Archiving in the Networked World: Open Access Journals

by Michael Seadle,
Editor, *Library Hi Tech*
Professor and Director,
Berlin School of Library and Information Science

Abstract

Type: research article

Purpose: The purpose is to investigate how extensively LOCKSS/CLOCKSS, Portico, and e-Depot provide long term digital archiving for the journals listed in the Directory of Open Access Journals (DOAJ).

Methodology: The article users publicly available online data, which are processed in a set of perl programs to measure number of DOAJ articles in the three archiving systems.

Findings: The findings show that only 8% of the DOAJ titles are in LOCKSS/CLOCKSS and only 5% in Portico. The findings also suggest that it could take 8 years to archive all full text DOAJ articles in e-Depot based on current plans.

Implications: The most important implication is that most open access titles listed in DOAJ currently have no effective long term digital archiving.

Introduction

The Directory of Open Access Journals (DOAJ) makes clear its commitment to archiving on its web site:

> "Long-term preservation of scholarly publications is of major importance for the research community. New formats of scholarly publications, new business models and new ways of dissemination are constantly being developed. To secure permanent access to scientific output for the future, focussed on the preservation of articles published in open access journals, a cooperation between Directory of Open Access Journals (DOAJ – www.doaj.org), developed and operated by Lund University Libraries and the e-Depot of the National Library of the Netherlands (www.kb.nl/e-depot-en) has been initiated." DOAJ, 2011

The agreement between the National Library of the Netherlands (the Koninklijke Bibliotheek or KB) and DOAJ dates from July 2009. To date the pilot project has involved only 170 journals, but the KB

plans to begin adding another 30 titles per week in the next months. In a year they should have over 1500. Ultimately the project should include all of the journals for which DOAJ has the full copy. The KB kindly supplied a list of journals for analysis. (Angevaare, 2011)

Since the KB project is only just starting and only covers full text content in DOAJ, it is worth looking at how many of the open access journals are currently in either LOCKSS (Lots of Copies Keep Stuff Safe) / CLOCKSS (Controlled LOCKSS) or Portico. Open access journals are not necessarily from small or independent publishers. Some Springer journals are listed, for example. But the overwhelming number of open access journals come as might be expected from small non-commercial publishers. These are exactly the sort of publishers that often lack the resources to implement their own archiving plan.

Long term archiving for open access journals is a key feature of the original Berlin Declaration (2003), which lists two "conditions" for open access contributions. One is of course "a free, irrevocable, worldwide, right of access..." and the other is:

> "A complete version of the work and all supplemental materials, including a copy of the permission as stated above, in an appropriate standard electronic format is deposited (and thus published) in at least one online repository using suitable technical standards (such as the Open Archive definitions) that is supported and maintained by an academic institution, scholarly society, government agency, or other well-established organization that seeks to enable open access, unrestricted distribution, inter operability, and long-term archiving."

Long term archiving is a term with many interpretations. Merely having a copy in a repository external to the original publisher offers a significant improvement in security, but does little for issues like integrity, authenticity, and potential format migration or emulation that define the contemporary discourse on digital archiving. LOCKSS / CLOCKSS, Portico, and the KB's e-Depot all address these issues, though in quite different ways. For that reason it is valuable for journals to be in more than one archiving system and certainly to be in at least one.

Research Question and Method

The research question in this column is: how extensively do LOCKSS/CLOCKSS, Portico, and e-Depot provide long term digital archiving for the journals listed in DOAJ? LOCKSS and CLOCKSS are generally treated as a single system for this study, because both use the core LOCKSS software, which is what matters for archiving purposes.

The method of analysis builds on the work done in the previous column (Seadle, 2011), which examined the overlap among journals between LOCKSS/CLOCKSS and Portico and found that 62% of the Portico titles overlap with titles in CLOCKSS or LOCKSS. The analysis also showed that, while both systems have a significant numbers of large publishers, Portico had at the time only 23 very small (single title) publishers and LOCKSS 312. On the basis of this data, it seemed plausible that LOCKSS might have more of the DOAJ titles.

The previous column compared the titles as text streams. Since then Portico colleagues have confirmed that the ISSN numbers in Portico come directly from publishers and are (or should be)

highly reliable. The same is true for LOCKSS. In fact a few journals in DOAJ do not have ISSN numbers and a few of the ISSN numbers in the other systems appear to be incomplete. This introduces some potential inaccuracy, but the number is too small to make a significant difference in the results. Matching on ISSN may be more accurate than matching on title, because of the data-cleaning problems discussed in the previous column and for that reason ISSN was used in this analysis.

The initial data were gathered by copying and pasting screenshots of DOAJ into files that were combined into a single text file. A research assistant did this work (she sensibly suggested that writing a harvesting program would have been better) and existing programs were modified to convert the screen data into a csv (comma separated values) file. In retrospect it would have been better to follow the research assistant's advice, or at least to copy the html source code, since the text-conversion process deleted the gif references "<img src=/doajImages/doajContent.gif border=0>" which turned out to be the only indication of full text and thus a key indicator for which journals the KB would eventually include in e-Depot. This meant that a harvesting program had to be written anyway to gather data about titles that would eventually be in e-Depot.

Data Analysis

The data analysis began with the screenshot file, which was processed with the program in Appendix 1 ("Data Extraction") to convert the information into a csv (comma separated values) file, using tabs rather than commas as the delimiter, since many titles had commas in them. The program put the ISSN first in each row for processing simplicity. Except for full text indicator, all elements in the record had a clear text marker. When, for example, a journal had no Creative Commons license, the word "license" did not appear.

It became immediately clear after the first sort that the file contained duplicates. Journals that fit more than one DOAJ topic area apparently had an entry in each. To correct for this the program in Appendix 2 ("Remove Duplicates") was used to remove duplicates based on ISSN matching. The resulting file had 5915 entries, while the DOAJ "new titles" search function claims 6205 entries from 1900-01-01 to 2011-02-28. The difference corresponds to the number of new journals added in the time since the data was collected. The data in the screenshot file are essentially a snapshot of DOAJ content as of the end of 2010.

The program in Appendix 3 ("Matching Program") first matched ISSNs from Portico and then from a combined and de-duplicated list of LOCKSS/CLOCKSS ISSNs against this file of DOAJ titles. The Portico and LOCKSS/CLOCKSS files were the same as had been used in the previous column. The dates on the original files suggested no updates since then. In each case two output files were created, one with matches between LOCKSS/CLOCKSS and DOAJ and a file with DOAJ titles that did not match, and also one with matches between Portico and DOAJ and a file with DOAJ titles that did not match.

The program in Appendix 4 ("Get URLs") harvested all of the URLs for the subject areas in DOAJ and put them as well as the number of titles in each area in a CSV file. Copying and editing the whole list with cut-and-paste would have taken too long, but editing the list to include "&p=" to get the additional pages for each group of one hundred entries was quick. Ideally the program would have followed the link and created these itself, but the more elegant solution is not always the easiest to implement in the short run.

The program in Appendix 5 ("Get ISSNs") then read the CSV file with the URLs, retrieved the HTML code, and searched it for: a) the full-text indicator and b) the ISSN. The ISSN numbers were then written to a new CSV file. This represented all of the titles that should be loaded into the e-Depot. Minor variations on the programs that I wrote to analyze results in the previous column compared the various CSV lists.

Results: KB

The first set of results have to do with the archiving project between DOAJ and the KB. Here are numbers that define the starting point:

- 6225 titles were in DOAJ as of 2 March 2011
- 3114 or 50% of these titles have full text
- 5836 titles were in DOAJ as of 31 Dec 2010
- 389 titles (6%) have been added to DOAJ in 2 months

If the 6% growth rate is not an anomaly (and past statistics suggest that it is not), it has significant implications for any archiving plan. The KB plans to archive about 30 journals per week, which would mean about 1560 per year, but if DOAJ were to continue to add titles at this rate, it would have at least 2334 additional titles each year, of which (based on the 50% statistic above) 1167 titles would be eligible full text works. In other words, progress toward complete archiving would take 8 years, as shown in table 1. (Note: this this calculation is conservative in making its projections based on the absolute number of titles added during the last two months rather than on a percentage increase, which would mean significantly more titles over time.)

YEAR	DOAJ FULL TEXT (1167 more each year)	KB ARCHIVES (1560 added yearly to e-Depot)	REMAINING TO ADD TO e-DEPOT
0 (March 2011)	3114	170	2944
1	4281	1730	2551
2	5448	3290	2158

3	6615	4850	1765
4	7782	6410	1372
5	8949	7970	979
6	10116	9530	586
7	11283	11090	193
8 (March 2019)	12450	12650	all archived

Table 1: archiving plan for DOAJ

It is important to remember that this calculation makes a number assumptions that may well not be true over time: the increase in the number of DOAJ titles will certainly vary, the percentage of full text titles may vary, and the rate at which the KB can include new titles into e-depot may not be precisely 30 per month. What this table demonstrates is not the future, but the prospect for the future if these current conditions hold roughly steady. Eight years is not a long time in terms of long term digital archiving, but it is long enough for small publishers to go out of business. It is one more reason why other long term archiving systems ought to be involved with the DOAJ titles.

Results: LOCKSS and Portico

The current involvement of LOCKSS/CLOCKSS and Portico with DOAJ titles is modest, but still somewhat greater than the 170 (3%) currently in e-Depot. As of the end of 2010:

- 445 (8%) DOAJ titles were in LOCKSS/CLOCKSS (356 are full text)
- 288 (5%) DOAJ titles were in Portico (270 are full text)

The overlap with the KB's plans is worth noting. Of the current 3114 DOAJ full text titles

- 181(6%) are in both LOCKSS/CLOCKSS and Portico

The unevenness of the coverage is striking: 6% of the full text (3%) of the total number of DOAJ titles are (or will be) archived in three leading systems, 5% will be in Portico and e-Depot, 8% will be in LOCKSS/CLOCKSS and e-Depot, and over 47% of the titles (mainly ones without full text) will not be in any of the three archives.

Of the titles with multiple coverage, those from just a few publishers play a significant role:

- 135 (30%) of the DOAJ titles in LOCKSS come from Hindawi Publishing Corporation
- 186 (42%) of the DOAJ titles in LOCKSS come from Biomed Central
- 10 (2%) of the DOAJ titles in LOCKSS come from Springer

- 159 (55%) of the DOAJ titles in Portico come from Hindawi Publishing Corporation
- 55 (19%) of the DOAJ titles in Portico come from Medknow Publications
- 62 (22%) of the DOAJ titles in Portico come from Copernicus Publications

In other words, three publishers account for 74% of the DOAJ titles in LOCKSS/CLOCKSS and three publishers account for 96% of the DOAJ titles in Portico. This suggests that the truly small open access publishers face a serious long term archiving exposure, unless they are ready to deliver full text to DOAJ, and even then the archinving will take time.

Reasons and Conclusions

None of this should come as a great surprise. Small and financially poor publishers do not fit the business model for Portico, which has long emphasized large publishers. Readers should not see this as a criticism of Portico. Portico charges only $250 per year for very small publishers and they offer a corporate-style service for publishers that want someone else to worry about the details of archiving. If Portico offered free services to penniless open access publishers, it would soon go out of business, which would not be good for others who depend on them.

LOCKSS charges nothing for publishers to participate. All that a publisher has to do is to put a simple digital manifest on their server that gives LOCKSS software an explicit permission to crawl their site and to harvest the content. Participating libraries must, however, elect to archive titles from these small open-access publishers. The problem may be that more libraries need to make that very modest time commitment to adding open access titles. Open access journals have no customers that pay subscription fees each year, which could well mean that they are outside the consciousness of many libraries. Digital storage space is not really a consideration. Storage costs have fallen to the point that adding text-based titles should have no noticeable financial impact. The LOCKSS software itself is of course fully open source and any group of persons or institutions with a concern to archive open access titles could set up a Private LOCKSS Network. The LOCKSS-based LuKII (LOCKSS und KOPAL: Infrastruktur und Interoperabilität) Project in Germany is currently focused on archiving data in open access repositories n Germany, but may at some point extend to DOAJ-type journals.

The KB's generosity in this situation is impressive and very welcome. The majority of the DOAJ publishers do not come from the Netherlands and in offering archiving services to them, the KB is providing a service to scholars throughout the world. The Netherlands today is a relatively rich country, but that does not mean that it is fair for scholars in other parts of the world to expect Dutch taxpayers to carry the burden independently forever.

A final conclusion is the need for more empirical data about what archiving systems are doing and how they are doing it. Only with publicly available data and with analyses that can be reproduced and tested can a genuine scholarly discourse about long term digital archiving take place. One of the

reasons to include the quick-and-dirty perl programs in the appendix of this and the previous column is to make the analysis fully transparent. Any reader can go to the same freely available online sources and use the logic in these programs to get the same answers -- or, if necessary, to find and correct errors in my results. Data are not always publicly available, of course. Publishers and (some) archiving services have an economic model that values secrecy and that is not likely to change in the near future. Nonetheless more data are available than are being analyzed and many publishers will cooperate in providing data, if asked. Future columns are likely to look at some of this data.

Acknowledgments

I would like to thank my research assistant Maria Yalpani for her help in preparing the data used in this column.

REFERENCES

Angevaare, Inge (28 Feb 2011), "DOAJ in KB e-Depot", private email

“Berlin Declaration”, 22 October 2003. Available (February 2011): http://oa.mpg.de/lang/en-uk/berlin-prozess/berliner-erklarung/

DOAJ (Directory of Open Access Journals), 2011, “Long Term Archiving”, Lund University Libraries. Available (February 2011): http://www.doaj.org/doaj?func=loadTempl&templ=longTermArchiving

Seadle, Michael, 2011, “Archiving in the Networked World: by the numbers”, Library Hi Tech, V. 29, N1. Available (February 2011): http://www.emeraldinsight.com/journals.htm?issn=0737-8831&volume=29&issue=1

Appendix 1: Data Extraction

```
# START PROCESSING
printf OUTPUT1
"ISSN\tTITLE\tPUBLISHER\tEISSN\tSTARTYEAR\tLANGUAGE\tCOUNTRY\tSUBJECT\tKEYWORDS\tCC
LICENSE\tEND YEAR\n";
while( $line = <INPUT1> ){
  chomp($line);
  $len = length($line);
  if ($len < 1) {
     $old_len = $len;
     if ($totcounter > 2) {
## WRITE RESULTS ##
        printf OUTPUT1
"$issn\t$title\t$publisher\t$eissn\t$startyear\t$language\t$country\t$subject\t$keyword\t$license\t
$endyear\n";
```

```
## RESET VARIABLES ##
        print "ISSN = $issn, TITLE = $title\n";
        $eissn = 0;
        $publisher = "null";
        $country = "null";
        $keyword = "null";
        $subject = "null";
        $startyear = 0;
        $license = "none";
        $endyear = "null";
        }
      next;}
    ++$totcounter;
    $first_five = substr $line,0,5;
    $first_ten = substr $line,0,10;
    $first_nine = substr $line,0,9;
    $first_eight = substr $line,0,8;

      if ($old_len < 1) {
        $title = $line;
        }
      elsif ($first_five eq "ISSN:") {
        $issn = "\'".substr $line,6,8;
       # $title = $prevline;  # pick up the title
        ++$title_counter;
        }
      elsif ($first_ten eq "Publisher:") {
        $publisher = substr $line,10;
        }
      elsif ($first_nine eq "Language:") {
        $language = substr $line,10;
        }
      elsif ($first_nine eq "Country: ") {
        $country = substr $line,9;
        }
      elsif ($first_nine eq "Keywords:") {
        $keyword = substr $line,10;
        }
      elsif ($first_eight eq "Subject:") {
        $subject = substr $line,9;
```

```
        }
      elsif ($first_five eq "EISSN") {
        $eissn = substr $line,6;
        }
      elsif ($first_ten eq "Start year") {
        $startyear = substr $line,12;
        }
     elsif ($first_eight eq "License:") {
        $license ="CC License";
        }
     elsif ($first_nine eq "End year:") {
       $endyear = substr $line,10;
       }
      else {$prevline = $line; }
    $old_len = $len;

} # end
```

Appendix 2: Remove Duplicates

```
while( $line = <INPUT> ){
  chomp($line);
  $len = length($line);
  if ($len < 1) {next}
  if ($prevline eq $line) {
     ++$count;
     print "Dup $count for $line\n";
     }
  else {
     printf OUTPUT "$line\n";
     $prevline = $line;
     }
  }
```

Appendix 3: Matching Program

```
while( $line = <INPUT1> ){
  chomp($line);
  $len = length($line);
  $issn_a = substr ($line,1,4);
```

```
 $issn_b = substr ($line,5,4);
 $issn = $issn_a."-".$issn_b;

 if ($len < 1) {next}
  ++$totcounter;
   $hit = 0; #reset hit counter
   $hit = grep (/$issn/, @list);
   if ($hit < 1) {
      ++$NOhitcounter;
      printf OUTPUT2 "$line\n";
      }
  if ($hit > 0) {
      ++$HITcounter;
      printf OUTPUT1 "$line\n";
     }

print "hit = $hit for DupCounter = $HITcounter or Nondup counter = $NOhitcounter for totalcounter =
$totcounter \n";
}
```

Appendix 4: Get URLs

```
while ($content =~ m/href=\/doaj\?func=subject\&cpid/g) {
  $pos = pos($content);
  # print "$counter\n";
   $url = substr $content,$pos+1,3;
   $size = substr $content,$pos+1,100;
  if ($size =~ m/\(\d/g) {
    $pos2 = pos($size);
    $size2 = substr $size,$pos2-1,3;
    }
  else {
     $size2 = "null";
    }
  if ($size2=~ m/j/g) {
     my $pos = pos($size2);
     $size2 = substr $size2,0,$pos-1;
     }
  if ($url =~ m/\>/g) {
    my $pos = pos($url);
   $url2 = substr $url,0,$pos-1;
    }
  else {
    $url2 =  $url;
```

```
  }
 printf OUTPUT "http:\/\/www.doaj.org/doaj\?func=subject\&cpid=" . "$url2, $size2\n";
 $counter++;
 }
```

Appendix 5: Get-ISSNs

```
while( $line = <INPUT> ){
 chomp($line);
 $line =~ m/,/g;
 $pos = pos($line);
 $url = substr $line,0,$pos-1;
 $url_counter++;
 print "COUNT $url_counter URL $url\n";
 use LWP::Simple;
 my $content = get $url || die "could not get $url";

 while ($content =~ m/doajContent/g) {
  $pos = pos($content);
  $stuff = substr $',37,8;
  $url_pt1 = substr $stuff,0,4;
  $url_pt2 = substr $stuff,4,4;
  $url_combo = "$url_pt1" . "-" ."$url_pt2";
  printf OUTPUT "$url_combo\n";
  $counter++;
  }
 }
```