Workshop Report

KEEPERS EXTRA
Developing the Keepers Registry Community

Taking the Long View: International Perspectives on E-Journal Archiving

8th September 2015, Edinburgh

Contents

1. Executive Summary

On September the 8th 2015, EDINA and the ISSN International Centre hosted a workshop designed to explore the challenges of increasing preservation coverage of e-journals and related digital resources. The workshop was attended by representatives of the agencies who report into the Keepers Registry, other national libraries, and related initiatives including the Digital Preservation Coalition, the Digital Curation Centre, and UNESCO. It was organised as part of the Jisc-supported Keepers Extra Project and the key objective of the day was to scope the challenges and barriers to improving preservation coverage and to explore potential for collaborative action at an international scale.

Initial discussion covered three themes: The Long Tail of Titles by Publisher, Archiving of a Nation's Published Heritage, and Archiving Different Types of Born Digital Content. Another theme, Ensuring Completeness of a Title's Issued Content, was proposed but not discussed in the event. The significant challenges that emerged were: ensuring systems are robust and interoperable, devising scalable processes for working with publishers, incentivising others to contribute and support agencies, identifying content streams and working with new content types.

Based on the priorities and suggested actions that emerged from these discussions, ideas were regrouped around two key topics: Interoperability for Sustainability, Issues of Scope and Focus. In practical terms, Interoperability for sustainability refers to the ways in which archiving organisations and related stakeholders might work together to design robust and scalable technical and organisation systems and processes. More broadly, it also refers to measures and actions that might enable longer term planning between agencies (e.g. data sharing and succession planning) and between national initiatives (e.g. actions to counter linguistic and cultural challenges). Issues of Scope and Focus, refers to the challenges of identifying content streams and working with new

formats, deciding between priorities and future research areas, and keeping up to date with related developments and initiatives.

The Keepers Extra project team is currently revising the project roadmap and drafting a proposal for activity in response to the workshop discussions. This will be circulated to participants soon. The following grouped suggestions emerged from the individual groups and whole workshop discussions, and have been identified as areas for further research:

- Interoperability of Licenses

Key objectives here are to establish common principles that would provide support to individual negotiations and potential legal challenges, make sharing of content more feasible, and help agencies plan for transfer and successions events. This work could also be used as a foundational statement to support emergent agencies internationally

- Pro-Active shared approach towards small publishers

Related to questions around licencing, there are challenges involved in working efficiently and effectively with small publishers. Key objectives are to establish a consistent and clear message about requirements for and expectations around archiving journals and other content. It would also be desirable to further explore the challenges around selection, identifying areas in which a common approach or principle may be useful and whether there effective ways in which the Keepers Registry might support this development.

- Guidance and Standards / Technical Sustainability

Key objectives here would be to develop common standards to ensure robust and interoperable systems, facilitate improved enumeration and description of parts of serial content, and enable better analyses of holdings coverage data. This work could also be used to encourage publishers to provide better data (including the use of identifiers such as ISSN and ISNI), to define and describe useful APIS to enable sharing of data between agencies and initiatives like the Keepers Registry, and to link preservation to access and entitlement rights data in useful ways.

2. Rationale, Objectives and Themes

The workshop addressed the preservation of the born digital scholarly record and electronic serials. It was organised as a part of the Jisc-supported 'Keepers Extra' project, and brought together international archiving agencies, representatives of national libraries from around the world, and related strategic initiatives. There were 25 participants, including representatives of the ISSN IC, EDINA, 8 of the 10 Keeper agencies, the Slovakian and Vietnamese national libraries, the DPC, the DCC, and the Netherlands National Commission for UNESCO that currently manages UNESCO's PERSIST project.

The challenge of improving preservation coverage and preserving the long tail has been widely recognised. However, as the authors of a 2005 Association of Research Libraries (ARL) statement noted, "responsibility for preservation is diffuse, and the responsible parties … have been slow to

identify and invest in the necessary infrastructure." (Waters et al) A decade later, analysis undertaken as part of the Keepers Registry has shown that over 80% of continuing resources assigned an ISSN have yet to be archived. This workshop was organised as the first step towards more co-ordinated international action to improve preservation coverage: participants discussed shared priorities and the various forms in which that action may take place.

The key objectives of the day were to scope the challenges and barriers to progress in improving archival coverage, to identify shared objectives and priorities among agencies, and to explore opportunities for coordinated activity especially around the 'long tail' issue, including scalability, data standardisation, funding and division of responsibility. We also hoped to establish commitment to potential future collaborative action agendas and consider which international bodies could assist and what kind of framework would be needed to support such activity.

On the basis of consultations conducted as part of the Keepers Extra project and through statements shared by the Keepers, we identified four thematic areas for discussion arising from shared priorities. These were:

1. Ensuring Completeness of a Title's Issued Content

Volumes and Issues; Granularity of Reporting; Quality of Metadata for Serials

2. The Long Tail of Titles by Publisher

Scalability; Cost; Working with Small Publishers; Technology; Questions around Responsibility

3. Archiving of a Nation's Published Heritage

Archival Coverage by Country; Role of National & Regional Bodies; Role of Legal Deposit

4. Archiving Different Types of Born Digital Content

Defining a corpus or content stream; Government Documents; News Media; New forms of media; Identifiers for content without an ISSN

3. Workshop Structure

After an initial welcome and update on the progress of the Keepers Extra project, the workshop was conducted in three sessions, designed to facilitate knowledge exchange, discussion, and the prioritisation of ideas. In session one, participants were invited to break into groups according to the themes and scope the challenges of each theme. They were then asked to make a pitch for the importance of these challenges. In session two there was further discussion of the pitched topics and consideration of potential actions and support required. This discussion was synthesised into suggestions. Then, in session three, participants explored how to go forward with these suggestions and the activities they were willing to be involved with.

4. Report on Discussions

Session 1

Presented with the four thematic choices, participants chose to group together themes 2 and 3 ('The Long Tail of Titles by Publisher' and 'Archiving of a National Published Heritage') under the more general rubric of 'The Long Tail', and to leave theme 1 ('Ensuring Completeness of a Title's Issues Content') off the table: it was noted that this theme is important and of interest to many, but too technical to make any headway in a workshop context. This meant the participants split into two groups: Group One worked through the challenges involved in increasing coverage of 'The Long Tail': Group Two focused on the challenges involved in 'Archiving Different Types of Born Digital Content'.

Group One: The Long Tail

Focusing on the questions of how to increase preservation coverage of the long tail, and responsibility and assigning responsibility this group discussed the central issues of scalability (organisational and technical), motivating publishers to do more, general approaches and legal concerns.

There was recognition that different agencies (archives and libraries) had different resources and capabilities and that some work might be shared to create efficiencies: a common approach or standard criteria among agencies for negotiating access to digital content could make talks with publishers easier, make ingest easier and combat the idea, sometimes voiced by publishers, that working with just one agency is sufficient.

Common approaches and policies would also help create an environment in which more easily scalable approaches (such as web-crawling) were able to be leveraged to make progress: i.e. if agencies commonly took material first and then later negotiated rights or had a takedown policy this would be less controversial than if one agency acted this way alone. There was recognition that regulatory change will be slow so a risk-taking approach of acting now and amending/taking down later is required. Small publishers often don't respond or don't want to sign a license, so agreed reasonable behaviour could lower costs and barriers to gathering their material. Infrastructure to engage with publishers/subject groups/creators could also potentially be shared. Agencies could negotiate for one another or for rights to share. Such approaches become especially important in cases such as succession and transfer of responsibility from one agency to another.

There is scope for better guidance for publishers to make it easier for agencies to manage acquisition and ingest at a lower cost ("if you want your stuff preserved, do this") which would also help publishers in the marketplace, as they are producing better quality metadata which is more saleable.

Incentivising others to take action was discussed as a key strategy: this could involve articulating chains of incentives and benefits (e.g. for publishers: promotion of material, credibility through working with recognised platforms, discoverability, long term link and data preservation) and/or working to create value that is beyond the preservation service, so that preservation happens as a by-product of activity such as access or research (eg. the HathiTrust Research Centre) and people come back to and care about the material preserved.

In terms of county by country segmentation, it was recognised that awareness is very low in countries outside Western Europe and North America and that articulating chains of incentives, as a first step towards raising awareness and generating action would be important. One suggestion was to get an organisation like UNESCO to take on responsibility for gathering and coordinating gathering of material around specific subjects like e.g. climate change. The delicacy needed to take responsibility without taking ownership was flagged up.

A general concern arising from all these discussions was the idea of interoperability, in linguistic, technological, organisational and legal terms. Legal and technical interoperability would be required to facilitate country by country segmentation / web-scale archiving.

In practical terms, legal interoperability would tie into common policies and approaches by creating an interoperability of licenses, in which agencies could negotiate for each other, take common approach to entitlements, establish rights to share between licences, plan for succession/transfer and disaster scenarios. This is more relevant for larger subscription publishers who will respond.

Technical interoperability would involve common agreement on how parts are enumerated and shared and the creation of well-described APIs (data held, metadata, content). This would enable support of developing countries, lower costs and provide a basis for guidelines for publishers.

Synopsis: Group One Pitch

This topic is worth taking action on because progress is achievable. We need to preserve national heritage and the long tail is part of this. The buzz word is interoperability. Firstly, economic – we need finance to accomplish this and so we need to work together. It's how our archiving systems have to work too. Economically this is the only way to go.

Beyond economics, interoperability in terms of technology is also crucial here. We have different archiving systems, it's crucial they can talk to each other. Systems with different assumptions can be a strength but we can ensure interoperability, building modules and software to pull things in. It's important to ensure things are robust when looking at the long term, interoperability will make it more robust.

The last of the core topics is cultural and legal interoperability. Action is being taken around the world, and we recognise that the only way to deal with the long tail effectively is by motivating communities to ensure the materials that they care about are archived and preserved. A localised approach is especially important in relation to publications in diverse languages, as different cultures and readerships may value different publications. Despite their regional variations, such approaches and actions need to be able to interact in terms of communications and practices which translate across borders/languages as well as in terms of sharing of data and coordinating effort. One important foundation for such coordination is legislation: different copyright systems need to be harmonised in order to make data sharing feasible. Across the EU we are using the same terms (harmonised form of copyright), but there are systems in other places which may conflict with EU regulations (US, AUS etc. UK is common law but subject to EU regulations and therefore harmonized).

Group Two: Archiving Different Types of Born Digital Content

This group focused on the issues involved in thinking beyond e-journals to different content types. The point was made that large publishers and 'big deal' journals are fairly well preserved and we have common approaches to such material: the bigger issues are around other born digital and web resident materials which are difficult to categorise and identify. E-Books, blogs, datasets all require thinking around preservation and long term access.

It was suggested that a notion of 'completeness of the scholarly record' would be interesting: You begin with the ISSN, collect the various forms of the journal article (printed content is not always the same as online content), and the data sets which underpin it.

There may be scope for guidance for creators and publishers about how to publish new types of scholarly content so that it can be preserved, and what methods make it very difficult. It was noted that this approach has been successfully undertaken by the ADS, for example. Publishers can sometimes force authors into publishing in a particular route, so that they don't have to deal with some of the more complex problems around media change, and the review process still favours books/journal articles, so preservation is still the 'endpoint' of a process that has lots of acknowledged challenges and issues.

The group considered the need for more explicit discussion of R&D in preservation, and ability to distinguish which problems require more thought and prioritise those, rather than being always caught up in the ever-present problems of archiving and increasing coverage of what we have already identified. It was agreed that some of the new content types are different enough that we don't know how we can archive the material, and that if we don't give R&D time to such issues they will be recurrent questions/challenges to preservation activity more broadly: all of these questions are continually raised and therefore we are continually diverted – we need to have a response, even if that is 'we're working on it'.

The issues involved in collecting dynamic content which changes over time were also considered: this may be versions of articles, but also online encyclopaedias and dictionaries (e.g. Dictionary of American Regional English by Harvard UP).

News media is an interesting example: it is increasingly personalised online so that there is now no 'daily edition' experienced by all readers. Perhaps this can be through of as two layers or things to be preserved: 1) the database, and 2) the instances or presentations of that resource. The latter is tricky and potentially huge: articles are one expression, but configurations of a webpage may be another. Given the importance of data and databases in providing evidence for arguments, they also need to be preserved.

National libraries have two groups: one team focused on e-journals (well-defined & published), one team focused on web archiving (different approach): linking between the two a good starting point.

It was acknowledged that there are a lot of communities looking after other media types and that rather than taking on everything, or necessarily trying to solve the problem, we might better learn from those other groups. Certain watch areas could be outlined, which could feed into planning. An example might be online educational resources, which are interesting as they go through lots of iterations and aren't often thought about as material to be preserved.

There was discussion of scope: resources are limited, so what should the Keepers Registry and archiving agencies focus on? Are there resources that can be relatively easily archived because they are similar to an e-journal? If so, are they in scope or not? Could the registry mechanism be repurposed for other media? Does a registry mechanism help to define scope of area, identify who is acting in this place?

The broader role of identifiers was discussed, the scope of the ISSN register, and the potential development of the ISSN number to extend to other kinds of materials. The question was raised as to how strongly the ISSN is linked to print--there are other standards for other media (e.g. ISAN for audio-visual, URI/URN, BUFEC for broadcast shows (e.g. Panorama) with episodes and issues. V-ISAN as a scheme for identifying those – again, could the registry be repurposed for these identifiers?)—and whether there are bodies of material that don't qualify for ISSNs but may be of interest. The prospect of using web-syntax to identify materials such as e-thesis was also raised, as was automating the assignment of ISSN to ensure more of the material that should be getting ISSN does.

The fact that the ISSN standard will undergo a revision in the near future was discussed, and that during this process there may be scope to think about what other content types the ISSN should describe.

Synopsis: Group Two Pitch

There are many problems in the area of digital preservation and if you take the narrow view and try to focus on journals, these problems will only continue to rear their heads: we need to find achievable goals and identify where we can we take action. Born digital content, digitized content, and web resident content: we need to consider it all in relation because increasingly trying to map control structures for things in digital form doesn't work very well. Things people understand as serials have online presences and not every reader will see the same content. Those looking at these from a web archiving perspective consider this from two points of views: firstly as databases to be preserved, and secondly as also constituted by a 'transactional archive' of things that have been viewed and/or customised content. Scholarly journals don't behave like this yet, but without constraining scholarly content, archiving agencies can provide guidance on what they can and cannot capture and preserve for the future and also provide warnings that experimentation may be problematic. We need to recognise a difference between dealing with what is happening now, and future R&D questions. What can existing mechanism cope with? What is currently in scope? What do we need to work out before other things come into scope?

Session 2

In order to guide the discussion in the second hour, we grouped the points made in the pitches into general topics and subtopics. These were:

Interoperability	**Scope and Focus**
Technical	Priorities versus R&D
Cultural and organisational	Roadmaps and planning
Economic	Guidance and helping others plan

The group began by considering aspects of Interoperability that can be achieved in the short term: here examples of collaborative work between the keepers and other agencies were noted. For example, linguistic differences can make international legal and technical interoperability of data difficult, but HathiTrust has been assisted with German language around copyright by NatHosting at FIZ Karsruhe. Likewise, a national initiative can work with LOCKSS, Portico and CLOCKSS to deal with different sizes of publishers, and to leverage and develop their technologies, or it can feed information to agencies to let them know instantly about particular entitlement rights to content as when it is triggered.

Interoperability of licencing would open the door to various forms of collaboration: standard agreements around needs and entitlements, and clarity on how we want to interact with one another. We all have licences already, so we could look at those to create model licences: we already have rights to keep and rights to access, but 'rights to share'? And 'rights to have audited'?

This raised the larger question of how agencies might operate to underwrite one another and guarantee stability, in terms of taking responsibility for sharing content and succession. In the event of an agency ceasing to function, the keeper agencies could work in concert in a similar way to national libraries, although this raises questions around how to transcend national/international legal barriers. What would be the route to make such an agreement or licence internationally recognised? Could it happen under ISO umbrella?

Legislative environments need to be conducive to digital preservation: agencies need to tell the law what it should be saying, but legislative change is slow, which is another reason common standards agreed between agencies are a good idea. There are limitations to legal deposit, which is heavily constrained by laws: small-scale operations may be able to move more quickly on this

It was noted that licencing changes would not necessarily have much of an impact on the long tail, as small publishers do not typically get into licence negotiations with agencies nor wish to sign licences. For this reason, it is perhaps better to talk about common principles or standards rather than licences. Completeness is more of a problem when publishers send material to agencies, less so when agencies take it from them through harvesting. So common principles could articulate and underpin practices such as harvesting without explicit permissions.

Standards for access need to be considered along with standards for preservation. Consideration of how the adoption of such standards /licences could be encouraged beyond Western Europe and North America would be important, and how they relate to new media forms.

In terms of technical interoperation, the strategy adopted by LOCKSS is reengineering to use different types of building blocks being used most commonly in the business world. This kind of approach could be adopted by others, making agency systems more interoperable. This might mean working more closely with commercial developers (for example, as the University of Michigan has done with Google: http://safecomputing.umich.edu/protect-um-data/google.php) There are risks involved in such an approach, but they are probably worth taking.

It was suggested that one small project in the area would be to test the movement of content between different agencies. In addition, it was agreed we need well-described APIs on the data that

is held, with metadata and content. David Rosenthal, in partnership with others at Stanford Libraries and the Internet Archive has begun to map out web archiving APIs needed for preservation. An early version of this work was presented at a workshop on preserving the long tail of scholarly publications held at Columbia. A guiding question here is "if we were going to make use of other material, what would our developer want to know?"

Moving on to the topic of R&D, it was noted that interoperability (of licences and of technology) might also be envisaged with those collecting other media. There are established projects on newspaper archiving in the US for example (IFLA has a news media section, which organises training and sharing of best practices for archiving news media). Collecting information about archiving projects and initiatives would be useful: tying back to the notion of 'watch areas'. In relation to educational materials, it was noted that there may be appropriate organisations to take a lead on preservation (e.g. universities could be directed towards taking responsibility for online course materials).

It was announced that the ISSN IC will begin work on revision of the ISSN standard in early 2016, and an important factor in that will be to adjust the scope of ISSN for digital documents and ongoing resources (databases etc.). It is intended to ensure more precision around how online materials are classified and identified: planning to launch revision in October, then go through the ISO process which will take 2 to 3 years of revisions. Call for participation for working group planned to go out in January 2016. All discussions from today will feed into revision. ISSN IC invites input and suggestions for what should be assigned ISSNs: any expert in a country that is a member of ISO TC 46 may be designated as an ISO expert and become part of this working group.

Session 3

The final session of the day was devoted to discussing whether there are shared objectives and priorities that we could collaborate on and what forms actions might take.

Steve Marks (UoT/Scholars Portal) offered to pull something together around licences, transfer and succession. Areas for research include how approaches to small publishers are made (Columbia has done some work in this area), best practice, form letters, co-ordination among Ivy Plus, IIPC (bringing into ANADP). Ted Westerveldt (Library of Congress) offered to be involved too.

Vincent Wintermans suggested that, although it could not be guaranteed, UNESCO may be willing to adopt such common principles as guidelines connected to *the recommendation on preservation and access to documentary heritage including digital heritage* that the organisation will probably adopt in November 2015.

EDINA has experience with APIs, and so work with others, such as LOCKSS, on defining needs would be possible.

William Kilbride offered assistance with guidance and standards, and advised that the DPC's TechWatch Report of Sept 2013 (Preservation, Trust and Continuing Access for E-Journals: http://dx.doi.org/10.7207/twr13-04) be revisited as it contains clear recommendations and was intended to inform the community: this could be a charter to push the agenda, and could form the basis for the common principles, if the recommendations are still coherent or current.

William Kilbride also offered to take a statement to the ALPS meeting, outlining how agencies would like to see publishers contributing: a 'digital preservation for small publishers' meeting will be set up early next year. Kate Wittenberg suggested that a main contribution for others would be priority setting: learned societies, disciplinary groups could pass information to agencies about what should be published.

There was also discussion around shibboleth (Michael Seadle, Steve Marks and EDINA) and ISNI (Andrew MacEwan and EDINA), with small groups offering to discuss and share knowledge further.

5. Next Steps

The success of the workshop and the willingness of agencies to commit their time and energy to the actions and initiatives proposed throughout the day showed enthusiasm for collaboration and recognition that opportunities to share knowledge, principles and practices would be of benefit.

Drawing on proposals discussed during the workshop, the project team are currently drafting a revised project plan for the remaining nine months of the Keepers Extra project.

We will be in touch with individuals soon to progress these plans further. If in the meantime you have any feedback or suggestions please get in touch.

Adam Rusbridge a.rusbridge@ed.ac.uk
Keepers Extra Project Manager

Lisa Otty lisa.otty@ed.ac.uk
Project Officer