



Scoping study for a registry of electronic journals that indicates where they are archived

Submitted to:	JISC
Version:	Final
Date:	14 January 2008

Department of Information Science,
Loughborough University
Ashby Road
Loughborough LE11 3TU
Tel: 01509 223064

Rightscom Ltd
Linton House
164/180 Union Street
London SE1 0LH
UNITED KINGDOM
Tel: +44 20 7620 4433
www.rightscom.com

Executive Summary

This study was commissioned by the JISC to examine the feasibility and possible scope of a registry to hold information about where e-journals are archived.

The project team carried out desk research and conducted twenty-two interviews with representatives of archiving organisations, national and university research libraries, publishers, and other organisations involved in the scholarly communication process. An interim report was distributed to all participants for comment and a workshop was held in London to discuss the findings. A full list of the organisations that kindly assisted the project is appended.

The research and especially the interviews have confirmed the assumption behind the project that there is a need for more information, and more easily accessible information, about where e-journals are archived. However, what has also emerged strongly is that this issue cannot be considered in isolation, either from the overall context of relationships within the scholarly communication system, nor from other initiatives being undertaken to improve information flows e.g. in relation to the transfer of journal titles between publishers.

The other clear conclusion is that the library community would expect a solution to the information problem which went beyond simply recording where a journal title was archived and tracking any changes over time. Librarians feel they need to understand and be reassured about what their access route would be to a title in the event of problems in their normal means of access.

The scope of this registry is specifically about where e-journals are *archived*, but discussion of the idea of a registry raised much wider concerns about the role of libraries as custodians of e-journals, and the difficulties they experience in guaranteeing uninterrupted access to them and in managing licences. It even led to a discussion about the cost of access to back archives, which is quite outside the scope of this project, but relates to the general concerns many libraries feel about how they should continue to ensure that they are custodians of their collections into the future. This illustrates how important it is to view the registry as part of an overall information ecosystem. In fact, it will be impossible to scope the registry without viewing its place within this system. Ultimately, what all libraries want is to deliver seamless and uninterrupted access to e-journals (and other material) for their users; this encompasses everything from day-to-day management of electronic resources (dealing with temporary technical problems, subscription administration, title transfers, changes in access models, authentication issues and complying with licence conditions) to being assured that content is preserved and accessible for the long term. There have been continuing innovations in this process ever since the beginning of electronic journals, including CrossRef and link resolvers to direct users to the 'appropriate copy'; this initiative should be seen as another element in that overall infrastructure, along with for example, the Reli licence registry pilot project currently being funded by the JISC¹.

¹ <http://www.lboro.ac.uk/departments/lis/disresearch/RELI/>

It also indicates the need to ensure that the necessarily limited role of the registry is properly communicated, if the decision is made to go ahead

The general issue of access routes has exposed some further questions about how complete and accurate the information record is (either on the library/consortia side or the publisher side) of contractual relationships formed over the period since e-journals began to be available. To take just one example, there are considerable ambiguities about rights of archival access in cases where libraries received an electronic copy of a journal 'free' with a print subscription. A registry in and of itself cannot be expected to solve these problems; at the most, all it can do is record the outcome. It is very important for all users to be clear about what the registry is and is not able to achieve at various stages of its implementation; in particular it is vital that librarians are not given the impression that because a title is in the registry, they will necessarily have access to it.

Librarians felt that they were most likely to consult a registry in situations where they were considering taking out or renewing a subscription; considering cancellation of a print subscription in favour of an e-only subscription; contemplating relocating or discarding print holdings. The vast majority of potential users of such a registry would be library staff in university and national libraries, though organisations licensing e-journals on behalf of the library community would also be likely to use the registry to check compliance with licence conditions.

One of the key benefits of a registry is perceived to be the exposure of gaps in archive provision. This was identified by all types of stakeholder: librarians would want to be alerted to risks to any of their holdings; publishers who are making provision would like to see their efforts recognised and pressure placed on publishers who are not making satisfactory arrangements; archive organisations would also benefit as that effect fed through to more demand for their services.

The drawbacks to a registry as a solution to the acknowledged information gap were mainly seen as ones of practicality (keeping the information accurate and up to date), trust (especially whether a national solution is appropriate, and conversely whether an international solution is feasible) and sustainability of the funding model. Other solutions were suggested, mainly involving either WorldCat or ERM vendors such as Serials Solutions. The latter were also suggested as a complementary part of a solution involving, but not limited to, a registry.

Apart from the desirability and feasibility of the registry, there is a separate question as to whether now is the right time for such an initiative, or if it is premature in relation to the problem of archiving and preservation itself; we have set out the arguments on both sides. We have also sketched out a possible phased approach to building a registry and also suggested a pilot stage which would give some indication both of the scale of the problem in terms of titles not being archived anywhere and of the real practical difficulties an effective registry would face in terms of data collection and quality.

Above all else, the question of the sustainability of a registry needs to be carefully considered. The primary purpose of a registry or registry-like service would be to direct libraries to the archives where particular journals are located and preserved; the archives themselves have to be sustainable over the long term, and to be of any use whatever, the registry must be equally long-lived.

Introduction

This study was commissioned by the JISC to look at the feasibility and possible scope of a registry of archived e-journals. The project has interviewed 22 people from a range of stakeholder groups: potential users (mainly university libraries and consortia); archiving organisations and national libraries (there is overlap in these); publishers; and a number of other groups concerned in one way or another with the journal supply chain (e.g. UKSG), with digital preservation, or with scholarly communication generally (e.g. RIN). Interviews have been carried out with international organisations as well as UK-based ones, as befits this very international community. Each interview lasted between 20-45 minutes. A core script was used, adapted to the various interest groups. An interim report was distributed to all participants for comment and a workshop was held in London to discuss the findings.

Background

Concerns about e-journal archiving have increasingly come to the fore in recent years, driven by the accelerating move towards electronic-only access to journals, and the potential divergence between the electronic and print versions of journals, as more features such as interactivity and links to underlying datasets are built into articles. Librarians are under increasing budgetary pressures, including the need to save physical space and cut labour costs, which is leading many to consider cancelling print subscriptions, moving print collections to remote storage, or discarding print archives. One of the principal barriers to moving towards e-only journal subscriptions is the level of uncertainty about the security of electronic archives, both in terms of effective long-term preservation, and in relation to the conditions under which material can be accessed by subscribing institutions. As part of a possible solution to these problems, this project seeks to investigate the feasibility and role of a registry of information about archived e-journals.

E-journal archiving services

A number of archiving services for e-journals now exist, and it is not the place of this study to examine them in any detail; this was done in 2006 by the CLIR report "E-journal Archiving Metes and Bounds: a survey of the landscape"². Its scope was very wide, including evaluation of the various archiving organisations on a number of criteria, and placing them in the context of legal e-deposit and open access repositories. The report also examined the degree to which publishers were participating in e-journal archiving services. It drew explicit attention to the problem that there is no central list of the titles being archived, as well as inconsistencies in the amount of information and the way it is presented among the archiving organisations. Among an extensive list of recommendations, the ones that are most relevant for this project are:

Recommendations: Academic Libraries and Organizations

5. Libraries should participate in developing a registry of archived scholarly publications that indicates which programs have preserved them. This registry can

² <http://www.clir.org/PUBS/abstract/pub138abst.html>

be used to identify gaps in publisher or content coverage. Models for such registries include the Registry of Open Access Repositories (ROAR) and ROARMAP.

6. Libraries should lobby e-journal archiving programs to participate in a network that shares information, codifies best practices, and promotes sufficient redundancy, and also shares responsibility for preserving peer-reviewed e-journals that are not currently included.

Recommendations: Publishers

- 1. Publishers should be overt about their digital archiving efforts and enter into archiving relationships with one or more e-journal archiving programs of the sort described in this report.*
- 2. Publishers should provide enough information to e-journal archiving programs to ensure that the scope, content, date span, and title coverage are adequately recorded.*
- 3. Publishers should extend liberal archiving rights in their licensing agreements with content aggregators and consortia. Digital archiving of e-journals should be a distributed responsibility.*

Recommendations: E-Journal Archiving Programs

- 2. Archiving programs should be overt about the publishers, titles, date spans, and content included in their programs. They should make this information easily accessible on their Web sites.*

A previous consultancy study undertaken by Maggie Jones³ on the subject of archiving of e-journals for JISC examined the then existing risks to the research record, the costs involved in archiving, and the potential for collaboration and central co-ordination. It made a number of recommendations, including on revisiting the archival clauses of the Model Licence, and maintaining a watching brief on third-party archive services, both of which have been progressed by JISC. It also discussed the LOCKSS initiative and recommended a technical analysis to examine its potential role in preservation.

A JISC briefing paper⁴ on journal archiving and preservation drew the following conclusions about the state of play of archiving services:

There are promising developments which make it simpler to make an informed choice than was the case when JISC commissioned a consultancy on e-journal archiving in 2003. Independent, trusted e-journal archiving services are emerging and those which offer the most promise for the UK environment should be supported. They are still at an early stage and the community needs to invest in different options in order to retain a vigorous and healthy diversity of services most likely to be able to accommodate the range of requirements.

The landscape is still rapidly evolving and it will be necessary to monitor developments closely and to encourage and facilitate regular communication between the three principal parties – libraries, publishers, and archiving services.

³ Jones, M, Archiving E-Journals Consultancy Final Report, October 2003
http://www.jisc.ac.uk/uploaded_documents/ejournalsfinal.pdf

⁴ JISC, E-Journals: Archiving and Preservation Briefing Paper
http://www.jisc.ac.uk/publications/publications/pub_ejournalspreservationbp

A report entitled *The E-Only Tipping Point for Journals*⁵ was published in December 2007 by the Association of Research Libraries. Among many other issues, the report discusses the question of preservation as a barrier to the transition to e-journals. It found that some libraries were participating in collective preservation activities involving print copies, but these depended on the continuing provision of print by publishers, and also did not take account of the increasing divergence between the print and electronic copies of journals. Librarians interviewed observed that it can be very difficult to discover what arrangements are in place to assure access:

Although the importance of preservation is widely acknowledged, license agreements do not always make it clear what responsibilities publishers are agreeing to take on. A collections officer observed, "Lack of clarity within licenses concerning ongoing access rights or archiving arrangements is a problem. To summarize that for our selectors often requires that we go carefully through the licenses and puzzle over what we may find there, or not find!"

The study also found that some libraries were pressing ahead with discarding print in the hope that solutions would be found, while others, in more research-intensive institutions which saw themselves as having a mission to maintain collections, were more cautious. They are therefore participating in initiatives such as LOCKSS or Portico. In the words of the report

As libraries implement e-only policies some are concluding that access to an aggregator's collection is adequate for non-core titles and they are canceling the print even though permanent access is not guaranteed—a potentially risky strategy. For others the stewardship role is a priority and they assume responsibility for insuring access to the content in either print or electronic form. They see the need to demonstrate that electronic archives can meet expectations when the need arises, and thus to "put the trust in trusted archive."

Registry models

There is really no single view of what constitutes a registry. Bodies describing themselves as such have been created for many different functions, for example, recording details of licensed drivers or practitioners, allocating domain names, recording consistent definitions of data (metadata registries). They may be actual physical locations where an act of registration takes place, or virtual entities. It is important in our view that the concept not be too narrowly defined in this case. For example, a registry tends to imply a monolithic, somewhat bureaucratic, central structure, perhaps not well suited to the distributed, interlinked nature of digital information. Several interviewees said things such as 'If by a registry you mean x, then I think it's a good idea....'

E-journal archives are themselves distributed; although one publisher expressed the hope that not every government would decide that it required a national archiving solution for e-journals, they recognised that communities feel more comfortable about something fairly local. Rationally, too, it makes sense (as the LOCKSS programme

⁵ Richard K Johnson and Judy Luther, *The E-only Tipping Point for Journals: What's Ahead in the Print-to-Electronic Transition Zone*, ARL, 2007. http://www.arl.org/bm~doc/Electronic_Transition.pdf

implies) to have your archival eggs in more than one basket. A registry might be best conceived as place where information flows in, is audited, and can be downloaded and incorporated into local stores, updated on a regular basis.

In a report of a meeting to discuss UK and US digital preservation initiatives, the following remarks were made which are of relevance here:

There was some discussion of organisational models, especially for shared infrastructure initiatives like the Global Digital Format Registry (GDFR), where issues of long-term hosting and control are of concern. Kevin Ashley raised again the free-rider problem, commenting that the UK Government might legitimately ask why it alone - through The National Archives - funds the PRONOM registry (<http://www.nationalarchives.gov.uk/pronom/>), when it is of wider benefit. David Rosenthal said that the information in registries needed to be both authoritative and up-to-date and wondered if this could be achieved in a distributed model.

Before looking in more detail at the problem in this case, we examined some existing registries with relevance to the project, including GDFR and Pronom, mentioned above:

In its recommendation that a registry should be established by libraries, the CLIR report cited above referred to ROAR and ROARMAP as examples of registries. ROAR has two functions:

- to monitor overall growth in the number of e-print archives and
- to maintain a list of GNU E-Prints sites.

The purpose of ROARMAP is to record the open-access policies of those institutions who are putting the principle of Open Access (as expressed by the Budapest Open Access Initiative and the Berlin Declaration) into practice. Universities and research institutions who officially commit themselves to implementing a systematic policy of open-access provision for their own peer-reviewed research output are invited to describe their policy in ROARMAP. The ROARMAP site states:

Signing will

(1) record your own institution's commitment to providing open access to its own research output,

(2) help the research community measure its progress in providing open access worldwide,

and

(3) encourage further institutions to adopt open-access provision policies (so that your own institution's users can have access to the research output of other institutions as well).

Roar and Roarmap are certainly registries within the scholarly communications arena, but since their focus is on *inviting* participation and since they also have a campaigning aspect, their relevance for the purposes of this project may, in practice, be small. This is because, although accuracy, timeliness and comprehensiveness are clearly desirable attributes of these registries, they are hardly critical, whereas for the proposed e-journal archive registry, these factors would be crucial.

We looked at other examples of registries; format registries such as PRONOM and the Global Data Format Registry, have obvious connections with digital preservation.

PRONOM is an online registry of technical information, a resource for anyone requiring impartial and definitive information about the file formats, software products and other technical components required to support long-term access to electronic records and other digital objects of cultural, historical or business value. It began as an internal resource for the UK National Archives and later became public. The web-enabling of PRONOM (PRONOM 3) in February 2004 represented the starting point for the development of PRONOM as a major online resource for the international digital preservation community. PRONOM 3 has been supported by a number of organisations and individuals. The latest version (PRONOM 4) marks a significant reworking of the underlying data model to allow the capture of detailed technical information on file formats and support future interoperability with other planned registry systems.

The Global Data Format Registry (GDFR) is an international project working to develop a model for a global file format registry. Its membership encompasses the international library and archival communities, including the National Archives. GDFR will be established as a distributed service in which participating research libraries, archives, and other organisations with preservation responsibilities can contribute, as well as use, format-typing information. It will provide sustainable distributed services to store, discover, and deliver representation information about digital formats. The Harvard University Library (HUL) has received a grant from the Andrew W. Mellon Foundation for a two year project leading to the deployment of the GDFR.

GDFR and Pronom are described in a number of sources as complementary.

The OCLC/DLF's Registry of Digital Masters was conceived as a common record of persistent digitised and born-digital materials held by libraries. Its main benefit is to help librarians to check if something in their own collections that they are considering digitising has already been digitised (and is being kept as a persistent object) by another library and whether it has been digitised in the format (for example with illustrations) which they require. Similar to a microform master, a master copy of a digital object serves as the preservation copy from which use or access copies may be made. Metadata may be used to describe both the master and use copy on the same record. A use copy of every registered object needs to be available, but not necessarily free-of-charge. Last year, LIBER and OCLC announced that they *"have agreed to exchange bibliographic records about digital masters. By this agreement, full information about digitized print material from both European and US libraries will be united in a central Registry of Digital Masters, which will be freely accessible for online searching. This collaboration is the first step toward a global registry."* EROMM will be the intermediary delivering European records.

The Library of Congress's NDIIPP is actively considering a registry of content from the participating partners in the programme. It is currently working on questions such as the granularity of information which would be in the registry and use cases. The incentive for participation is the usefulness of sharing such information among the partners in the network, which include not just libraries but commercial organisations and public television. They emphasized to us the need for keeping barriers to participation low.

The European Register of Microfilm Masters (EROMM) is another example of a registry which gives librarians the opportunity to avoid duplicative work. The EROMM database allows microforms and digital forms to be located and accessed: requests are forwarded online to the institution that owns the master copy of a microform, and hyperlinks point to the web location of a digital reproduction. EROMM lets microfilming and digitisation programmes in libraries know whether selected books have previously been microfilmed or digitised elsewhere. Members of the Consortium of European Research Libraries (CERL) are allowed free read-only access to the EROMM database - provided one institution in their country provides data to EROMM and is a member (as is the case for most CERL members).

It may be worth mentioning the Romeo service here, now under the auspices of Sherpa, which records publishers' policies on self-archiving. There are several reasons for considering Romeo. Firstly, it has been suggested to us that archive information could be integrated into it. Secondly, the white, yellow, blue and green classification or something similar might be a useful template for the way information is presented in the archive registry. Thirdly, there was some criticism of it for being inaccurate, used as an example of problems that could arise if this registry was unable to establish good mechanisms for updating.

An article by Lorcan Dempsey in August 2006, entitled "Registries: The Intelligence in the Network"⁶ perhaps comes closest to the model of a registry which is most appropriate in this context. He cites the case of DNS and how without it, "the burden of data collection and configuration on each organisation would be that much greater and the overall efficiency of the network would be much reduced" before going on to say that "this is exactly the situation we are in with higher level network services where we have no such directory services. Increasingly, library applications need to know about a variety of entities. We are used to thinking about information objects (books, journals, maps, etc). What about institutions (suppliers, libraries, etc), policies (e.g. ILL policies), licenses, collections (databases, special collections, summary level descriptions of archival collections, and so on), and services (addresses and interface details for machine users, and descriptions for human users)? The absence of appropriate directory services for each of these reduces the efficiency of the network." He points out that metadata is required to do this, and that indeed, a great deal of metadata exists, but it is "scattered across many systems and services. It may be hardwired into particular applications and not be more generally available....This is where discussions about directories or registries come in."

Key Issues as seen by interviewees

Definition of an e-journal archive

One of the main recurring issues in discussions about archiving concerns what exactly an e-journal archive is and we explored this with all the interviewees. Everyone was aware of the fact that the term 'e-journal archive' was open to several different interpretations. The main perceived 'divide' was over perpetual access/post-termination

⁶ <http://orweblog.oclc.org/archives/001105.html>

archives for provision of continued access to previously subscribed content versus 'dark' archives storing long-term preserved copies, whose contents would only be accessed under strict conditions, known as 'trigger events', for example, following a sustained breakdown in provision from the publisher. (According to the CLIR report, these archives are more properly called 'dim', as a true 'dark' archive would never allow direct access). In some ways this divide is artificial; firstly, perpetual access archives cannot really be considered as such if there is no preservation strategy and practice in place. Secondly, some of the archiving organisations provide both perpetual post-termination access (Portico, OCLC) and long-term preserved 'dark/dim' archives. Nevertheless, this is a perceived distinction among many in the user community.

Other types of archives were also described:

- LOCKSS boxes, which preserve a local archive of a library's holdings, available to the library in the event of failure of their usual access route;
- national library archives held under legal e-deposit laws or by virtue of voluntary deposit by publishers. These may be accessible, but only to users with physical access to a reading room, as the terms of use are based on 'walk-in' access only;
- archives created locally for the purposes of distribution with the consent of publishers
- publishers' own internal archives

The JISC Briefing Paper cited above has this to say on terminology:

*The terms 'perpetual access', 'archiving', and 'long-term preservation' are sometimes used interchangeably. **Perpetual access** is most commonly associated with e-journal licence clauses designed to provide assurance of continued access to subscribed material in certain circumstances, including post-cancellation. **Archiving** describes the process and procedures whereby e-journal content may be managed for the short or long term. **Long-term preservation** refers to the processes and procedures required to ensure content remains accessible well into the future, regardless of any technical or organisational changes.*

Broadly speaking, potential users of the registry and other interviewees whose main role is to broker access to e-journals, while recognising the importance of preservation, are very much focused on the conditions under which they can gain access to archives. Doubts were expressed about whether the 'dark' archives could ever be accessed in any imaginable real-life scenario, which called into question their usefulness. One interviewee said she expected access to be opened only in the event of a third world war, in which circumstance getting hold of an e-journal would not be uppermost in her mind.

Since then, a trigger event has taken place involving Sage and Portico⁷; Sage has decided not to continue publication of the journal *Graft: Organ and Cell Transplantation*,

⁷ <http://www.portico.org/news/112807.html>

and because it is archived in Portico, that section of Portico's archive is now 'light' for the purposes of giving member libraries access to this journal, regardless of whether they subscribed to the journal or not. However, access is only to Volumes 4-6, as the journal was previously published by another publisher, a good illustration of the complexities of such arrangements.

The archiving organisations, on the other hand, together with some other stakeholders, would argue that an archive cannot be properly regarded as such unless it carries out effective preservation. Another issue raised was whether legal e-deposit archives should be included in the registry; clearly they are archives, but from the point of view of university librarians they have serious limitations, both in terms of coverage being limited to publications in the domestic territory, and very limited access.

Another issue that was highlighted was that of trust in the archive, and the question of independent certification of the quality of archives was raised. Clearly, this would affect decisions about whether to include an archive in the registry, as inclusion would implicitly confer a certain status on the archive, simply by virtue of defining it as an archive. The CLIR report discussed the issue of certification. The RLG/NARA trusted digital repository (TDR) certification checklist defines a set of management policies that establish the characteristics of a repository for digital preservation. The registry rules would have to determine whether only archives meeting a quality standard (which is not yet in place) should be included, or perhaps include information on others but identify them in a different way. The RLG/NARA work has been built upon by the Certification of Digital Archives Project, which tested auditing methods on several participating archives, including the KB e-Depot, Portico and LOCKSS. The project listed the general factors to be considered when arriving at metrics for assessing and certifying archives. These are as follows:

Organization: The mission and solidity of the organization that supports the repository will affect the repository's prospects for continuity. Repositories vary, from those created for the express purpose of preserving content for academia to those embedded within scientific, publishing and aggregator organizations. It is important to know the extent to which preservation is integral to the parent organization's mission, and how important the repository functions are to that organization's revenue stream.

Governance and Accountability: The governance of the organization that supports the repository determines which communities interests will drive the activities of the repository. How accountable is the organization to the user community, and in what ways is that accountability assured? Conversely, how accountable is the organization to the producers or publishing community?

Content: What content is maintained by the repository and what are its critical characteristics? The extent and scope of the journal titles, databases, and other materials archived should be listed, or easily discovered, and verifiable. What mechanisms are in place to ensure the continued deposit of the listed content, and prevent its withdrawal by the publisher?

Ingestion: Trustworthy repositories will disclose specific data on the form and functionality of the content ingested. Most archives reformat or normalize content in order to limit the cost of managing and migrating complex formats. Normalization may make the archived content look or behave differently than it does when delivered directly to users by producers or publishers. Clarity about the nature and degree of normalization can provide a sense of the scale of investment the library and/or the repository will have to make, if any, to provide an acceptable level of functionality in the future.

Technical Systems and Data Security: The most obvious indicator of the reliability of a repository is the stability and robustness of the technical infrastructure that supports digital preservation. Factors here include whether the repository system conforms to the Open Archives Information Systems Reference Model, to various system security requirements and standards developed in government and other domains, and whether the policies and methods for backup, redundancy, authentication, distribution of functions and services are clear and conform to accepted best practices. Also important is the scalability of the system. Is the repository likely to be able to accommodate new and complex forms of content and functionality?

Cost Structure and Distribution: The costs of a repository can be structured and distributed in several ways, with differing implications for future costs to the library. The repository may assess the library or users a combination of initial capital fees and ongoing maintenance fees, or simply a subscription fee. Some costs might also be borne by the publisher of the archived content. While there are limits to how precisely a repository can project future fees in advance, libraries should be clear about the cost drivers (such as amount and complexity of content, frequency of migration, royalties to content publishers, etc.) and how the costs are distributed in the event of changes in those drivers.

Rights: Repositories should disclose documentation of the rights they hold to deliver the content in the event of failure by the producer or publisher, the duration of the grant of those rights, and whether said rights are transferable. Such documentation should be clear about what constitutes failure. Failure is often defined as when a publisher no longer offers the content, but drastic subscription price increases, the decision to make the content available only as part of a larger, prohibitively priced bundle, and similar events can also put content out of reach of libraries.

Results and Outputs: Longevity and performance are important indicators of the reliability of a repository. While digital preservation is only just emerging, organizations and systems that have demonstrated competency in effectively fulfilling preservation functions are likely to continue to support persistence

Subsequently various international organisations met and agreed certain core requirements:

The attendees identified what they believed were ten basic characteristics of digital preservation repositories:

- 1. The repository commits to continuing maintenance of digital objects for identified community/communities.*
- 2. Demonstrates organizational fitness (including financial, staffing structure, and processes) to fulfill its commitment.*
- 3. Acquires and maintains requisite contractual and legal rights and fulfills responsibilities.*
- 4. Has an effective and efficient policy framework.*
- 5. Acquires and ingests digital objects based upon stated criteria that correspond to its commitments and capabilities.*
- 6. Maintains/ensures the integrity, authenticity and usability of digital objects it holds over time.*
- 7. Creates and maintains requisite metadata about actions taken on digital objects during preservation as well as about the relevant production, access support, and usage process contexts before preservation.*

8. *Fulfills requisite dissemination requirements.*

9. *Has a strategic program for preservation planning and action.*

10. *Has technical infrastructure adequate to continuing maintenance and security of its digital objects.*

The key premise underlying the core requirements is that for repositories of all types and sizes preservation activities must be scaled to the needs and means of the defined community or communities

The London workshop for this project also discussed the question of whether e-journals in all 'archives' should be included in the registry, regardless of the status of the archive. It was agreed that they should, but that quality indicators should be displayed, almost certainly based on the progress of the work referred to above, which has resulted in the creation of the Trustworthy Repositories Audit and Certification: Criteria and Checklist⁸. This would reassure both libraries and content owners about the quality and the sustainability of the archives. The registry itself cannot be an auditor but it should accurately reflect the auditing and certification that is happening.

Is there an information gap?

The first issue on which we consulted our interviewees was whether in fact there was a problem that needed addressing, whether by a registry or any other solution. *Almost everyone agreed that there was*, both in terms of an overall lack of information about where e-journals are archived, but more particularly, the difficulty of finding the information across a range of possible sources. People felt that quite possibly the information existed somewhere but that it would require visits to a number of different sources to confirm it. There is also an expressed need for holders of LOCKSS boxes to be able to make comparisons between what is in their boxes and their overall holdings.

If a journal is part of a Nesli deal, the information about post-cancellation access ought to be discoverable in the small print, but not the information about dark archiving. There is now a clause in the model licence to apply from 2008 which requires publishers 'from time to time' to inform subscribing institutions about where their dark archive or archives are located, but this is not yet in operation. That also leaves out journals which are not in a Nesli deal. Nesli is being extended to small and medium sized publishers, and it was also felt that this might expose more issues on archiving and information about it.

However, some potential users felt confident that they would be able to locate the information as to where a particular title was archived quite easily – if a registry was to be useful it would have to contain a lot more information than that, or be part of a wider solution. Others felt that the effort and cost required to create a registry would be out of proportion to its utility; users would be unlikely to consult it very often, and they could actually find the information elsewhere, albeit not so easily.

⁸ <http://www.crl.edu/PDF/trac.pdf>

In terms of knowledge about e-journal archiving services, Portico seemed to be the organisation with the most 'mind share'. No one mentioned their holdings comparison service, however, which allows both participating and non-participating librarians to submit spreadsheets of their holdings and receive information comparing their holdings with those of Portico, with supplementary information on the status of the titles i.e. whether the process is at the stage of an agreement only, or if the titles have been fully ingested and archived.

Is a registry the answer?

Opinion was divided on whether a registry was the optimum solution to fill the information gap. The library community and the publishers were broadly in favour; some of the archiving organisations expressed reservations. The key benefits were seen as:

- it would be a one-stop-shop for librarians, compared with looking in licences, on publishers' websites, contacting subscription agents or aggregators, or searching individual archiving organisations' websites
- It would allow organisations such as JISC Collections to include a stipulation in their licences for publishers to supply information about archives to a single place, rather than requiring them to tell institutions individually
- It would put pressure on publishers to make archive provision for their titles, as the gaps would be much more visible to the community (this argument came from publishers as well as librarians)
- It would encourage transparency and the provision of detailed information by the archiving organisations about their holdings

Those who were less enthusiastic about a registry advanced a number of arguments, many of which concern the quality of implementation, but some are more 'in-principle' objections:

- Setting up new bodies should be avoided if possible, as the landscape is already very crowded
- Its costs would outweigh its benefits
- It would be hard to keep a registry up to date and accurate, as tracking any information about e-journals is notoriously difficult: "e-journals are slippery things" as one interviewee put it
- It may be outside the normal workflow and infrastructure
- It is not clear who would pay in the long term, and a registry tracking preservation archives must be sustainable to have any value whatsoever
- The money would be better spent on something else
- It's a distraction from the real issues over access and ownership of collections
- It's premature, as the problem of archiving of e-journals itself is far from resolved
- Sustainability of such a registry could be very difficult to ensure, and without that, it would be a waste of time, money and effort

What information should a registry contain?

There were two main points of view about the information which people would expect such a registry to contain. The first is that the registry should be 'lightweight' in terms of information. Proponents of this view argue that all the registry needs to have is the simple information that 'x title, for y volumes and issues, is archived in z'. The user will then be pointed to z to check other information, including the conditions of access, the format in which the title is stored e.g. is the original 'look and feel' preserved, and the preservation strategies being followed. The advantage of this approach is that limiting the scope of the registry will allow it to focus on what it is uniquely for, keeping the costs down and making it more likely that the information is accurate and timely.

The second view is that information on how the content can be accessed is crucial. It is not simply the absence of a central directory of where titles are archived that is problematic to many libraries, it is also that they find the information about the terms under which access is allowed confusing and obscure. Proponents of this view also argue that users don't really care if something is in an archive unless they can get at it. There are further layers to be unpicked here: a registry could *in theory* fairly easily record the generic information about formats and strategies, as well as the 'trigger events' under which a 'dim/dark' archive would be opened, though maybe it would not be sensible to do this at the title level. However, the latter is not perhaps as straightforward as it appears. These 'trigger events' differ across the archiving organisations, with several, according to the CLIR report, not able to spell out policies on access to journals where copyright had expired, for example. More importantly, the CLIR report uncovered some issues about whether content would be made available to non-participating (but subscribing) libraries, even after the trigger event occurred, let alone to those with no current subscription. The licences which publishers have made with the different archiving organisations are based on a common core with each organisation, but there can be publisher-specific variations.

The danger would be for the registry itself to confer a level of reassurance about access which was not accurate, and given that this would be one of the key reasons for consulting the registry, this could be a significant problem.

What is much more difficult is to allow a library to see its **own** route to accessing a copy in terms of its actual contractual rights. One of the most worrying gaps in information in this whole scene appears to be complete and accessible records of contracts and licence terms, on both sides of the contractual relationship (consortia and institutions, and publishers) over the period since e-journal subscriptions began. Not only are there apparently gaps, there are also ambiguities: for example, what are the archival rights, if any, attached to an electronic copy which came 'free' with a print subscription? What are the contractual rights for institutions party to a consortial deal where the consortium no longer exists? What is the contractual position for an institution which has merged with another? There are also issues about rights to give access to an archive between societies and publishers; publishers who have paid for its digitisation may assume they have rights to sell the archive, but this can be disputed, especially when a society decides to change publisher.

Essentially, such issues have been and will continue to be settled by negotiation. It may be that bodies such as JISC Collections and their negotiating agents Content Complete

(as well as, or perhaps in concert with, similar organisations internationally) can achieve some collective agreements on certain points rather than institutions themselves having to negotiate individually with publishers. In any case, a registry can only capture this information after the process of negotiation; it cannot itself solve these issues.

A further point which needs to be emphasised is that it is just as important, if not more so, for a registry to be capable of being used to identify titles which are not archived anywhere, and are therefore at risk.

We consider later on the question of where the registry should obtain the information and also how it should be presented.

How should a registry be set up/operated?

There were a number of suggestions made about this, including the JISC itself as an obvious candidate; a board representing the various stakeholder interests; and that CrossRef would be a suitable host for the registry, as it has the trust of publishers and has some of the important information needed as a result of receiving notifications of title transfers for DOI purposes. Content Complete, which negotiates the Nesli deals in the UK, was also suggested, as it has the appropriate relationships with the publishers.

One of the key issues for the registry is the degree to which it is envisaged as primarily for the use of UK librarians, or as an international resource. Clearly, e-journals and many e-journal archiving organisations are international, and it does not make a lot of sense for every country to build its own registry. However, there was a definite sense among UK interviewees that its main purpose would be to serve UK institutions and their researchers, and it could be made available to others as a by-product. Conversely, it was felt by several US interviewees that, welcome as the initiative was, US institutions might have some difficulty with a UK-based registry, and that there was very likely to be much more of a problem with other regions e.g. Asia. This parallels the archives themselves, with a marked preference in many cases for solutions close to home.

However, a crucial issue is that to build a comprehensive picture of where e-journals are being archived requires the co-operation of the archiving organisations, most of whom are based outside the UK. It would make no sense to try to base the registry solely on what is published in the UK, especially as there is no legal basis yet for e-deposit in the BL archive.

What would undermine trust in the registry?

At the top of everyone's list was poor quality of information and lack of timeliness. For most people, there seemed to be little concern about governance, in the abstract, but in practice, it must be assumed that this would be important. The registry would need to be governed in such a way as to secure and maintain the trust of both the library community and publishers.

Alternatives to a central registry

A number of interviewees who thought that setting up a registry was not a good idea suggested alternatives.

- Several people mentioned OCLC's WorldCat as a location where archive details could be added. But other interviewees questioned how widely WorldCat is used, and if OCLC would be seen as a disinterested party
- ERM vendors: these are localised to a library's collections, so just adding another field could enable them to do the job. One interviewee suggested that librarians who are already paying for these services ought to be "knocking on their doors" to demand that this information be included. But while there is certainly a powerful logic to this argument, it would not help libraries that do not yet have an ERM. It appears that even many of those who have them haven't got the staff resources to populate them. The objection was also raised that the ERM vendors are commercial organisations, whereas the solution should be a community one
- A general suggestion that the function of the registry should be attached to something else that already exists in order to leverage existing organisations and infrastructure e.g. Romeo (SHERPA), SUNCAT. This is again a very attractive option in theory, if the chosen organisation were willing and the choice commanded widespread support and confidence

Potential users and presentation

The vast majority feel that it would be likely to be used only by librarians with responsibility for serials or for all e-resources, which carries implications for the interface. Librarians are used to dealing with complexity. Only one interviewee thought end users should have access. There may be other occasional users e.g. JISC itself. Alerts would be a useful feature, for example showing a change in the risk profile of a title. Some quality indicator(s) should be displayed next to an archive service, as outlined above.

Stakeholders

Throughout the report, we have referred to different groups of stakeholders and interviewed representatives of the main groups with an interest in the registry. The following table brings together a list of these stakeholders, their interests and possible roles in the registry.

Table 1: Stakeholders and possible roles in the registry

Stakeholder	Interest(s)	Possible role(s)	Costs incurred, if any
National libraries	As archiving organisations, registry would help to motivate publishers to make archiving arrangements; the registry may also help to reduce the burden on archiving organisations to communicate holdings and other information to individual libraries	Provision of data to registry in capacity as archiving organisations; role in establishing registry; ongoing governance role	Data feed involving initial and some ongoing technical staff involvement; senior staff involvement in establishing and governance of registry
Third party archiving organisations	Registry would help to motivate publishers to make archiving arrangements; the registry may also help to reduce the burden on archiving organisations to communicate holdings and other information to individual libraries	Provision of data to registry; role in establishing registry; ongoing governance	Data feed involving initial and some ongoing technical staff involvement; senior staff involvement in establishing and governance of registry
Publishers	Would help to speed the transition away from print through reassurance to libraries. Would help to ensure that publishers not making proper archiving arrangements were pressured to do so	Support of concept vital, may decide to collectively undertake establishment of the registry under auspices of e.g. CrossRef	If collectively established, costs would be borne; if not 'owned' by publishers, they could be asked to contribute funding or sponsorship for registry
University libraries and organisations grouping libraries (e.g. SCONUL) and representing researchers (e.g. RIN)	One stop shop to check if and where journals holdings are archived; later, registry may be part of infrastructure to provide seamless access to appropriate copy. Collateral for decision to move e-only/store print off site etc.	User requirements, testing; information on holdings and licences in later stages as part of overall solution	Staff costs in attending meetings and testing
JISC Executive	Safeguarding investment in e-journal content; helping librarians to do their jobs efficiently	Initiating, co-ordinating, representing UK HE, funding	Initial funding, possibly long-term contribution to sustaining registry
JISC Collections	Ensuring the material paid for is available for permanent access to UK researchers and students; ensuring licence terms are adhered to	Negotiating licence terms, user requirements	Staff time
Consortia (international)	Ensuring the material paid for is available for permanent access to researchers and students; ensuring licence terms are adhered to	Negotiating licence terms, user requirements	Possible contribution to sustaining registry
Subscription agents	Not clear – service to customers.	Providing historical licence information (but may be providing direct to archiving organisations in future?)	Costs in data provision – no corresponding benefit?
SUNCAT, Cross Ref, Ringgold	None directly	Providers of triangulating data to registry to enable	Technical staff involvement; involvement in

		creation of 'endangered species' list; possible involvement in registry set up/governance	meetings
EDItEUR	None directly	Onix PL use	Involvement in meetings, technical resource

Scenarios and use cases

Scenarios

The following scenarios were identified as threatening continued access to e-journals, both current and archival:

- Journal title transfers between publishers. This is a very important issue and there are already projects addressing this situation and the challenges it presents. Though much of the immediate focus may be on ensuring uninterrupted access to current content, there are equally important issues related to archival access
- Mergers of publishers, which give rise to contractual issues about archival rights e.g. the transfer of Academic Press titles to Elsevier
- Business failure, in particular, fears over the future of smaller publishers
- Technical failure is perceived as less of a threat – at least for the time being

The registry is likely to be consulted in the following circumstances:

- Before discarding or relocating print or taking a decision to move to e-only – this would be to make the case for these moves, decisions which would involve faculty at some level
- Before initiating or renewing a subscription deal – this would be to inform an internal library decision
- Active management of subscriptions to ensure uninterrupted access
- Auditing to check licence terms are being observed from the point of view of accountability for public money – this would be a use by JISC or similar organisations with responsibility for concluding major licensing deals

It was felt that it was unlikely that the registry would be used during the normal course of a subscription. The case where a library might check if there was an archive before carrying out preservation work in-house was not seen as at all likely; no library we spoke to felt that they had the resources or skills to undertake such a task themselves.

Use Cases

- A library is considering moving to an e-only subscription for a journal to save space and labour costs. They need to make a case for this to the faculty committee. They know that they have rights to access the publisher's electronic archive through their current 'print-and-e' subscription but they want to check that the archive is being preserved and in what format, if they needed to access it in the event of problems with the normal delivery via the publisher's server. The senior acquisitions librarian logs on and checks the registry for the information about where the title is archived and then checks the archive provider's site for further information (or gets this from the registry itself). If there is no record in the registry, the librarian contacts the publisher or subscription agent to question its arrangements
- A library is about to renew a 'big deal'. They check the information they currently have in their licence on post-termination access, which is provided via the publisher's own site. They are concerned about what would happen if this route were not available through business or technical failure, so the e-resources librarian logs on to the registry to check on back up arrangements
- A society journal title changes publisher. Publisher A had arrangements with an archive service; under the terms of the contract, the archive, once ingested, stays at the archive service, but Publisher B, which has acquired the journal, has no such arrangement. An alert goes out from the registry to librarians who have requested them about the change of status of the title
- A licensing body such as JISC Collections wants to check if there are any journals for which it has negotiated access to on behalf of the community that are not being properly archived and preserved

How a registry might work

Clearly this depends on the chosen scope of the registry. The registry or registry-like service could be conceived as a 'phased model' with each phase satisfying a cumulative hierarchy of needs, mitigating different degrees of risk, and therefore requiring different levels of information and drawing them from different sources.

The need for a pilot stage

Before going on to outline the phases of a registry operation proper, it is worth considering the value of a pilot stage. This pilot stage would garner the same information as would be required for Phase 1 (see below), namely the holdings of e-journals in all the archiving services (currently ingested and also 'in the pipeline'), with ISSNs, title, volume and issue level data, and compare them with a master list (we are provisionally suggesting SUNCAT). This would yield an 'endangered species' list of e-journals which are currently not apparently being archived anywhere. The list could be subdivided into: 'archived, preserved but inaccessible' e.g. in dark archives, subject to 'trigger events', 'archived, preserved and accessible under specified conditions' e.g. in national library e-deposit archives, and 'not archived at all'. A further refinement might be to ask subject

specialists to gauge the risk to key journals in their fields; the overall ecology might look positive but viewed from the niche, it may look much less so. There is evidently a particular concern about archiving and preservation arrangements for many OA journals, for example.

The pilot stage results would be open to the stakeholders but there would be no external access to the repository itself. Publicity for the headline results could be undertaken, if mutually agreed, following private assessments of validity and also approaches to publishers who are apparently not participating in any archive programmes.

Though the infrastructure for holding and analysing the data should be constructed using widely-adopted repository tools such as Fedora, which could be re-usable for the registry proper, the process of obtaining the data and manipulating it could involve much more manual intervention at this stage than would be possible for an on-going registry.

What are the advantages of such a stage?

- The 'endangered species' list would be an extremely useful output in its own right, driving the preservation agenda itself, and giving an overall mapping of the size and nature of the problem
- It would help to ascertain whether or not the registry is premature in relation to the underlying issue of preservation
- This stage would be likely to command the support of all stakeholders, even those sceptical of the value of a registry, because it would increase transparency and drive 'business' to the archive organisations
- The archive organisations' co-operation is critical for a registry to succeed and they are perhaps the least enthusiastic about a registry
- It would be valuable in preparing for the registry, both in terms of building consensus and getting a more detailed sense of the data quality and interoperability problems a registry would face, as long as it was crystal clear that the pilot stage was a learning and mapping exercise

The 'Registry'

Phase 1 would contain the minimum information for the 'lightweight' specification, namely where journal titles, down to the year, volume and issue level, are archived. It was initially envisaged that this information would come from publishers. In the UK, a model licence condition which required publishers to submit this information to the registry on an annual basis to link with the renewal cycle would act as the leverage for this, but there would need to be a compliance check. However, the archiving organisations made the point that the data they receive from publishers is of variable quality, and has changed over time. One told us that bibliographic data is usually pretty accurate because it is used on the publishers' own website, but that:

"There is a lot of variability in the way the metadata is packaged. Particularly with the larger publishers who have been putting material online for a long period of time, you see four, five, six different flavours. The early years were particularly messy. The big challenge is to normalise it all."

That suggests it makes more sense to obtain the necessary subset of data for the registry from the archiving organisations rather than from the publishers, since it would have to be cleaned up all over again. As an example of what has to be done, Portico states in a description of its ingest processes⁹:

One of the interesting characteristics of e-journal content is that descriptive metadata is abundant; in some cases there is too much metadata. E-journal articles supplied in marked-up SGML or XML (either full text or headers) normally have all the descriptive metadata clearly identified: author, title, journal, volume, issue, date, etc. In some cases publishers even include extra metadata not used directly in the article such as previous titles by which the journal was known or the identity of the copy editor or the date on which the proofs were mailed to the author. Some of this additional metadata is really the publisher's own business process data, not part of the published article. After consultation with the publisher we will remove that non-content information during conversion.

Assuming that information was obtained from all the archiving organisations, gaps could then be identified, by comparing records in the registry with some triangulating source of data, such as the holdings of major research libraries through SUNCAT records, discovering titles which were not listed as being in any archive. As SUNCAT also includes records from the CONSER and ISSN Register databases, this would represent a very good source for an 'authority list' against which to measure the holdings of the archive organisations.

It should be noted here that even data such as the ISSN of a journal is not necessarily unproblematic. In the article referred to above Portico comments:

The descriptive metadata that goes into the Portico METS files is extracted from the NLM DTD article instances. It is then run through a light-weight automated curation process in which it is checked for required values such as ISSN and date of publication and ISSN and journal titles are validated against the master list of journals for which we have archiving agreements. This assures us that we are archiving only content for which we have a contract and also identifies cases where a title change has occurred if we have not already been notified of that by the publisher.

However, though this process validates the ISSN against the agreements between Portico and the publisher, it does not necessarily ensure that the ISSN is initially correct. We know of only one publisher which claims to have accurately mapped the ISSNs of every journal it has ever published.

This is used as an example of how in practice, it can be very difficult to ensure the accuracy of even apparently simple e-journal metadata.

The next phase (Phase 2) would make information available about generic conditions of access, (the 'trigger events'), the 'quality indicators' and more detail on the content and formats of the files stored. One archive organisation told us that libraries had not yet asked for information at the article level, but the degree of granularity of information required would need to be investigated further.

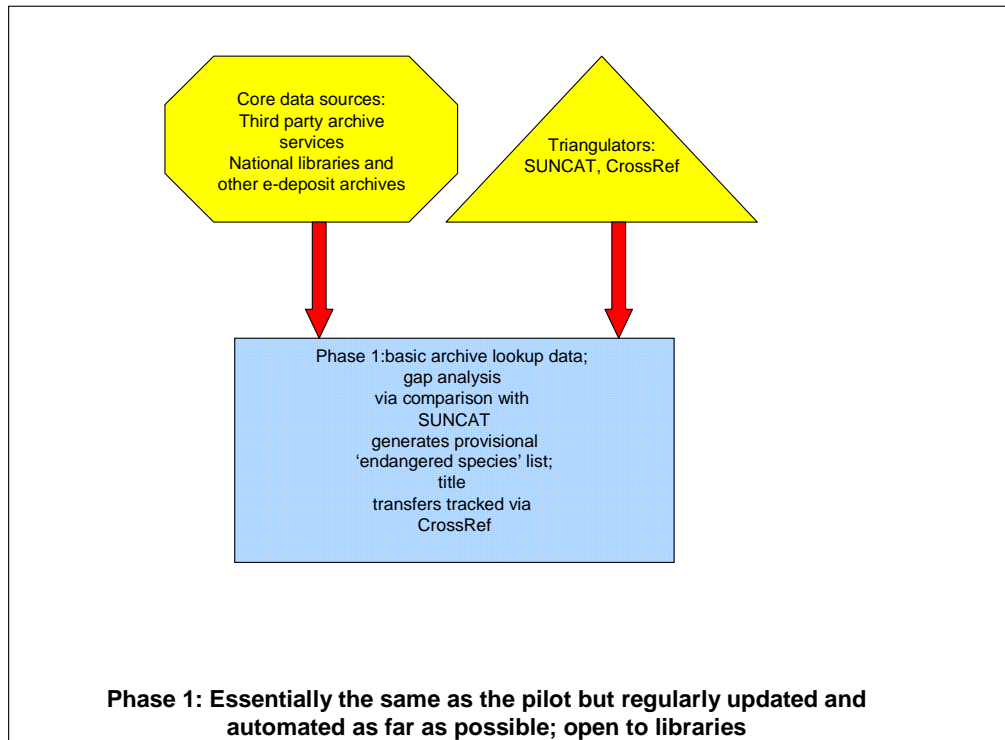
⁹ Owens, Evan Automated Workflow for the Ingest and Preservation of Electronic Journals. Portico, 2006 <http://www.portico.org/news/Archiving2006-Owens.pdf>

The final phases would be able in some way to indicate access routes. Each library has a specific set of entitlements based on its own historical contractual arrangements (and its participation in third party programmes such as Portico) though clearly there is a good deal of information for the Nesli deals, for example, which is common across all participating libraries. Information from the registry could be downloaded by a library and combined with locally held rights information, either in an ERM or in a simpler form such as spreadsheets. Alternatively, the registry could create a report specifically for the library based on information supplied, for a fee, on a service bureau basis. However, the apparent gaps and ambiguities in the historical record in relation to the archival rights of particular institutions cannot be remedied by a registry, but only by negotiation, with the results recorded in the registry.

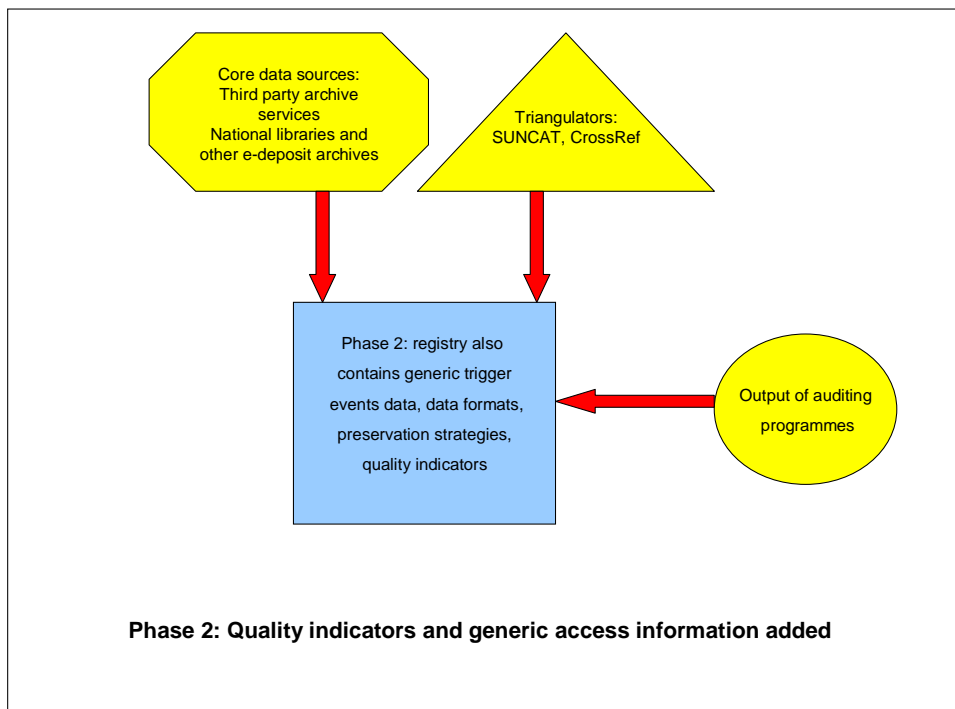
Perhaps eventually there would be a way to integrate these information sources completely. The most interesting suggestion so far has been a lightweight registry which integrates with several other sources/services to provide a fuller solution involving Onix for Licensing Terms (specifically Onix for Publications Licenses or Onix-PL) expressions, CrossRef (especially the journal transfer project), the Ringgold institutional identifier, and the ISSN. But Onix-PL is not a 'magic bullet': the licence information applying in the given situation must be known and the relevant part of the licence converted into an appropriate formal expression. At the moment, Onix-PL is still in development (though a great deal of progress has been made) and the work of actually encoding the relevant information from licences would be a daunting task. However, it is clear that there is demand from the community for ways of automating current licence information and displaying it to users (which is being explored in the current RELI pilot project) and also for managing the history of licences in a more efficient way than the present files and spreadsheets. Again, this is not something which would be sensibly attempted purely for the purposes of this registry, but as part of the overall attempt to make managing e-resources more efficient and less time-consuming through shared effort and infrastructure.

For any phase of this model to work at all there will be a need for auditing tools and triangulation of information. Some of the information could be fed directly to the registry – for example by the third party archive services or by CrossRef. Other information might be regularly harvested from sources complying with OAI-PMH. It would be extremely unrealistic in view of the diverse nature of the information sources, the sheer volume of journal titles and the movements that regularly occur such as changes of title and publisher, to expect that this process is going to be easy or seamless.. Each phase of the registry's implementation would require the definition of a data set, and investigation of the standards by which information would be communicated between entities. The phases are illustrated in the following diagrams:

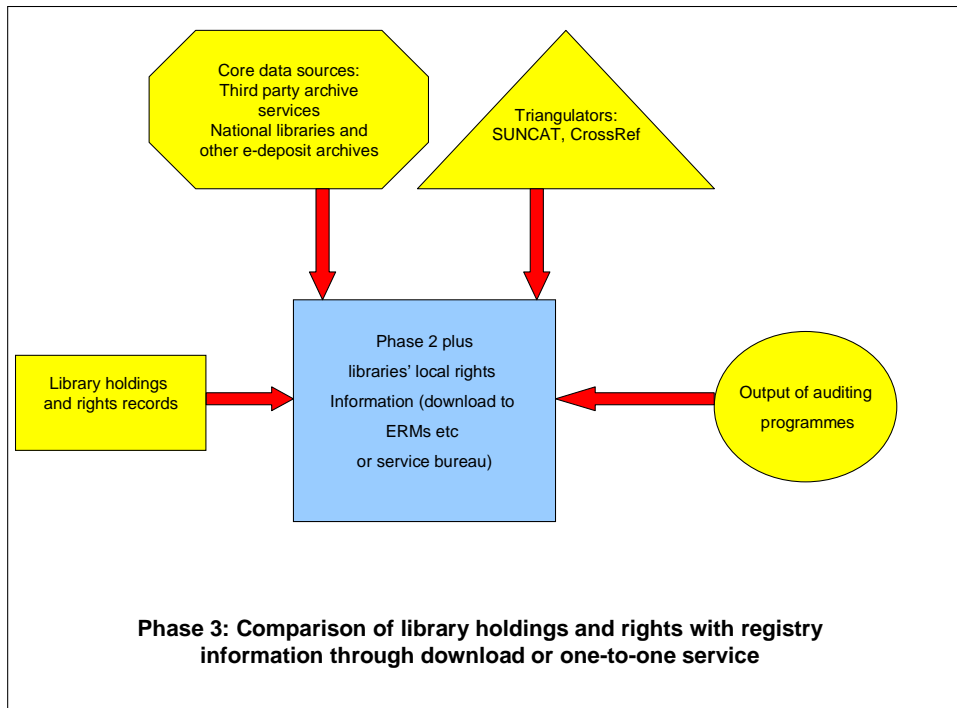
Phase 1



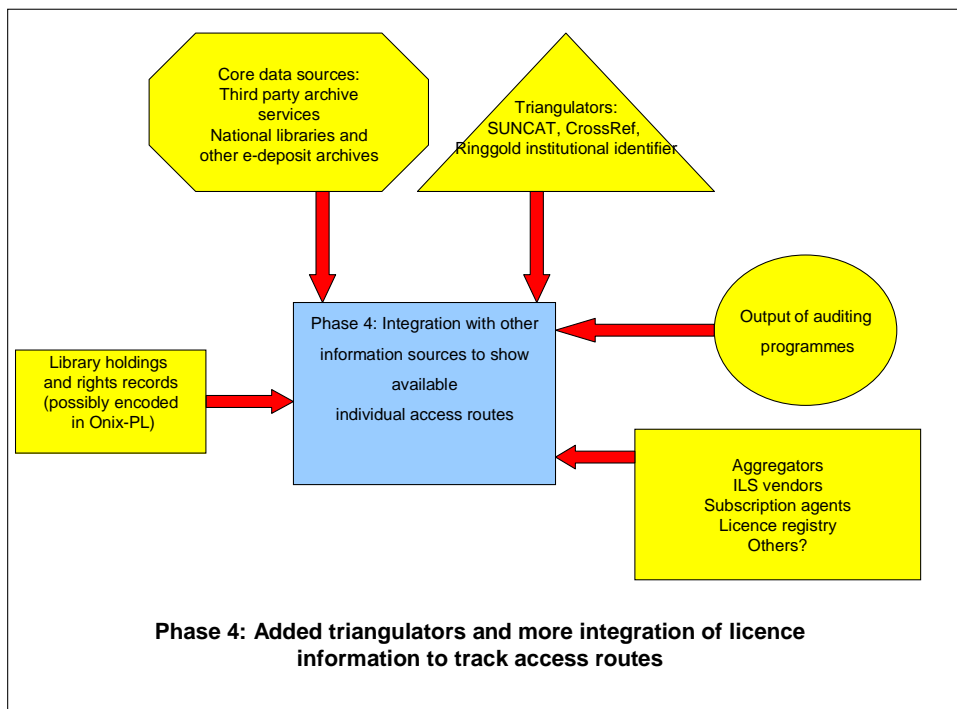
Phase 2



Phase 3



Phase 4



Registry models – some considerations

There are some generic issues concerning registries which need to be taken into account when considering how to establish and operate this registry:

- Motivation/leverage for participation is absolutely critical or the registry will be trapped in a vicious circle of poor data and low usage. Some registries rely on incentives e.g. in the case of EROMM, participation by at least one library in a country gives free access to all of them; some registries rely on the force of law, some on the desire to maintain a momentum behind a movement and be publicly identified as part of a movement (ROAR). Some factors for motivating participation in this registry have already been mentioned, from the use of leverage (making information supply a licence condition) to moral pressure (exposure of having made no satisfactory archiving arrangements) to positive reputation enhancement (being seen publicly to be acting in the best interests of customers);
- Currency of information: if the information in the registry requires updating, then there needs to be a regular cycle of information submission, and some kind of pressure (if only moral) applied to ensure compliance or harvesting and validation/auditing arrangements which are equally satisfactory. In the case of 'campaigning' registries, it is important for them to be recording upward trends in participation, so while accuracy may not be critical, they can still wither if they don't ensure they are visible
- Barriers to participation must not be too high but high enough to ensure quality otherwise there will be a major effort needed to clean up the information that is supplied
- Sustainability, including costs and revenues, usage and perceived value. Registries can be set up with project funding, but to be of any use – especially if they are concerned with persistent digital objects and preservation – they must have sustainable funding models. This does not necessarily imply a revenue stream other than public funding, but if a registry is to rely permanently on public funding, then it has to demonstrate usage and usefulness. It could be argued that some level of charge on the customer base would give a clear indication of whether the service is indeed proving useful or not, and therefore provide evidence for the case for public funding to continue. On the other hand, customers may be reluctant in this case to pay now for something that they perceive as having only long term relevance.
- Context: registries are part of a general information ecology that is evolving; consideration needs to be given both to interoperability and how potentially the registry fits in with other systems and services employed by the target user group, for example, facilitating information feeds into the registry, and allowing users to download information from it to go into other systems.

Timing

Some interviewees expressed the view that establishing a registry is premature, as the main issue facing the community is the primary one of ensuring that archiving and preservation is taking place. This encompasses not only the question of whether all publishers are taking the necessary steps to ensure access to their content in the long term, but also the technical hurdles facing the archiving organisations and their own long term sustainability. For example, CLOCKSS recently stated that:

The CLOCKSS initiative is funded by participating publishers and library organizations, as well as by a grant from the National Digital Information Infrastructure and Preservation Program (NDIIPP) via the U.S. Library of Congress. The grant is intended to finance CLOCKSS through a mixture of ingest fees from publishers and revenue from an endowment raised from voluntary contributions over the next five years. The need to secure long-term sustainable funding for CLOCKSS will be one of the key strategic issues facing the Board in 2008.

It is therefore worth setting out the pros and cons of establishing a registry now rather than later.

Now

- A registry will reveal the gaps that exist in provision which will be valuable in itself
- This could put pressure on publishers who are not taking satisfactory measures to secure access to their e-journals in the long term
- It could spur the archiving organisations to be more transparent in the information they make available as to their holdings, their processes and policies (filling the gaps identified by the CLIR report)
- It could take a long time to build consensus for an international registry, so the sooner the process starts the better

Later

- Waiting to establish a registry will allow the main focus of activity to be on the expansion and sustainability of the archive initiatives themselves
- The data available from the archive organisations will be more extensive and of a higher quality later; some archiving organisations are still in their pilot stages
- More e-deposit laws will be in place and there will be more practical experience of their operation
- Onix for Publications Licenses will be more fully developed and deployed
- There are no *particular* factors which dictate urgency in establishing the registry

The suggestion we have made for a pilot phase could prove useful in bridging the gap in views over the timing of a registry initiative.

Sustainability

Costs

Here we outline a very broad cost estimate for setting up and operating a registry over three years. Note that the annual costs do not relate directly to the Phases proposed above as it is beyond the scope of this report to undertake the detailed estimating that would be required for such a mapping to be accurate. Instead, we assume here that the cost of developing the additional content for the different phases is spread across the three years, and any technical development required to move to Phase 3 and 4 is implemented in Year 2 (the main technical infrastructure being laid down in the Setup period).

Year 1 is where most effort is required to ensure that the registry achieves a critical mass of information and usage.

Roles required

The main roles identified are:

- IT
- Registry management
- Promotion
- Project management

Although some roles could be carried out by the same person, they are sufficiently differentiated to be considered separately.

The first and last are self-explanatory; the second and third may need some explanation.

Registry management

Development of the registry requires expert decision-making about issues such as information architecture, metadata, standards and communications protocols that goes beyond the technical development level. The registry manager would also act as the intellectual leader of the registry project, and be responsible for strategic and policy aspects of its development. The registry manager will also co-ordinate an expert or advisory group (any standing governance body would be run by the project manager). The role has been costed as half-time as it is unlikely that anyone equipped to carry it out would be interested in undertaking it for anything less. However, it might be possible to turn it into a full-time role by combining it with either the promotion role, or more likely the project role (but not both).

Promotion and administration

Although the registry manager would be responsible for much of the high-level communication work necessary to develop the registry, a more focused role is also required to maintain constant contact with the organisations and institutions that are the sources of much of its content, and also with the users and potential users. This would include setting up, maintaining and managing e-mail listservers, running a website for the project with regularly updated lists of content and sources, details of access conditions and trigger events etc.

Table 2: Cost estimates

	Capital – hardware (see note 1)	Capital – software (see Note 2)	Services costs (see Note 3)	Personnel – IT (see Note 4)	Personnel - registry management (see Note 5)	Personnel - promotion & administration (see Note 6)	Personnel - project management (see Note 7)	Total	Comment
Setup		£2,000	£1,000	£12,500	£8,750	£10,000	£5,250	£39,500	IT development, setting policies and strategies, extensive promotion to organisations
Operations - year 1	£3,333		£1,500	£2,500	£38,500	£22,000	£2,800	£70,633	1/2 time registry management; 1/2 time promotion to libraries and publishers; IT maintenance
Operations - year 2	£3,333	750	£1,500	£6,250	£38,500	£10,000	£2,800	£63,133	Need for promotion development additional development to add functionality required for later phases
Operations - year 3	£3,333		£1,500	£2,500	£38,500	£10,000	£3,500	£59,333	
	Note 1	Hardware costs are amortised over 3 years							
	Note 2	Software costs are written off in year incurred							
	Note 3	Service costs are for bandwidth, IT facilities etc							
	Note 4	Assumed cost/day		£250					
	Note 5	Assumed cost/day		£350					
	Note 6	Assumed cost/day		£200					
	Note 7	Assumed cost/day		£350					

Funding

There were a variety of suggestions about how to fund the registry, an issue which is clearly critical.

- Fees could be paid by the publishers; however, the objection was raised that this would simply be passed on to libraries by big publishers and contribute to current financial problems for small ones
- A flat annual fee from HE institutions of no more than £2000
- JISC alone
- JISC could initially fund it, then invite others in as well e.g. European consortia
- Sponsorship e.g. 'This registry is brought to you by (insert 'large journal publisher' or even 'Google Scholar')
- JISC Collections initially funds the registry, and then gets affiliate fees for extending it to other countries or selling/renting the infrastructure
- It could be run as a subscription service by a commercial entity such as a subscription agent or ILS vendor
- Fees could be paid for additional services by the registry, which librarians could carry out themselves for free but choose not to

Some of these funding sources are mutually exclusive, but others are not e.g. sponsorship could co-exist with public funding.

Conclusions

1. There is an information gap that needs filling
2. There is broad support for a registry of some kind, though with significant dissenting voices, especially among the archiving organisations
3. Some of the dissent concerns the timing of such an initiative, and we have set out some of the factors relating to that above
4. We have not encountered anything which indicates that a registry would not be feasible. There are however a number of barriers to achieving it
 - a. Costs and resources
 - b. Consensus building about what exactly its scope should be
 - c. Consensus building about its national or international ambit
 - d. Consensus building about its governance
 - e. There will be different levels of barrier to address depending on how ambitious its scope is and what level of risk it is designed to mitigate (see table 1)
5. The main structural question to be addressed is whether the registry is a central service with clear governance or a set of functions distributed among different organisations. It may begin as the latter and then evolve into the former, or remain as a 'bricolage' of stitched-together elements. The downside of that is clearly that information could fall through the gaps, but the advantage is that it builds on what already exists

6. The user community definitely wants a service which will not simply give them a level of reassurance that a particular title is being archived, but will provide detailed information about the conditions under which it can be accessed **by them**
7. That suggests either a complex registry solution or, more practically, a simpler registry existing alongside complementary services.
8. The danger of just building a lightweight registry initially, even if there is provision to grow it later, is that expectations in the library community will be raised and then disappointed, leading to a lack of usage. This could be offset if the registry was created as an extension of a well-known and well-used resource such as CrossRef or SUNCAT
9. Discussions about establishing a registry could perhaps be inhibited by the current environment of mutual suspicion and mistrust between publishers and the library community; however, this was not evident in the particular interviews we carried out with publishers or librarians, and there are powerful mutual interests in improving information flows about e-journals
10. In our view, there will sooner or later be pressure to expand the scope of the registry to encompass not just e-journals but e-books, and probably other outputs of e-science. As scholarly outputs become more natively electronic there will inevitably be more blurring of the edges between books and journals, which have been defined by their printed forms, but are increasingly breaking free of them. It may quite quickly come to seem strange to confine the registry to e-journals
11. Sustainability is a critical issue: this is only worth doing if it can be sustained in the long-term

Table 3: Barriers and benefits

	Organisational barrier	Technical barrier	Other barriers	Benefits
Pilot	Co-operation of archiving organisations	Data quality	Cost of one-off exercise	Improved knowledge of overall provision of archiving arrangements; pressure on publishers who are making no provision
Phase 1	+agreement to move to registry proper; scope and governance arrangements	+updating mechanisms (communication, interoperability)	+Ongoing funding to be agreed for registry development and maintenance	+libraries more aware of balance of overall risk in cancelling print
Phase 2	+ acceptance of quality indicators by all stakeholders	+availability of output of auditing processes		+libraries more aware of particular risks associated with particular archiving solutions
Phase 3	+Relationships with libraries	+gaps and ambiguities in the licence history records		+libraries able to understand access routes

Phase 4	+involvement of other organisations in the supply chain	+availability of Onix-PL	+effort and cost of encoding licence information	+libraries more able to take decisions about e-only
---------	---	--------------------------	--	---

Recommendations

1. The JISC needs to decide whether doing something, even if initially, (or perhaps eventually), it falls short of what librarians feel they need, is better than doing nothing
2. Our view would be that it is certainly worth carrying out the pilot stage leading to an 'endangered species' list, which would contribute to the overall progress towards preservation as well as providing concrete experience in obtaining the data
3. Detailed discussions with the archiving organisations listed in the CLIR report would be a logical next step before embarking on the pilot stage
4. Following the learning and mapping process of the pilot, it would then be possible to have a more informed view of the next stage
5. If the JISC does decide to move ahead with an actual registry or registry-like service, the next major decision would be whether to conceive of it as a 'virtual' entity, whose function is to act as a framework linking a number of different elements together or as a single identifiable body
6. Consultations with organisations which have been suggested to us as having an important role to play such as CrossRef, Ringgold, EDItEUR, SUNCAT, OCLC and some of the key ERM vendors would illuminate the question of how best to implement a registry or registry-like service
7. The JISC would also need to decide whether its primary focus is to serve the UK research community and let international partners join in if they wish, or whether to conceive of it as a potentially global solution, which would mean getting international involvement from the beginning. Clearly that is also a question affected by decisions on how to implement it, for example, if CrossRef were to be the host and operator, it would naturally be an international solution. It seems to us that in order to gain the co-operation of the archiving organisations based around the world, which would be vital to its utility, the registry/registry-like service would have to be conceived as something which would serve the whole international scholarly community

Appendix A

List of participating organisations (interviews and/or workshop)

Archiving organisations/national libraries

The British Library
CISTI
CLOCKSS/LOCKSS
Kopal/DDB
KB e-Depot
National Library of Scotland
NDIIPP (Library of Congress)
OCLC
Portico
UK PubMed Central

Publishers/publishers' organisations

ALPSP
Cambridge University Press
Elsevier
Oxford University Press

University Libraries

The Bodleian Library
The Open University Library
The University of East Anglia Library
The University of Sheffield Library

Other stakeholders

CrossRef
DPC
JISC Collections
Ontario Scholars Portal
Research Information Network
UKSG