JISC Final Report

Project Information			
Project Identifier	*To be completed by JISC*		
Project Title	Preservation of Publications in Education (POPE)		
Project Hashtag			
Start Date	01/11/11	**End Date**	29/02/12
Lead Institution	Institute of Education (University of London)		
Project Director	Bernard Scaife		
Project Manager	Bernard Scaife		
Contact email	b.scaife@ioe.ac.uk		
Partner Institutions	N/A		
Project Web URL	http://ioepope.wordpress.com/		
Programme Name	JISC Grant Funding 12/11: Digital Infrastructure Portfolio		
Programme Manager	Neil Grindley		

Document Information			
Author(s)	Bernard Scaife		
Project Role(s)	Project Manager		
Date	16/05/12	**Filename**	POPEfinalreport_20121024.doc
URL	http://ioepope.wordpress.com/2012/10/24/final-post-and-outcomes		
Access	This report is for general dissemination		

Document History		
Version	**Date**	**Comments**
V1	**16/05/12**	
V2	**25/10/12**	**Final**

Table of Contents

1 Acknowledgements 2
2 Project Summary 2
3 Main Body of Report 2
4 Conclusions 6
5 Recommendations 7
6 Implications for the future 7
7 References ..... **Error! Bookmark not defined.**
8 Appendices 9

1 Acknowledgements

Preservation of Publications in Education (POPE) was part of the Digital Infrastructure Portfolio programme and was funded by JISC.

2 Project Summary

POPE was a four month JISC project which aimed to preserve up 5000 UK government publications which were currently linked to via the library catalogue and are rapidly disappearing from the Web due to link rot. This was in order to support research into education between 1990 and 2010. Further details about its aims and objectives can be seen in the grant submission[1].

3 Main Body of Report

3.1 Project Outputs and Outcomes

Output / Outcome Type *(e.g. report, publication, software, knowledge built)*	**Brief Description and URLs (where applicable)**
Documents referring to 6,331 Eprints documents preserved	10,568 files representing 6,331 EPrints records were identified and preserved in DERA (http://dera.ioe.ac.uk). 5,000 EPrints were specified in the deliverable
Reintroduction of lost documents to the digital arena by scanning print versions	Done
Preservation report – DROID and plan	Done
Evaluation	Included within this report

[1] http://ioepope.files.wordpress.com/2011/10/aims-and-objectives-included-in-the-pope-jisc-bid.pdf

3.2 How did you go about achieving your outputs / outcomes?

Initially, we had planned to check the links in the metadata and import those documents where the links were still live and change the status of the remaining records in order to work on these as a second sequence. These records went into a two-stage workflow in which after the document had been downloaded, regular cataloguers checked metadata before making the actual record live. However, we decided to begin by experimenting with additionally making a very quick search on Google to see whether the broken link was going to be easy to correct. This involved a simple copy and paste of the article title and looking through the first page of results only. Whilst doing this, we were lucky to discover some URL patterns which indicated that we had some groups of records which had been sourced from one site but where the base URL of the website had been subtly altered, which in turn had broken the links. In this case, the person doing the work would note this, meaning that in some subsequent records, simply by correcting the base URL, the documents could be imported without further work. It was further decided that the secondary checks were not yielding any corrections of metadata and that the records should be made live immediately once it was discovered that the links were live and therefore the document downloaded. The is pushed the throughput from 11 items per day to 112 record per day by the start of month 2 (well exceeding the 83 per day required to keep the project on track).

During the second sequence (links which were not *easily* discoverable), we allowed for some more in depth searching on the web to try and find the document. 50% of the remaining documents were now located and were imported to Digital Education Resource Archive (DERA)[2]. These normally took between 4-6 minutes per record, but sometimes as long as 30 minutes.

The remaining documents had the paper printouts retrieved and scanned. These were then imported to DERA.

In terms of evaluation, we costed the entire process of migrating and enhancing the records from the Library Management System to DERA. We kept statistics about the number of records which were preserved by category (e.g. links still live; links easily found; links difficult to find; and links not found). Finally we have made recommendations for future projects conducting similar work.

3.2.1 Statistics

6331 EPrints preserved representing 10,568 documents (average 1.66 documents per record). Of these, 92% of links were found to be not working. The majority of multiple document records were from Department for Children, Schools and Families (DCSF) consultation documents which included responses and other related forms.

Number of documents where links made live = 5715 (90%)

Number of documents with format issues (in html) 204 (3%)

Number of documents with unresolved copyright issues 343 (5%)

Duplicates discarded = 69 (1%)

Number of documents where links found to still work = 506 (8%)

Number of documents where links fixed after second sequence and more in depth searching = 700 (11%)

[2] http://dera.ioe.ac.uk/

Number of documents where links not found and rescans performed = 69 (1%)

3.2.2 Costings

Workings are in Appendix A.

Average cost per record to preserve an EPrint = £1.86
Easy to find or working link preservation cost = £1.32
Difficult to find preservation cost = £3.30
Cost to rescan = £28.71
Cost to pdf a single html document = £0.95
Cost to pdf a multiple html document = £34.50

3.2.3 Copyright

We found that we had underestimated the number of documents that would need further due diligence work performed in order to proceed with making those documents accessible. There were 125 records of this sort. These were mainly research documents from DCSF in which there had been some form of collaboration either with Universities or individuals. This means that the Open Government Licence[3] (by which the majority of our source material had been licensed) did not apply. The reason we were not aware of this was that MARC21 library catalogue records are not primarily concerned with recording intellectual property rights information. In these cases, the documents have been preserved, but cannot be released to the public until we have conducted due diligence to ascertain whether the rights holders allow this.

We are operating a notice and takedown policy on DERA[4] to ensure that any due diligence decisions which later turn out to be incorrect can be easily rectified by removing the document. This is working well. We have had one enquiry concerning a document which was “draft” but not labelled as such and available on the public web.

3.2.4 Techniques for Speeding up the retrieval of documents with broken links

The order of search repository choice is important in order to give oneself the most likely chance of finding a broken link quickly. We have discovered that the following strategy worked best for us

a. Try the National Archives UK Government Web Archive[5] (if the url we have is an original one as sometimes they have been taken from a third party site which was also hosting the content. In this case, we would be looking for the original)
b. Try Google (which does not index UK Government Web Archive material)
c. In some instances using the Internet Archive “Wayback” machine[6] can help
d. The British Library’s UK Web Archive (http://www.webarchive.org.uk): Particularly useful for Quality Assurance Agency for Higher Education (QAA) - both for England and Scotland; and Welsh material;

We discovered that using the above technique would reduce the cost from £3.30 to £1.32 per record which highlights that using library staff who are trained to know which resources to search in order to

[3] http://www.nationalarchives.gov.uk/doc/open-government-licence/
[4] http://dera.ioe.ac.uk/policies.html
[5] http://www.nationalarchives.gov.uk/webarchive/
[6] http://archive.org/web/web.php

find things can substantially reduce the cost of retrieving material which is referenced via a broken link.

3.2.5 Format issues

We had not expected to find links to documents which were actually multiple html web pages. These do not lend themselves to uploading as full-text documents. In general, this problem only occurred in items between 1999 and 2003. After this date, most documents were in either pdf or Word format. In cases of mixed or html formats, we retired the documents to be dealt with later as it was not clear how they could be "attached" as documents in EPrints.

We then sampled 61 of 204 html formatted source documents with a view to further investigating these. It was found that 36% of these linked to single pages which contained the entire entity. A freeware pdf creator was used to print the screen to a machine readable pdf which was then ingested to the system in the usual way. 64% of the sample referred to splash page type documents where links led to subsequent html pages (for example each chapter or appendix).

For multiple html link documents, we counted the links per document and averaged then which resulted in an average 36 documents per EPrint record. This was beyond the scope of the available project time, so we have costed the work involved to assist future project planning.

Another issue which was outside the scope of this project was that not all the pdf files are machine-readable. Flat image files will not have their content included in the full text search on EPrints. We were keen to ensure that as preservation remained the major driving force of this project, so the files were taken "as is" and no machine manipulation took place subsequent to their retrieval.

3.2.6 Scanning

When scanning our print backups, we had to take some basic digitisation decisions in a very short time-frame. Our aim was to provide something that was as good a surrogate for the original electronic document from which they had been printed as possible recognising that this is not primarily a high-specification digitisation project but one which must deliver accurate results economically. We therefore decided to scan documents as greyscale unless they contained colour. We used 300dpi / 24-bit resolution (colour) and 300dpi/8-bit (for greyscale). We used Adobe Acrobat X to convert the scans to OCR (so that they could be searched as full text on DERA) and produced wherever possible an output format of PDF/A in order to meet known digital preservation standards[7] as far as possible. We conducted some sampling of OCR by looking at 4 documents in detail against their originals to check for inaccuracies. Results are shown in Appendix B.

3.3 Immediate Impact

The work has meant that we no longer need to check and correct external urls for official publications in our Library Management System. We now point users to the DERA Eprints system if they are searching for this class of documents which allows them to conduct full text searching over the corpus as opposed to metadata-level searching only.

DERA stats have shown a clear increase in visits since December 2011. Whilst it is not possible to isolate this entirely to the POPE project, as the majority of our publicity on DERA has focussed on this, it is likely to correlate. From December 2011 to April 2012, the statistics are as follows:

[7] http://www.digitalpreservation.gov/formats/fdd/fdd000125.shtml

Month	Visits	% increase on Dec 2011
Dec 2011	5150	0
Jan 2012	9868	91.6
Feb 2012	11017	113.9
Mar 2012	13699	166
Apr 2012	13734	166.6

The wider community will be able to benefit from the lessons learnt from this project when conducting similar projects. Additionally, there have been a number of organisations which have ceased to exist since this project ended and for which the IOE is now the custodian of those records. This will benefit the audiences which those organisations previously served. For example, BECTA[8]. A crude measure which Google Analytics shows is that 797 clicks on the Becta Organisations list in DERA have been made between Dec 2011 and Apr 2012.

3.4 Future Impact

The future impact of POPE is that it will link into our strategy to allow education domain content to be cross searchable via a single system and ultimately for this material to be classified using our London Education Thesaurus[9] vocabulary. It will mean that people can find related material on a topic and we expect this to drive up usage of our content and refine relevance of results returned to users. We have made a start in this area with the JISC OEM-UK[10] project which releases our content as open licence linked data.

We will continue to measure the impact of the project by measuring statistical use of the data, whether in DERA or any future cross-resource discovery system. We also have plans to better track citation usage and impact within our domain and citation count increases in the preserved DERA material will be part of what is monitored in order to see whether this increases.

3.5 Business Case

The business case for preserving publications in this way is as follows:

1. By liberating these records from a library management system, one can remove the need to spend large amounts of staff time monitoring and collecting broken links.
2. By replacing broken links with full text publications, the search experience can be enrichened for the user which feeds into the mission of most libraries to “aid discovery”.
3. The preservation and access to these assets allows them to be treated as a collection for the domain opposed to a list of recommended links. This enhances the value of the domain host organisation to its community and raises its profile.
4. The cost over time of maintaining broken links on legacy material outstrips the cost of fixing the problem up-front and continuing using a more sustainable model.
5. Technology such as EPrints and other open source repository software does much of the hard work and allows smaller organisations to enter the arena of digital preservation without large overheads.

4 Conclusions

Relevant to indicator: G = General; W = Wider community; J = JISC

[8] http://www.education.gov.uk/aboutdfe/armslengthbodies/a00192537/becta
[9] http://amzn.com/0900008180
[10] http://ioeoemuk.wordpress.com/

a. It is possible to preserve a large bank of material using current EPrints technology and liberate the Library Management System to do things it is better suited to. (G,W)
b. Usage of resources previously linked to can be increased by importing them and exposing them in ways which allow for better control of metadata format, full text searching and inclusion in search engine listings. (G,W)
c. Libraries which do not have the technical skills to batch migrate records from their Library Management System to a repository may find this model difficult to replicate without paying for consultancy. (G,W,J)
d. Libraries are well placed to retrieve documents which have broken links due to their specialist searching skills.
e. There is a trade-off between the cost of up front staff resource versus the cost of ongoing maintenance which if made can benefit an organisation both in terms of overall cost and benefit to the user. (G,W)

5 Recommendations

Relevant to indicator: G = General; W = Wider community; J = JISC

a. Organisations which host externally linked documents should ensure that as a minimum, they have simple Apache re-directs in place in order, for example, that removal of a www from the URL prefix does not break batches of URLs for no good technical reason. (G,W)
b. Subject domains should consider using a similar model to preserve official publications in their arena. (G,W)
c. JISC should consider how to ensure that silos of official publication repositories do not develop if a better national discovery strategy might work, possibly via the Discovery Programme. (J)
d. Considering IPR factors more closely before starting the project may help with allocating time resources more effectively. (G)
e. A community based minimum preservation standard for re-digitising government publications which libraries hold as “final” hard copy printout for hosting on repositories would be useful. (J)

6 Implications for the future

6.1 OCR the flat files

Whilst we would not want to remove the original files which have been preserved (as they are the source files without any post-processing machine manipulation applied), it would be useful if we had surrogates for any of those which are flat files in order that they can become machine readable. This would need to be a batch routine in which the relevant files were automatically identified, OCR’d and the second copy loaded alongside the first in its relevant EPrints record (possibly with the option to hide one or other from the end user). We will try to identify these files using Preserv and Droid which are both EPrints plugins.

6.2 Remove the content from the silo

The progress of the Discovery project and our linked-data JISC project are opportunities to allow data to be cross-searched and linked in new and exciting ways. For example, we would like to be able to re-index the preserved publications using the London Education Thesaurus (LET) which would allow it to sit alongside content from other databases which had been classified under the same term.

6.3 Preservation Plan

The Preserv2 plugin on EPrints was run against the POPE content alongside the DROID plugin which attempts to identify file formats and versions in use. The following table shows the results of this:

Format	Version	Filecount
Hypertext Markup Language	4.01	4
Extensible Hypertext Markup Language	1	770
Portable Network Graphics	1	2
OLE2 Compound Document Format	-	950
Microsoft Powerpoint Presentation	97-2002	29
Acrobat PDF 1.0 - Portable Document Format	1	1
Acrobat PDF/X - Portable Document Format - Exchange 1:1999	-	6
Acrobat PDF 1.1 - Portable Document Format	1.1	14
Acrobat PDF 1.2 - Portable Document Format	1.2	817
Acrobat PDF 1.3 - Portable Document Format	1.3	1250
Acrobat PDF 1.4 - Portable Document Format	1.4	3208
Microsoft Office Open XML	2007	1
Acrobat PDF 1.5 - Portable Document Format	1.5	879
Acrobat PDF 1.6 - Portable Document Format	1.6	1176
Acrobat PDF 1.7 - Portable Document Format	1.7	22
Graphics Interchange Format	1987a	23
Microsoft Word for Windows Document	6.0/95	10
Graphics Interchange Format	1989a	241
Microsoft Word for Windows Document	97-2003	502
JPEG File Interchange Format	1.01	6
Rich Text Format	1.5	6
Acrobat PDF/A - Portable Document Format	1	73
Hypertext Markup Language	-	1095
UNKNOWN (DROID found no classification match)	-	5
Plain Text File	-	1
Stats+ Data File	-	8
Microsoft Word for Macintosh Document	6	2
Cascading Style Sheet	-	1
ZIP Format	-	48
JavaScript file	-	1
Microsoft Web Archive	-	4
Hypertext Markup Language	4.01	4

The preservation strategy and plan for the IOE Library and Archives will include migration of the older formats above to a more modern format (retaining the original) ensuring that the material which has been preserved in POPE remains viable in the long term. This is viable because there are a relatively small number of these, the majority being already a reasonably up-to-date version of Acrobat PDF. We will also make an annual appraisal of file formats to ensure they are not coming to end of life.

6.4 Sustainability

The IOE Library has undertaken to ensure that all new publications which are accessioned to the library will also be preserved (where licences allow) in DERA. This is only possible because we have saved the time we spent fixing broken links and can now use it more profitably.

6.5 Information on Project

The Wordpress site http://ioepope.wordpress.com/ will remain in place and future updates and contacts will be placed there.

7 Appendices

7.1 Appendix A: Costings

From LMS to preserved.

Pre-POPE:

Export from LMS 6331 records (Scripting) approximately 6 hours: Grade 9 * 6 hours = ~£200
Auto-load to DERA: Negligible

POPE:

Amended base URL and used this in archive £1.32 per record * 5275 recs (sub total; £6963)
Searched via Google: £3.30 per record * 700 recs (sub total; £2310)
Digitised back-up £28.71 per record * 69 (sub total; £1981)

Total cost to IOE of digitising/preserving the 6331 POPE records £11,454 (we got £9,600 from JISC for employing staff)

7.2 Appendix B: OCR sampling on four randomly chosen documents

D5310 http://dera.ioe.ac.uk/5310/

1. OCR process ignored all reversed text (eg white text on dark background). 4 pages, including the cover page, are affected.
2. For the text that is successfully OCR'd, it generates a few spelling errors, about 10 for this document. Some of the errors involves addition of blank space where there is none in the original document for example ***detail ed*** instead of ***detailed***, **gu idance** instead of **guidance**

D7160 http://dera.ioe.ac.uk/7160/

1. A rather mixed result for this document. For some pages, we get very good OCR (complete paragraphs without errors) and on some we get about 1 recognition/spelling error per line of text
2. Original document had been printed 2 pages to a sheet, which may explain higher rate of errors compared with other documents
3. Generally, special symbols e.g. degree have been recognised correctly
4. Italicised and underlined words generally recognised incorrectly

D5470 http://dera.ioe.ac.uk/5470/

This is a plain text file with only simple play on formatting, graphics and tables. Very accurate OCR.

D5510 http://dera.ioe.ac.uk/5510/

Another simple text file with columns and minimalist text formatting. Very accurate OCR.