Archiving in the Networked World: By the Numbers

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

Abstract

Type: research article

Purpose: The purpose is to investigate: 1) how many journal titles are both in LOCKSS and in Portico? 2) what is the relationship of small publishers to LOCKSS/CLOCKSS and Portico? 2) what is the relationship of large publishers to LOCKSS/CLOCKSS and Portico?

Methodology: The article describes how data from Portico, LOCKSS, and CLOCKSS was cleaned and analyzed using Perl programs to discover duplications.

Findings: The findings show a significant overlap among the archiving systems. It also shows that Portico has no bias against small publishers and that large publisher are as willing to choose the LOCKSS software as to choose Portico. LOCKSS does, however, archive many more small and arguably endangered publishers and may be the only economically viable choice for them.

Originality/value: The push for greater transparency has made more and more data available. Both LOCKSS and Portico deserve commendation for providing the detailed lists of titles and publishers on which this article was based. Such data gives the library community an opportunity to build decisions about the long term digital future on firm and verifiable ground.

Keywords: Archiving, Data analysis, Publishing, Statistics

Introduction

During discussions in Germany about a National Hosting" plan a number of questions arose about which publishers work with LOCKSS (Lots of Copies Keep stuff Safe) and which with Portico. (For background about the National Hosting process, see Seadle (2010).) One person suggested that Portico was not open to hosting small publishers. Others wondered about how willing large publishers were to work with LOCKSS, since a rumor had circulated that large

publishers may be unwilling to work with LOCKSS.

Four days later during the CLOCKSS (Controlled LOCKSS) [1] board meeting in New York in November 2010, the representative of major publisher said that it made sense to use more than one archiving system for long term security. [2] Erik Oltmans (2006) and Hilde van Wijngaarden from the Koninklijke Bibliotheek in the Hague, Netherlands, had expressed a similar opinion in an article four years earlier:

> "Publishers are not likely to deposit their material at an infinite number of digital archives. They probably want to sign archiving contracts with a limited number of institutions around the world to deposit their materials – partly to spread their risks and partly for geopolitical reasons."

Because of these questions, it made sense to investigate the following questions:

1) **Duplication**: how many journal titles are both in LOCKSS and in Portico?
2) **Small publishers**: what is their relationship to LOCKSS/CLOCKSS and Portico?
3) **Large publishers**: what is their relationship to LOCKSS/CLOCKSS and Portico?

Data Sources

Both Portico and LOCKSS offer detailed information about which journals and which publishers use their services. All downloads of data were made between 3 to 10 December 2010.

The Portico data was found on the Portico website [4] as a "downloadable list of committed e-journal titles and holdings". The file came with four tables: introduction, summary, details, and definitions. The table labeled "details" contained publisher, title, society, print ISSN, e-ISSN, PCA (post-cancellation access), status, and holdings. The entries under publisher, title, and the ISSN numbers were clear. Society appeared to be the name of the publishing society. Status and holdings detailed whether a title was "queued" for processing or "preserved". If the title had been "preserved" then the holdings were listed in detail. For this analysis, only the columns for title and publisher were used. All content, whether queued or preserved was treated equally as being in Portico. The Excel spreadsheet was reduced to just the detailed publisher and title, the order of the columns was reversed, and the file was converted to a CSV (comma separated values) file for processing. In addition to 11,958 e-journal titles, the Portico site also lists "E-book Titles (65,986)" and "D-Collections (39). The e-book and d-collections were not considered in this analysis. No obvious date in the file or on the website indicated how old the data might be. The presumption was that it was relatively current.

The LOCKSS data was found on the LOCKSS website [5] under "Publishers and Titles" with a choice CSV, Excel, OpenDocument, or PDF format. The website also offered an overall metric

that LOCKSS has “over 6,600 e-journal titles from approximately 450 publishers”. The file had titles and publishers from established publishers comparable to the list from Portico. The file included a single table that listed publisher, title, ISSN, and e-ISSN. The columns for title and publisher were reversed and the ISSN columns removed for processing.

LOCKSS also preserves works from the Digital Federal Depository Library Program. These are essentially e-book materials and were not included in the analysis. The website also refers to six “Private LOCKSS Networks”, some relatively large such as the MetaArchive [3]. These have been left out of the current analysis since they deal with a broader range of material than journals from established publishers.

The LOCKSS site also offers information about CLOCKSS, which uses the LOCKSS software and the same preservation techniques, but has a different business model that results in a dark archive that does not allow automatic access to those with established access rights, as is the case for global LOCKSS. Since the contents in CLOCKSS are of the same type as in Portico and in global LOCKSS, it seemed reasonable to include the CLOCKSS data in the analysis. The CLOCKSS website offers a list of “Committed E-Journal Titles” in the same layout as for LOCKSS, but only in Excel and PDF format. The Excel format was used and transformed to a CSV file. As with Portico, no obvious date indicated how old the file might be and the presumption was that it was relatively current.

Data cleaning

The data from all three sources showed careful attention to detail and appeared to need little or no cleaning. After some analysis, however, it became clear that some matches were failing because Portico had added comments to distinguish titles with very similar or identical titles and different publishers. The LOCKSS and CLOCKSS data did not add any comments and used the publisher name in the next column to distinguish titles. The number of affected titles was relatively small, only 822 our of 11628 titles in Portico or about 7%. Nonetheless it was more than could reasonably be cleaned by hand. A short Perl program was used to remove them. The core routine of the program may be found in Appendix 1. The program is included for those who want to check for logic errors or to reproduce the analysis.

This program read each line from the Portico file (INPUTP), looked for the characteristic open-parenthesis (“that began a comment and the close-parenthesis“)that ended it. A check of the LOCKSS and CLOCKSS files found no titles with parentheses in their name, which made it likely (though not certain) that none of the comments within the parentheses belonged to actual titles, though the possibility for non-duplicate titles could not be excluded. In any case this would understate rather than overstate the number of matches. Other data irregularities such as typing errors could also affect all three source files. An alternative solution would have been to match on ISSN numbers, if the numbers were matched automatically and not entered by hand (which

could be more subject to data entry errors than the titles).

Question 1: Duplication

The three CSV files were processed with a set of simple Perl programs running under Mac OS 10.6.5. The first program only checked for titles in Portico that were also in LOCKSS. To simplify processing, the LOCKSS file was reduced only to titles and then loaded into an array (@list = <INPUTL>;). The Portico file was then read line by line, the title stripped out, and then searched for a match against the array of LOCKSS titles using the code in Appendix 2.

The *OUTPUTND* file was for non-duplicates and *OUTPUTD* was for duplicate titles. The $NDcounter and $Dcounter tracked the number of non-duplicate and duplicate titles to ensure that the totals added correctly. The whole line from the Portico entry (including title and publisher) was written to the respective file. A minor change in the names of the input and output files allowed the same program to run against a file of CLOCKSS titles only. A similar change could establish the amount of duplication between LOCKSS and CLOCKSS. Combining the CLOCKSS titles with the titles from LOCKSS and removing duplicates with the Perl program in appendix 3 allowed a final run of the program looking for duplication between Portico and LOCKSS and CLOCKSS together.

The results from the analysis show significant overlap among the archiving systems, even though the matching process likely understated the amount of duplication, since only a perfect match counted from the program's viewpoint.

- 7256 Portico titles (62% of the Portico total) are identical with titles in either CLOCKSS or LOCKSS (69% of the combined LOCKSS and CLOCKSS total).
- 3242 CLOCKSS or LOCKSS titles (31% of the combined LOCKSS and CLOCKSS total) are not in Portico.
- 4372 (38%) Portico titles are not in CLOCKSS or LOCKSS.

Part of the overlap comes from a few large publishers, who use both systems.

- Of the 11,628 Portico titles, 2474 or 21% come from Elsevier.
- Of the 11,628 Portico titles, 1033 or 9% come from Springer.
- Of the 7403 CLOCKSS titles, 1970 or 27% come from Elsevier.
- Of the 6598 LOCKSS titles, 2460 or 37% come from Springer

If anything this trend toward duplication seems to be increasing. No large publisher has pulled out of an archiving system and more appear to be hedging their bets with both Portico and either LOCKSS or CLOCKSS. Interestingly there is relatively little overlap between LOCKSS and CLOCKSS, only 253 clear matches. This suggests that publishers recognize that the

LOCKSS software lies behind both LOCKSS and CLOCKSS and that they choose whether or not they want a dark archive.

Question 2: Small Publishers

The definition of a large or small publisher could depend on a number of measures. For this analysis I have used a simple metric based on the number of titles in the archive: publishers with fewer than 30 titles in an archiving system count as small publishers. This is obviously a flawed definition, since a large publisher could contribute only a few titles to one or another archiving system. That appears, however, not to be the case. This definition could tend to elevate the number of small publishers and to understate the large ones.

To simplify the analysis, a program was used to count how many titles belonged to each publisher in the LOCKSS, CLOCKSS, and Portico lists. The logic of the program is in appendix 4. The program isolates the publisher's name in the CSV file, and counts the entries until the next publisher name appears. The routine in appendix 4 is for LOCKSS. The same routine repeats for Portico and CLOCKSS. The programs writes the total number of journal titles and the publisher name to a file.

The program provided the following results:

- 433 of the 450 publishers in LOCKSS have fewer than 30 titles in the system (96%).
- 15 of the 30 publishers in CLOCKSS have fewer than 30 titles in the system (50%).
- 66 of the 111 publishers in Portico have fewer than 30 titles in the system (60%).

By this measure, Portico is certainly not unfriendly to small publishers, but LOCKSS has a comparatively overwhelming number of small publishers. CLOCKSS clearly works more with larger than smaller publishers, which makes sense, since it was created at the initiative of some large publishers.

If single title publishers in the archiving system were used as the definition of a small publisher, the split becomes more extreme:

- 312 of the 450 publishers in LOCKSS have only 1 title in the system (69%)
- 4 of the 30 publishers in CLOCKSS have only 1 title in the system (13%).
- 23 of the 111 publishers in Portico have have only 1 title in the system (21%).

None of these measures suggest that Portico is unfriendly toward small publishers, though clearly small publishers are not its focus. The measures do suggest that global LOCKSS has a strong focus on small publishers, and these results make sense in terms of the concern among global LOCKSS staff that very small publishers are especially endangered.

Small publishers are not the cheap to work with. A JISC report (Beagrie, 2008) says:

> *"The profile of costs across functions within the national data centres we interviewed appears to be very consistent. It was notable that they all believed their accessioning and ingest costs were higher than ongoing long-term preservation and archiving costs."*

The table accompanying this text lists access and ingests costs at 42%. The acquisition and ingest process is publisher dependent because of ingest format issues. Archiving these small and (arguably) publishers is troublesome and expensive compared to the low handing fruit of large publishers, where a singe ingest process handles large numbers of titles. In emphasizing small publishers global LOCKSS has made a service choice, not an economic one.

Question 3: Large Publishers

The same data can be used to look at the relationship between archiving systems and large publishers. If 100 titles is used as the cutoff for large publishers, the results are:

- 8 of the 450 publishers in LOCKSS have 100 or more titles in the system (2%)
- 9 of the 30 publishers in CLOCKSS have 100 or more titles in the system (30%).
- 19 of the 111 publishers in Portico have 100 or more titles in the system (17%).

These figures suggest that CLOCKSS is more strongly oriented toward large publishers than Portico and that, while LOCKSS has large publishers, its focus has been elsewhere. If the cutoff is 1000 titles instead of 100, the results change very little.

- 2 of the 450 publishers in LOCKSS have 1000 or more titles in the system (1%)
- 4 of the 30 publishers in CLOCKSS have 1000 or more titles in the system (13%).
- 4 of the 111 publishers in Portico have 1000 or more titles in the system (4%).

It is clear that large and very large publishers such as Elsevier and Springer have no fundamental dislike for the LOCKSS software. The evidence suggests that large publishers trust LOCKSS/CLOCKSS and Portico equally.

Consequences

The details of these statistics are by no means completely reliable. They rely on matches based on assumptions that the entries in the tables for over 25,637 titles are so accurate, that the matches performed against them are reliable. It is important to remember that the processing understates duplication. Nonetheless a few striking results deserve comment.

The first of these results is the degree to which publishers, especially large publishers, are

ready to work with multiple archiving systems. That makes sense in several ways. The cost of belonging to CLOCKSS, LOCKSS, or Portico may seem significant to libraries, but for large publishers it is small change. These companies are not specialists in digital archiving. They may incline toward the marketing arguments from one or another system, but they also know from their own business experience that the claims of software suppliers do not always match their promises, especially over long periods. Claims are easy to make in marketing presentation, but harder to prove. Is Portico's migration system really a solution that will make materials available in 100 years? Or are the measures that LOCKSS uses to preserve the bitstream really necessary for content to be there after a century? When the costs of participation are low enough, publishers have no reason not to spread their risk.

Smaller publishers are in a different situation. For them the membership costs in multiple archiving systems are sums as significant as for libraries – and for the very small publishers even more problematic. Very small publishers also have no scholarly or engineering basis for choosing one or another system. The preference for LOCKSS should be seen in significant measure as a financial decision. LOCKSS is cheaper and any archiving system is better than none. This does not mean that small publishers avoid Portico, only that Portico's prices offer them less incentive in a market where they have no other readily reliable evidence on which base their decision.

It is a reasonable assumption that some archiving system is better than none, especially for small publishers that lack the technology infrastructure to survive a serious server crash or the financial resources to guarantee a long term presence for digital materials. One of the clear results of the data in the analysis in this article is that LOCKSS is the only archiving solution for the vast majority of these small publishers. Few librarians serious expect a large publisher like Elsevier or Springer to vanish overnight, but the danger for a publisher with only one or even fewer than ten titles is historically much larger. This means that, in the real world at present, the only archiving system that genuinely protects endangered content is LOCKSS – if only because it is the only system that they can afford.

CONCLUSION

The explicit goal of this article has been to present data to help answer three questions that have come up during discussions about national hosting and and other venues. A broader and perhaps more important goal has been to base the discourse on publicly available data. Publicly available data about archiving systems and publisher choices are not always easy to discover, but the push for greater transparency has made more and more such data available. Both LOCKSS and Portico deserve commendation for providing the detailed lists of titles and publishers on which this article was based. Such data gives the library community an opportunity to build decisions about the long term digital future on firm and verifiable ground.

Notes

[1] http://www.clockss.org/

[2] Board meetings are confidential. For this reason the reference is anonymous.

[3] http://httpwwwhttp.httpmetaarchivehttp.httporghttp/http

[4]http://httpwwwhttp.httpporticohttp.httporghttp/httpdigitalhttp-httppreservationhttp/httpwhohttp-httpparticipateshttp-httpinhttp-httpporticohttp/httpparticipatinghttp-httptitleshttp/http

[5]http://httpwwwhttp.httplocksshttp.httporghttp/httplocksshttp/httpPublishershttp_httpandhttp_httpTitleshttp

[6]http://httpwwwhttp.httpclocksshttp.httporghttp/httpclocksshttp/httpParticipatinghttp_httpPublishershttp

References

Beagrie, Neil, Julia Chruszcz and Brian Lavoie (12 May 2008). Keeping Research Data Safe a Cost Model and Guidance for UK Universities. Charles Beagrie Limited. A study funded by JISC. Available (10 December 2010): JISC

Oltmans, Erik, and Hilde van Wijngaarden (2006). The KB e-Depot digital archiving policy. *Library Hi Tech* 24, no. 4, pp. 604-613 Available (5 December 2010): Emerald.

Seadle, Michael (December 2010). Archiving in the networked world: LOCKSS and national hosting. *Library Hi Tech* 28, no. 4. pp. 710 - 717. Available (5 December 2010): Emerald.

Appendix 1

Logic from the perl program to clean data.

```
my $counter = 0;
while( $line = <INPUTP> ){
  chomp($line);
  $len = length($line);
   if ($line =~ m/\(/g) {
      $pos = pos($line);
      print "POSITION is $pos\n";
      $line =~ m/\)/g;
      $pos2 = pos($line);
      $cleanline1 = substr $line,0,$pos-2;
      $cleanline2 = substr $line,$pos2 ;
```

```
    ++$counter;
    print "Counter = $counter\n";
    printf OUTPUTND "$cleanline1$cleanline2\n";
    print "Line = $cleanline1$cleanline2 \n";
    }
  else {printf OUTPUTND "$line\n";}
 }
```

Appendix 2

Logic from the perl program to match data.

```
my $Dcounter= 0; # dup counter
my $NDcounter= 0; # non dup counter
my $totcounter = 0;
while( $line = <INPUTTP> ){
 chomp($line);
 $len = length($line);
 print "$line\n";
 $commaloc = index ($line,",");
 $publisher = substr ($line,$commaloc);
 $title = substr ($line,0,$commaloc);
 if ($len < 1) {next}
  ++$totcounter;
#
#   ($hit > 0) means a match and a dup title
#   ($hit < 1) means no match and no dup.
#
   $hit = 0; #reset hit counter
   $hit = grep (/$title/, @list);
   if ($hit < 1) {
      ++$NDcounter;
      printf OUTPUTND "$line\n";
  }
  if ($hit > 0) {
      ++$Dcounter;
      printf OUTPUTD "$line\n";
  }
}
```

Appendix 3

Logic from the perl program to remove duplicates.

```
my $prevline = " ";
my $line = " ";
my $count = 0;
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 4

Logic from the perl program to count the number of titles per publisher.

```
printf OUTPUT "LOCKSS\n";
print "LOCKSS\n";
while( $line = <INPUTL> ){
  chomp($line);
  $len = length($line);
  $commaloc1 = index ($line,"\",");
  $publisher = substr ($line,$commaloc1 + 2);
  print "Publisher is $publisher\n";
  #$publisher =~ tr/, "/ /;
  $title = substr ($line,0,$commaloc1);
  if ($len < 1) {next}
  if ($totcounter == 0) {
    $prevpub = $publisher;
    }
   ++$totcounter;
   if ($prevpub ne $publisher) {
     printf OUTPUT "$pubcount,$prevpub\n";
     $prevpub = $publisher;
```

```
  $pubcount = 1;
   }
else {
  ++$pubcount;
}
```