Tales from The Keepers Registry
Helping to ensure ease & continuity of access to digital resources

Peter Burnhill
EDINA (JISC & University of Edinburgh)

Digital Future and You: Library of Congress, Washington DC
10 December, 2012
Three Stories To Tell & Pictures To Paint

(based on an invited paper being accepted for *Serials Review*)

1. The Keepers Registry (to monitor progress)

2. A Large Volume of Serial Issues

3. The Scholarly Web (the now & the future then)

+ some asides on the way: eg, ‘seriality’
Monitoring Preservation of (Digital) Serial Content

Helping to ensure continuity of access

Identifying the stream

http://www.flickr.com/photos/shinez/5000985919/
Central Task for Libraries: To ensure researchers, students and their teachers have ease and continuing access to online scholarly resources.

- **Ease**: usability
- **Access**:
  - restricted
  - open
- **Continuing**:
  - access to back content
  - long-term preservation

The Keepers Registry as monitor upon e-journal archiving.
The Basics of Our Shared Concern

[The Good News] 😊

What was once availably locally <on shelf, in drawer/datacentre> is now online & accessed remotely, anytime/anywhere

- Studies show that scholarly literature is now nearly all online
- Libraries moving to an e-only environment for journals

[The Bad News] ☹️

The role of libraries as trusted keepers of information and culture has been disrupted

✧ Libraries no longer take physical custody of digital
  - Publishers license content online remotely

✧ Risk of loss for future scholars, citizens & our children
Our Shared Understanding

✧ World heritage & scientific understanding is global

✧ Scholarship & science has global literature
  ✧ Researchers in any one country are dependent upon content written and published in other countries

✧ International/National Reports & Activity: 10 Years On
  ✧ Archiving E-Journals (JISC: Maggie Jones, 2003)
  ✧ Archiving Electronic Journals (L. Cantara (Ed) DLF/CLR, 2003)

Increasingly born digital, or re-born digitized

Need to ensure continuity of access
The real heroes in the story are ... 

The Keepers, organisations that act as our digital shelves:

① web-scale not-for-profit organizations
   * e.g. CLOCKSS Archive & Portico

② national libraries
   * e.g. British Library, e-Depot (Netherlands) & National Science Library of China

③ library consortia
   * e.g. HathiTrust & Global LOCKSS Network
Sidebar note on National Libraries

Should we wait upon Legal Deposit?

- 94% of libraries have some form of legal deposit for print.

- Only 44% national libraries had legislation in 2011 for e-books or e-journals; expected to rise to 58% by June 2012.

- Only 27% [expected to rise to 37% by June 2012] actually ingesting via legal deposit

- Total national libraries collecting = those 14 via legal deposit + 9 by other means (Netherlands, UK/BL & Switzerland have voluntary deposit)

- 2 (KB e-Depot & BL) participate in the Keepers Registry
  - Only when the other 21 join will all know about their activity

from presentation, CENL 2011 Survey by Lynne Brindley to CDNL Annual Meeting Puerto Rico, 15/8/11
“Digital information is best preserved by replicating it at multiple archives run by autonomous organizations”

B. Cooper and H. Garcia-Molina (2002)
A Registry to discover ‘who is looking after what’

- Idea mooted in UK Report (JISC: Maggie Jones, 2003/4); Call in USA/Canada for “clarity of public statement by each agency or through a registry” (CLIR Report, 2006)

- UK scoping study recommended an e-journals preservation registry be built (JISC: Rightscom/U. of Loughborough 2007)

- JISC funded EDINA & ISSN-IC as partners to Pilot an E-Journal Preservation Registry Services (PEPRS)
  - Phase 1: August 2008 – July 2010
    ‘investigate, prototype and build’ [evaluation in Feb. 2010]
  - Phase 2: August 2010 – July 2012
    ‘preparing for service & governance’
SERVICES: user requirements

E-J Preservation Registry Service

E-Journal Preservation Registry

METADATA on extant e-journals

ISSN Register

METADATA on preservation action

Digital Preservation Agencies
e.g. CLOCKSS, Portico; BL, KB;
UK LOCKSS Alliance etc.

Piloting an E-journals Preservation Registry Service

Data dependency
Partners have 15+ years of association

- ISSN International Centre (ISSN IC, on behalf of Network)
  linked to national libraries and publishers
  - an intergovernmental institution governed by statutes/convention between UNESCO & France (as host country)
  - coordinates the ISSN Network of national centres, operating an automated system for the registration of serials, via assignment of International Standard Serial Numbers (ISSNs)
    http://www.issn.org

- EDINA linked to needs of research & teaching
  - part of The University of Edinburgh (Scotland, UK)
  - designated in 1995 to act as a national data centre by JISC, the ICT agency for UK universities and colleges
    http://edina.ac.uk
EDINA

EDINA’s Mission: We develop and deliver online services & digital infrastructure for UK research and education.

Our Vision: To be integral to the quality & productivity of research & education, in the UK & beyond

*part of The University of Edinburgh & the Jisc Family*
Extracts of various Letters, approving of the STATISTICAL ACCOUNT OF SCOTLAND, and urging a Perseverance in that Undertaking.


I CANNOT but express myself highly pleased with the undertaking in which you are engaged, (that of drawing up the Statistical Account of Scotland), and give my best wishes for its success. I am full persuaded, that when enlightened men, will take the trouble to examine so minutely into the state of society, as your inquiries seem to go, it must result in greatly ameliorating the condition of the people, promoting the interests of civil society, and the happiness of mankind at large. These are objects truly worthy the attention of a great mind, and every friend to the human race, must readily lend his aid towards their accomplishment.
No XX. Extract of a Letter from his Excellency John Adams, now President of the United States of America, dated Philadelphia, 2d March 1793.

I RECEIVED yours, with your plans for a natural history of sheep, and the Statistical Survey of Scotland. You could not have made a wiser choice. The natural history of that animal, so useful to man, must be extremely important, as well as extremely curious; and a detail of particulars, relative to your native country, must be interesting to all but especially to the inhabitants of it. We, in this country, enjoy a delicious tranquility at present, and if your European fermentations should not disturb us, shall continue to be happy.
EDINA’s Mission: We develop and deliver online services & digital infrastructure for UK research and education.

Our Vision: To be integral to the quality & productivity of research & education, in the UK & beyond

part of The University of Edinburgh & the Jisc Family

Jisc reformed to become a charity governed by UUK, GuildHE & AoC; JISC was an arm of Government funding

Previous JISC Mission: To provide world-class leadership in the innovative use of information and communication technology to support education, research & institutional effectiveness.

With Vision of easy & widespread access to information and resources, anytime, anywhere; a vision with technology and information management at the heart of research & education.
The Keepers Registry

- Initial scope was upon what is published in digital form
  - in serials and periodicals (journals, magazines, newspapers etc)

- Content of significance for each & every country:
  - publishers & archiving agencies are international

We can already report some good progress, having launched a Beta service a year ago, as noted on our blog, http://thekeepers.blogs.edina.ac.uk/
The Keepers Registry

Supporting long-term access to journal content

http://thekEEPERS.org

Discover who is looking after what e-journals

Search

Input a search term (e.g. free text, or a title or ISSN(s))

Search

Glossary and definitions

Current Statistics

- Serials reported as 'preserved': 18,402
- Serials where action is 'in progress': 11,819

We include a title count where at least one agency has taken the reported action on one or more volumes. See the FAQ for more information.

News and Events

17th July 2012: The Keepers Registry's first Pecha Kucha outing is a winner! [View]

16th July 2012: Your one minute starts now! [View]

June 2012: Draft Inclusion Criteria released for review. [View]

June 2012: ONIX for Preservation Holdings. [View]

FAQ for news on additional functionality that is planned at keepers.org

Learn more about the archiving agencies in The Keepers Registry who have stepped forward to take responsibility for ensuring long-term continuity of access to global scholarship.

- British Library (Last updated: 06 Sep 2012)
- CLOCKSS Archive (Last updated: 13 Aug 2012)
- e-Depot (Last updated: 24 Apr 2012)
- Global LOCKSS Network (Last updated: 14 Aug 2012)
- HathiTrust (Last updated: 01 Sep 2012)
- National Science Library, Chinese Academy of Sciences (Last)
- Portico (Last updated: 06 Aug 2012)

The most up-to-date information is available on the agency website. The date represents when the Registry was updated.

This is a Beta release from the PEPRS project. We are also exploring governance and seeking sponsorship for this global facility. Your feedback is welcome.

Supported by

JISC

Acknowledgements

ISSN

EDINA
Reminder about the real heroes in the story

The Keepers of e-journal content

And others, as they tell all via the Keepers Registry ...

http://thekeepers.blogs.edina.ac.uk/2012/06/27/draft-inclusion-criteria-released-for-review/
Sidebar note on monitoring their progress ...

- How should we measure progress?
  - 18,400 serial titles reported as being ‘preserved’

- How many e-serials are there?
  - ISSN Network has issued c.100,000 ISSN for online resources
    * 20% of those by ISSN-US (Lib.Congress);
  - c.30,000 refereed scholarly journals (SerialsSolutions)

- Cross-checking the e-serials (having ISSN) that a given university library cares about:

<table>
<thead>
<tr>
<th>University Library</th>
<th>% ‘Preserved’ by 1 or more</th>
<th>% ‘Preserved’ by 3 or more</th>
<th>% Not known as ‘Preserved’</th>
<th>Total having a valid ISSN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Columbia</td>
<td>26</td>
<td>12</td>
<td>74</td>
<td>58,882</td>
</tr>
<tr>
<td>Cornell</td>
<td>28</td>
<td>13</td>
<td>72</td>
<td>54,698</td>
</tr>
<tr>
<td>Duke</td>
<td>22</td>
<td>10</td>
<td>78</td>
<td>61,682</td>
</tr>
</tbody>
</table>

- But only partial coverage of issued content!
  - Volumes and Issues are missing from archival holdings
Example search for: *Origins of Life*

1 hit found
Current search: Origins of life (Key Words)

<table>
<thead>
<tr>
<th>Title</th>
<th>ISSN</th>
<th>Publisher</th>
<th>Current extent of archiving</th>
<th>Archiving agency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Origins of life and evolution of the biosphere</td>
<td>1573-0875 (Online); 0169-6149 (Print)</td>
<td>Springer</td>
<td>Preserved: v. 36, 40-41</td>
<td>CLOCKSS Archive</td>
</tr>
<tr>
<td>Kluwer Academic Publishers</td>
<td></td>
<td></td>
<td>Preserved: v. 27-37</td>
<td>e-Depot</td>
</tr>
<tr>
<td>Springer Netherlands</td>
<td></td>
<td></td>
<td>Preserved: v. 37-41</td>
<td>e-Depot</td>
</tr>
<tr>
<td>Springer Science+Business Media</td>
<td></td>
<td></td>
<td>In progress: v. 1-present</td>
<td>Global LOCKSS Network</td>
</tr>
<tr>
<td>Springer Science+Business Media</td>
<td></td>
<td></td>
<td>Preserved: v. 27-42</td>
<td>National Science Library, Chinese Academy of Sciences</td>
</tr>
<tr>
<td>Springer Science+Business Media</td>
<td></td>
<td></td>
<td>Preserved: v. 39, 41</td>
<td>Portico</td>
</tr>
</tbody>
</table>

... but coverage of volumes is partial & patchy

This e-journal is being archived by 5 archiving agencies ...
2. A Large Volume of Serial Issues

1. Identification

- focus when building the Registry was online content that had an ISSN, as project convenience
  - Incorrect, erroneous or missing
- We knew that different ISSN for digital and for print
  - But value of ISSN-L as kernel field to link across ISSN records
- Used HathiTrust to extending scope to include digitized journals
  - Rules for assigning ISSN to digitised content from print journals
  - same ISSN applies to all digital versions of an online resource
- Noted ISSN assignment rules for integrating resources (websites and online databases that change over time)
Sidebar note on Identifiers

• **Current Publisher Titles:**
  - Relatively low numbers of ISSN assignments for titles from India and China/Hong Kong, as reminder of number of ‘hidden’ e-journals & the like.
  - c.20% of 100,000 issued by US ISSN Centre; UK ISSN Centre assigned c.10%; centres in Netherlands, Germany & Brazil each assigned c. 4.5%. The ISSN IC assigns c.3.5%.

• **Dead Publisher Titles:**
  - Estimated to be 250,000 digitised serials in HathiTrust
  - Any ‘body’ (not just a publisher) can apply for ISSN
  - Responsibility for assigning ISSN is with the national ISSN Centre for the country that organises the digitisation
1. Identification

2. Recording the extent preserved
   - Issues about issues: a Universal holdings statement
     * Stewardship: print archiving as much as digital archiving
     * ‘Middle child syndrome’ ALA Annual Holdings Update Forum, 2011
   - Interoperability standard for Issues and Volumes
     * ONIX for Preservation Holdings (ONIX-PH); KBART

3. Variability in publisher information
   - Transfer of title between publisher
   - Publisher naming
     * Not critical at ingest into Keepers Registry, but poor browsing
   - Identifiers for Publishers (ISNI, builds on VIAF)
     * International Standard Name Identifier (ISO 27729),
ISSN Register typically has only publisher at time of ISSN assignment

Archiving agencies (and therefore Keepers Registry) deal with current publisher
### 'as RDF Triples'

<table>
<thead>
<tr>
<th>Entity</th>
<th>Predicate</th>
<th>Entity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Publisher</td>
<td>Is (authorized by)</td>
<td>Issuing_Body</td>
</tr>
<tr>
<td>2 Serial</td>
<td>Belongs to</td>
<td>Issuing_Body</td>
</tr>
<tr>
<td>3 Issue</td>
<td>Is part of</td>
<td>Serial</td>
</tr>
<tr>
<td>4 Article</td>
<td>Is part of</td>
<td>Issue</td>
</tr>
<tr>
<td>5 Volume</td>
<td>Is attribute of</td>
<td>Issue</td>
</tr>
<tr>
<td>6 Table_of_Contents</td>
<td>Is part of</td>
<td>Issue</td>
</tr>
<tr>
<td>15 Patron</td>
<td>Is (authenticated) member of</td>
<td>Library/Institution</td>
</tr>
</tbody>
</table>
3. The Scholarly Web (the now & the future then)

a) **Preserving the digital**

b) **Digital Fixity & Copy of Record**
   - “the property of being unchanged between two points in time”, (Caplan, 2006)
   - Abrams & Rosenblum (2003): use of XML as archiving format
   - Rosenthal (2010): format obsolescence as rare problem that happens infrequently to a minority of unpopular formats
   - “the sine qua non is that the original bits be preserved”.
3. The Scholarly Web (the now & the future then)

a) Preserving the digital

b) Digital Fixity & Copy of Record

c) Citation in Serials of Serial Content

– Citation of sources is a fundamental part of scholarly discourse.

– expectation that the sources with which scholarly statements are made can and should be checked by others, with the potential for their re-use for the purposes of reproducibility of results.

– traditionally, sources relate only to statements made by other scholars and to some extent the evidence from which these statements are drawn.

• Presumption that the cited (printed) material – typically an article – sat on a shelf that could, with tedium and delay, be obtained on Inter-Library Loan
3. The Scholarly Web (the now & the future then)

a) Preserving the digital

b) Digital Fixity & Copy of Record

c) Citation in Serials of Serial Content

d) **New Models of Scholarly Communication**
   
   - Then we all began to apply electricity & get digital ...
Libraries and Publishers provide framework … the traditional ‘middleware’/infrastructure’

The Scholarly Web (the now & the future then)

Author (article)

Publisher
article → serial → issue

Value-add £ services

Licence

Licensed Online Access

ILL/docdel

Institutional arrangement

£

Library (serial)

Reader (article)

article is the ‘information object of desire’

Form a £ Economy
Mixed Modes in Scholarly Communication

Informal: ‘invisible college’ and the ‘gift economy’

Author (article) → Publisher (article → serial → issue) → repositories

Mixed Modes in Scholarly Communication

Peer review → learned society → E-prints → free2web access

License

Licensed Online Access

ILL/docdel

Institutional arrangement

Library (serial) → Reader (article)

Formal Economy

'Open Access' → 'Open Access' repositories

'Open Access' repositories

££
3. The Scholarly Web (the now & the future then)

a) Preserving the digital
b) Digital Fixity & Copy of Record
c) Citation in Serials of Serial Content
d) New Modes of Scholarly Communication
e) **Cataloguing and Archiving The Web**

Reduced ambition: assignment of identifier at points of (re-)issue

- Internet Archive (http://archive.org).

**Seriality:** identify stream of issued content

- Put simply, there can be fewer ISSN than DOI

Reminder about ease of access ...

Memento (http://www.mementoweb.org/)
f) New Scholarly Objects, with Links

Beyond enhanced publication & supplementary data (behind the graph)

- *Research Objects*: rich aggregations of linked content with suitability for re-use by machine. Archived Objects & Publication Objects which “are intended as a record of activity, and should thus be immutable” and citable.” (Bechhofer, et al, 2010)

- *Compound units*: “aggregations of distinct information units [] represented (by OAI-ORE) [to] enable it to be accessed and processed by machines & agents” (Van de Sompel & Lagoze, 2007)

“two traditions, or mentalities, even cultures, co-exist” (Buckland, 1998):

(i) Approaches based on a concern with documents, with signifying records: archives, bibliography, documentation, librarianship, records management, and the like; and

(ii) approaches based on finding uses for formal techniques, whether mechanical ... or mathematical (as in algorithmic procedures).
3. The Scholarly Web (the now & the future then)

a) Preserving the digital
b) Digital Fixity & Copy of Record
c) Citation in Serials of Serial Content
d) New Modes of Scholarly Communication
e) Cataloguing and Archiving The Web
f) New Scholarly Objects, with Links

g) Web Citation

- Range of scholarly assets published & referenced is increased
- Dynamic. What was at the end of a given HTTP URL/URI at the moment of citation can and does change, or even cease to be, when scholars wish to look up the citation.

This is referred to as ‘citation rot’.

- Study by Sanderson, Phillips & Van de Sompel (2011) found that 28% of the resources referenced by articles in an institutional repository had been lost and that 45% (66,096) of the URLs [in arXiv] that were found to still exist had not been archived.

- Expect announcement of Mellon funding of *Time Travel for the Scholarly Web* (TT4SW) as joint UoE/LANL project 😊
3. The Scholarly Web (the now & the future then)

a) Preserving the digital
b) Digital Fixity & Copy of Record
c) Citation in Serials of Serial Content
d) New Modes of Scholarly Communication
e) Cataloguing and Archiving The Web
f) New Scholarly Objects, with Links
g) Web Citation

h) Points of Issue

assignment of identifier at points of (re-)issue

- *Nature Precedings* acted as preprint repository for Life Sciences community. An integrating resource, assigned ISSN, 1756-0357.

- the long running *arXiv*, a pre-print repository for physics and comp.sci. (*and 5th in terms of h5 Index, being highly cited*) has yet to be assigned ISSN
Principles into practice  [even for unique/special objects]

1. Assign an identifier at ‘point of issue’ [ISSN for the stream]
2. Archive routinely (preferably have others/peers do that for you too)
3. Tell someone what you are doing (and how) [e.g. Keepers Registry]
4. Publish terms of access (now and when triggered as orphaned) [OA]

Make Copies, Establish Safe Places & Monitor Progress
- What is different about the digital includes the ease with which digital content can be copied, cheaply and exactly.
- Strategy for a safe places network in which the responsibility for custody of the new digital content is shared.
Development Roadmap

Brief History

The Keepers Registry Beta service provides easily accessible information about the archiving arrangements for electronic journals, and is an output of the JISC funded project, Piloting an E-journals Preservation Registry Service (PEPRS). The project commenced in August 2008 and is currently funded until July 2012. In April 2011 the PEPRS Beta service was launched with limited functionality. In October 2011, an enhanced service was re-launched as the Keepers Registry.
Down in dingle dell, four Bad Fairies dwell
Named Neglect, Decay and Loss
And WorryMuch

Come away, O human child!
To the waters and the wild
With a faery, hand in hand,
For the world's more full of weeping
than you can understand.

The Stolen Child, WB Yates
I came to tell you a story, of Fairies & The Keepers Dot Org
Your shelves don’t hold those e-journals, And what is online might, without trace, disappear!

Beware of those three Bad Fairies, Neglect, Decay and Loss.

Seek out The Good Fairies, Who will act as your digital shelves.

But what do you really know? Beware of two more Bad Fairies Called Ignorance & Missing Metadata

Looking for a Happy Ending? Then join in with The Keepers Registry At thekeepers DOT org
Use Seriality for Preservation Monitoring

Thank you for listening

p.burnhill@ed.ac.uk

http://thekeepers.org

http://thekeepers.blogs.edina.ac.uk/

http://www.flickr.com/photos/shinez/5000985919/