs. non-OA articles in the same journals. *D-Lib Magazine*, 2004:10 (6). http://www.dlib.org/dlib/june04/harnad/06harnad.html
23. Harnad, S., Carr, L., Brody, T. and Oppenheim, C. Mandated online RAE CVs linked to university eprint archives: enhancing UK research impact and assessment. *Ariadne*, 2003:35. http://www.ariadne.ac.uk/issue35/harnad.
24. Eprints.org. Registry of Departments and Institutions who have adopted an OA self-archiving policy. http://www.eprints.org/signup/fulllist.php
25. Eprints Handbook: Actions for Departments to Achieve Open Access. http://software.eprints.org/handbook/departments.php
26. Self-Archiving FAQ for the Budapest Open Access Initiative (BOAI). http://eprints.org/self-faq
27. OSI Eprints Handbook. http://software.eprints.org/handbook/
28. RoMEO self-archiving policies by publisher. http://romeo.eprints.org/publishers.html
29. RoMEO summary statistics. http://romeo.eprints.org/stats.php
30. Swan, A. and Brown, S. JISC/OSI journal authors survey report. 2004. http://www.jisc.ac.uk/uploaded_documents/JISCOAreport1.pdf
31. Swan, A. and Brown, S. Authors and open access publishing. *Learned Publishing*, 2004:17 (3), 219-24. http://www.keyperspectives.co.uk/OpenAccessArchive/Authors_and_open_access_publishing.pdf
32. Cliff, P. ePrints UK – architecture: 1.032003, based on project proposal. 2003. http://www.rdn.ac.uk/projects/eprints-uk/docs/technical/architecturev1.032003/
33. Martin, R. ePrints UK: developing a national e-prints archive. *Ariadne*, 2003:35. http://www.ariadne.ac.uk/issue35/martin
34. van Veen, T. and Oldroyd, B. Search and retrieval in the European Library: a new approach. *D-Lib Magazine*, 2004:10, Feb. http://www.dlib.org/dlib/february04/vanveen/02vanveen.html
35. Macleod, D. Google launches research archive project. *Guardian Unlimited*. 2004. http://education.guardian.co.uk/higher/news/story/0,9830,1191090,00.html

Alma Swan and Sheridan Brown
Key Perspectives Ltd
48 Old Coach Road, Playing Place
Truro TR3 6ET, UK
Email: aswan@keyperspectives.co.uk

Paul Needham
Information & Library Services
Cranfield University
Cranfield, Bedfordshire MK43 0AL
UK
Email: paul.needham11@btinternet.com

Steve Probets, Adrienne Muir, Charles Oppenheim, Ann O'Brien, Rachel Hardy and Fytton Rowland
Department of Information Science
Loughborough University
Loughborough, Leicestershire LE11 3TU
UK