Piloting an e-journals preservation registry service (PEPRS)

PETER BURNHILL
Director
EDINA National Academic Data Centre

CO-AUTHORS:
FRANÇOISE PELLE (Director) and **PIERRE GODEFROY**, ISSN International Centre;
FRED GUY, **MORAG MACGREGOR**, **CHRISTINE REES** and **ADAM RUSBRIDGE**, EDINA

The need for a registry of archived scholarly publications features in various reports on digital preservation and archiving of e-journals. Online access to scholarly journals brings many benefits. However, there are genuine concerns about long-term access, given threats from natural disaster, human folly and technological failure. Fortunately these concerns are being recognized and a number of organizations are stepping forward as archiving agencies, although as yet with no readily accessible means to discover who is looking after what, to what purpose and with what terms of availability.

This paper describes a project to pilot an e-journals preservation registry that is being carried out by a UK national academic data centre and the international standards body for serials. Funded by JISC for the UK, various open issues include whether this should be regarded as an international initiative and whether it should help address related concerns in the library community for information on post-cancellation access.

Introduction

First, the good news: the shift to e-journals means that researchers and students in UK universities have access to many of the journal articles they need, anywhere, anytime, to read online or to download for later use. Notwithstanding inequities of access due to economic disparity, the pervasiveness of the web makes plain the extent to which this is global, for both access and publishing, even if subscriptions and much digital library activity are institutional and national.

Next, the bad news: the worry that what is now in digital form may not always be available. Digital archiving of e-journal content has become a topic for policy agenda across all sectors in the global serials industry, as well as in research and academic funding bodies nationally. First, there are initiatives that focus on digital preservation, risks associated with computing failure (bit rot, format obsolescence), natural disaster (earthquake, flood, fire) as well as human foibles (business decision-making, financial loss, criminal/political action, operator error). Second, there is now focus on other aspects of continuity of access, with libraries conscious that worsened finances may force them to cancel subscriptions, and that their role as service provider for back issues of journal content is undermined by their dependence on having to rent online access.

In the UK, JISC, acting on behalf of the funding councils for higher education, and other organizations, such as the Digital Preservation Coalition (DPC)[1] and the Digital Curation Centre, have emerged as advocates for action, commissioning or generating studies, reports, guidelines and tools. A review carried out by Maggie Jones in 2002/3 identified many of the salient issues[2]. The UK serials community is not alone in having to confront this bad news; every academic policy community internationally faces comparable challenges. In the US, Kenney et al in the CLIR Report[3] in 2006, reviewed the emergence of agencies willing to take

on digital preservation responsibilities. In turn, JISC commissioned an analysis of the CLIR Report and its implications for the UK, in another review by Maggie Jones in 2007[4].

In a more recent review from 2008, Terry Morrow et al[5] examined the policies and practices of six of the agencies having particular relevance to the UK noting differences in purpose. These six included LOCKSS[6] (Lots Of Copies Keep Stuff Safe), CLOCKSS[7] (Controlled LOCKSS), Portico[8], the e-Depot[9] (from the Dutch national library, Koninklijke Bibliotheek, KB), OCLC Electronic Collections Online (ECO)[10], and the British Library e-Journal Archiving Programme. They were selected partly on the basis of seven indicators of viability used within the CLIR Report[11]. The mission of some agencies, like the CLOCKSS collaboration begun initially by six 'steward libraries' and 12 of the largest publishers, is long-term preservation of the scholarly record; for some others, their main emphasis is on 'perpetual access' – this latter term encompassing both back-up for short-term failure or continuity of access to back content in an e-journal after the cancellation of a current subscription ('post-cancellation access').

Both Jones[12] and Kenney et al[13] had stressed the need to know who is doing what, for whom and under what terms of access, either in terms of clarity of public statement by each agency or through a registry by which it would be plain what content was being archived, and therefore what was not. In her review for JISC of the CLIR Report recommendations, Jones[14] commented that a registry "would be an incredibly valuable resource … However, it would also be a difficult task, as the report points out". She noted that it seemed an appropriate collaborative task and looked to the British Library, JISC and the DPC to discuss how to take it forward, noting models for registries for open access repositories.

Scoping a digital preservation registry for e-journals

The JISC response was to commission a 'scoping study for a registry of electronic journals that indicates where they are archived'. The authors of that study, Sparks et al[15] interviewed a range of stakeholders in the UK, including representatives from national and university libraries, publishers and archiving organizations. On the central question,

the study concluded that there was need for readily accessible information about whether and where e-journals were archived. However, differences of view emerged in the interviews and workshop about the context in which such a network-level registry should sit, and about matters related to timing, implementation and sustainability. The authors commented, "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"[16].

On the matter of organization, there was discussion of the relative merits of a centralized service versus a collection of distributed services, concluding that "the function of the registry should be attached to something else that already exists in order to leverage existing organizations and infrastructure"[17]. A pilot stage was recommended, followed by phased development of the registry using SUNCAT, the UK serials union catalogue[18], as a potential master list against which to compare the current and planned 'holdings' of archiving services, noting that SUNCAT included records from the ISSN Register that could be used to establish which 'titles' were covered by the various archiving services, and which were not.

A pilot project for a e-journals preservation registry service

In August 2008, JISC began funding a project to pilot an e-journals preservation registry service (PEPRS). This is being carried out by EDINA, the UK national academic data centre based at the University of Edinburgh, and the ISSN International Centre (ISSN-IC), building on a long-standing association.

The aim of this two-year project is to investigate, build and pilot an online facility, based upon an authoritative registry. Its purpose should be to enable librarians and policy-makers, principally in the UK but also worldwide, to ascertain the archival provision for an e-journal: which archiving scheme has been used, including access/release arrangements; which e-journals are not (yet) within an archiving scheme.

The requirements are that it should be accurate, up to date, comprehensive; possible to allow other

serials lists and serial union catalogues to cross-reference registry entries in order to identify 'endangered' e-journals; free at the point of use ('freely-available'). A review is scheduled in February 2010 to decide whether to support the preservation registry as a production service.

The project is under way but still at the design and build stage. Contact has been made with three types of digital preservation agency: network-level organizations; national libraries; and library co-operatives. Thus far the following have indicated willingness to work with the project as candidate archiving agencies: CLOCKSS and Portico; British Library and e-Depot; UK LOCKSS Alliance[19]. Co-operation with OCLC[20] is also being established.

Preliminary design considerations

A registry must have certain desirable qualities as well as functionality if it is to be effective and to command respect. Review of work already cited has helped define a list of those qualities. These include being comprehensive in coverage, accurate (with due provenance) and up to date, in terms of relevant e-journals and of information about the e-journal content under stewardship of archiving services.

The general approach in the project has been to devise an architecture that can deliver these qualities for a set of services based on the registry. It also needs to use standards-based interoperability in order to support machine-to-machine use as well as a web-based user interface. This suggests a modular architecture with respect to function and authoritative sources of information. The exact form of network interoperability (search or harvest/push or pull) has yet to be determined.

The registry could be thought of as a matrix that correlates preservation information against known e-journals. A key feature in the abstract data model, illustrated in Figure 1, is the explicit dependence upon two remote external information sources: metadata on e-journals, the rows of the matrix; metadata on archival action for each serial title is generated from the archiving agencies. Use of persistent and internationally accepted identifiers, both for e-journals and for archiving agencies would clearly benefit the registry. As seen in Figure 1, and discussed below, the metadata on e-journals is sourced from the ISSN Register. Thus far there is no registration or assignment of identifiers for archiving agencies.

It is envisaged that the registry might maintain a minimal record against each known e-journal title, indicating action (or not) by each registered

Piloting an E-journals Preservation Registry Service
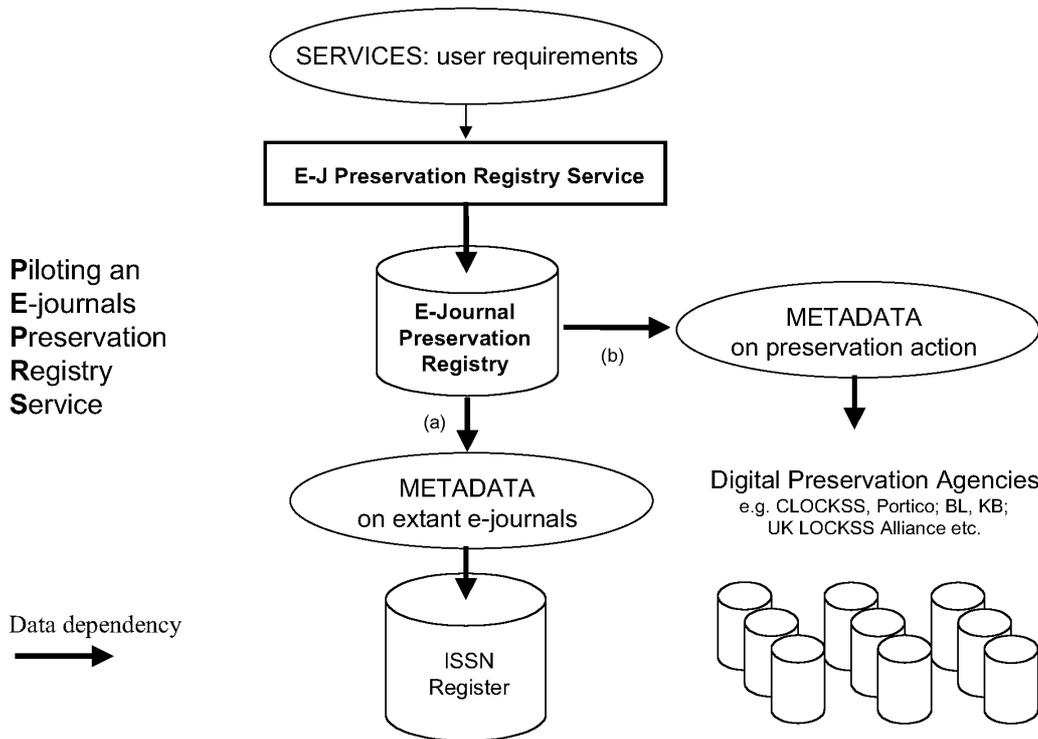
Data dependency

Figure 1. Abstract data model

preservation agency and providing telematic link, typically in the form of a URL. The intention is that this would support both up-to-date report by preservation agencies and also keep an historic record of the statements made. Issues to do with choice of fields and vocabulary are discussed below. Verifying specific fields of information is the responsibility of the source best placed to deliver up-to-date information. Technologically, this exploits the 'always on' presumption of the Internet, whilst also using cache technology to manage risk of failure. Organizationally, this makes plain that the registry is not itself an audit authority.

It is recognized from the outset that archiving organizations are presently striving to make themselves financially viable. An important part of the project therefore is to identify and propose technical and organizational means of reducing burdens, and thereby costs, whilst satisfying the overall aim for the online facility.

On the assumption that the operation is low cost, it is also envisaged that the basics of the registry services should be free-at-the-point-of-use, either globally or restricted to members of the UK academic and research community. This latter issue is one of several 'open issues', and is for JISC to address if it decides to support the registry service, requiring some form of authentication/authorization control unless the registry is regarded as its contribution to global infrastructure for its externality value.

The partnership of EDINA and ISSN-IC lends an e-journals preservation registry some useful qualities, directly or in prospect. Clearly, the ISSN-IC is the international standards agency for serials, and location of the registry in a national data centre (EDINA) assists evaluation of qualities for a well-connected and well-maintained platform providing 24/7 access against a variable load, both for web access and for machine-to-machine interoperability.

Serial identification and the challenge of enumerating extent of serial content

A registry benefits from use of persistent and internationally accepted identifiers: use of the ISSN and the ISSN Register is therefore a key feature, together with inclusion of the ISSN-IC as partner in the PEPRS project. The ISSN Network manages the international standard numbering system for serials, of which serials in electronic format are a proper subset.

Through co-operation with publishers and intermediaries and special efforts by the ISSN-IC and the ISSN Network, ISSN records now exist for over 60,000 e-journals (strictly, serials in electronic format), with more in prospect. This is a critical mass of serial titles for project purposes and is the basis for a successful production service, surely representing the most comprehensive record of the world's scholarly journals published in electronic format, both subscription and open access journals – the *Directory of Open Access Journals* (DOAJ) requires ISSN assignment. Put another way, it could be argued that any e-journal worth preserving ought to have an ISSN.

The scope of first test implementation will be limited to serials with an ISSN but this could not be the case for a production service. For the ISSN-IC, one of the positive outcomes of the project could be the opportunity to improve the coverage of the ISSN Register.

Separate ISSNs are assigned to the print and electronic versions of a title; these versions are now being associated through a recently introduced Linking ISSN (ISSN-L)[21]. A workflow has to be devised to manage several possibilities that could occur with archived content. The most obvious instance is where an ISSN has not yet been assigned for an e-journal title being archived, or an ISSN exists only for the print version. The intention is that the workflow would result in the assignment of the 'e-ISSN' for use in the e-journal preservation registry (see Figure 1) and hence by the archiving agency. There is also the instance where an ISSN has been assigned to the e-serial, but the archiving agency only had knowledge of the ISSN for the print edition. Possible use of the ISSN-L in the project is being considered. One open issue which is to be addressed is how to take note in the registry of a 'digitized journal', the digital surrogate of a serial which was published in print format. Ordinarily that would not have been assigned an 'e-ISSN'; as that will often be for older serial content, an ISSN may not even have been assigned to the print format.

Librarians will want to know the extent of preserved content for a given title: the article that is the ultimate information object of desire from the researcher's point of view. Determining the appropriate metadata on serial content is a problem that has been relegated to the second phase of the PEPRS project. Provisional thinking is to create four date fields for each e-journal: earliest and

latest known date of issue in digital format; earliest and latest known date of issue archived. It may be possible to do better, but this type of information is problematic without known and available publication pattern for each title. There is likely a role here for co-operation with others on this, such as Onix for Serials[22] and CrossRef[23].

Developing a shared vocabulary for archival action

The main areas of policy interest are being established as part of the review of user requirements for the registry service (Figure 1). The term 'archiving action' signals another open issue for PEPRS: should the scope be limited to digital preservation, or should the scope be widened to take note of growing concern among libraries and their patrons about 'perpetual access'? Regardless, and sticking to the original brief to focus on digital preservation, it is clearly important to be able to make plain the form of archival action being taken by a given agency and the corresponding terms of access or release of that content: for example, whether this is long-term preservation with triggered general ('open access') release, or whether this is post-cancellation' access dependent upon prior subscription.

These policy areas need to be resolved into defined fields of metadata. The next step is the development of a common vocabulary for entries to assist users of the registry service who will want to compare attributes of preservation actions. This vocabulary needs to be sensible for users of the registry service but must also be sensible for the archival agencies as the system depends upon their self-statement of policies and the like, including terms of access to the content archived.

Concluding remarks

Project work is under way to provide an online facility that aims to bridge the information gap about preservation, and perhaps other archival action on e-journal content. This is our first full report to the UK serials community of what we intend – in much the same way that SUNCAT was first reported[24]. Indeed, this is one of a set of strategic initiatives being taken in and around SUNCAT, as the UK serials community begins

to exploit the value of a national union serials catalogue, which typically has three sets of serial level information: identification; bibliographic description; holdings and location. There are other sets of information to add, or to link to, at serial title level, including preservation status. This is also one of a set of strategic initiatives for the ISSN-IC. More generally, there is agreement about the need to link to up-to-date information about the current status of a serial, including changes of publisher: that too could be held and accessed as authoritative network-level information.

A number of open issues have been identified in establishing a registry service for e-journal digital preservation. Two open issues of major significance are to do with scope. The first of these is the international character of the proposed register and registry service. The implicit focus, given funding by a UK funding source, is to serve the needs of the UK research community. But it is clear that both the e-journals of interest and archival action extend beyond the UK and any one national boundary. International involvement seemed sensible from the outset. As the authors of the scoping study noted, "It seems to us that in order to gain the co-operation of the archiving organizations 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"[25].

The second is whether and how to extend the scope of the proposed registry service to include librarians' concerns about continuity of access, both of open access materials and of post-cancellation access to subscribed content, in the information on archival action. It could be that such information is for the archival agencies to report upon beyond the scope of this registry. The complication comes with maintaining or linking to a record not only of current subscription status, but of past subscription status, combined with the fact that subscriptions are most often determined by national rather than international contract. This suggests that while the PEPRS project might aim to be international for information on preservation, it should either limit its role for subscribed content to the UK only, working to support NESLi licensing[26], for example, or else make its functionality available to third parties who can meet the needs of other countries more appropriately.

There is also an additional set of open issues to do with sustainability of the registry and its services.

The scoping study had observed, "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"[27]. Discussion of that set of issues here would be premature, but these will have to be addressed before February 2010 when PEPRS will be reviewed by JISC to assess whether and how to make the transition into a production service according to some agreed business plan and form of governance – ensuring and maintaining trust across the serials community. That said, this seems a suitable open issue upon which to end.

References

1. Digital Preservation Coalition (DPC) http://www.dpconline.org (Accessed 11 February 2009)

2. Jones, M, 'Archiving E-Journals Consultancy (Final Report)', October 2003; available from http://www.jisc.ac.uk/uploaded_documents/ ejournalsfinal.pdf (Accessed 30 January 2009)

3. Kenney, A R, Entlich, E, Hirtle, P B, McGovern, N Y and Buckley, E L, 'E-Journal Archiving Metes and Bounds: A Survey of the Landscape', September 2006; available from http://www.clir.org/PUBS/abstract/pub138abst. html (Accessed 30 January 2009);

4. Jones, M, E-Journal Archiving (Review and Analysis of the CLIR Report *E-Journal Archiving Metes and Bounds: A Survey of the Landscape*), February 2007; available from http://www.jisc.ac.uk/whatwedo/programmes/ preservation/ejournalarchiving.aspx (Accessed 30 January 2009);

5. Morrow, T, Beagrie, N, Jones, M and Chruszcz, J, Tee Em Consulting, 'A Comparative Study of e-Journal Archiving Solutions', May 2008; available from http://www.jisc-collections.ac.uk/news_and_ events/news_articles/ejournals_pr.aspx. (Accessed 9 February 2009)

6. http://www.lockss.org/ (Accessed 11 February 2009)

7. http://www.clockss.org/ (Accessed 11 February 2009)

8. http://www.portico.org/ (Accessed 11 February 2009)

9. http://www.kb.nl/dnp/e-depot/e-depot-en.html (Accessed 11 February 2009)

10. http://www.oclc.org/electroniccollections/ (Accessed 11 February 2009)

11. Kenney, A R, ref. 3

12. Jones, M, ref. 2

13. Kenney, A R, ref.3

14. Jones, M, ref.4

15. Sparks, S, Look, H, Muir, A and Bide, M, 'Scoping study for a registry of electronic journals that indicates where they are archived', Rightscom and Loughborough University, 2008; available from http://www.jisc.ac.uk/media/documents/ programmes/preservation/ejournalregstudy.pdf (Accessed 30 January 2009)

16. Sparks, S et al, ref.15, p.3

17. Sparks, S et al, ref.15, p.17

18. http://www.suncat.ac.uk (Accessed 11 February 2009)

19. http://edina.ac.uk/lockss/ (Accessed 11 February 2009)

20. http://www.oclc.org (Accessed 11 February 2009)

21. See the ISSN website: http://www.issn.org/2-22637-What-is-an-ISSN-L.php (Accessed 11 February 2009)

22. http://www.editeur.org/onixserials.html (Accessed 11 February 2009)

23. http://www.crossref.org/ (Accessed 11 February 2009)

24. Burnhill, P, Halliday, L, Rozenfeld, S and Kidd, T, 'SUNCAT: a modern serials union catalogue for the UK', Serials, 2004, 17(1), 61-67; available from http://www.metapress.com/content/ c6y0ltjtxlhrgfr9/ (Accessed on January 30, 2009)

25. Sparks, S et al, ref.15

26. The NESLi licensing scheme is managed by JISC Collections http://www.nesli2.ac.uk/model.htm (Accessed 11 February 2009)

27. Sparks, S et al, ref.15, p.3

■ **Peter Burnhill**
Director
EDINA National Academic Data Centre
Causewayside House
University of Edinburgh
160 Causewayside
Edinburgh EH9 1PR
Scotland, UK
Tel: +44 (0) 131 650 3301; fax: 3308
E-mail: p.burnhill@ed.ac.uk
URL http://edina.ac.uk

■ **Francoise Pelle**
Directeur
ISSN International Centre
20, rue Bachaumont
75002 Paris, France
Tel : +33 1 44 88 22 20; fax : +33 1 40 26 32 43
E-mail: pelle@issn.org
URL http://www.issn.org

To view the original copy of this article, published in *Serials*, the journal of the UKSG, click here:

http://serials.uksg.org/openurl.asp?genre=article&issn=0953-0460&volume=22&issue=1&spage=53

The DOI for this article is 10.1629/2253. Click here to access via DOI:

http://dx.doi.org/10.1629/2253

For a link to the table of contents for the issue of *Serials* in which this article first appeared, click here:

http://serials.uksg.org/openurl.asp?genre=issue&issn=0953-0460&volume=22&issue=1