

Web Library: Data Movement Spring 2007 Report

Dmitriy Shtokman (ds346)
Cornell University
Advisors: Ruth Mitchell, William Arms

Table of Contents

	Abstract.....	3
1.	Data transfer process.....	3
2.	DP Crawl Status.....	4
3.	DV Crawl Status.....	4
4.	ED Crawl Status (amzn)	7
5.	EB Crawl Status (verification)	9
6.	File Name Patterns.....	11
7.	Download Problems Encountered.....	12
8.	Conclusion.....	13
9.	Terminology.....	13
10.	Acknowledgements.....	14
11.	References.....	14
	Appendix I: Cumulative Transfers to Date.....	14

Abstract

The purpose of this report is to describe the work completed on transferring data from the Internet Archive in California to the Web Laboratory machine in Cornell during the Spring 2007 semester. Data Movement Team is responsible for moving specified subsets of archived data to make it available for research purposes here at Cornell. Once the data has been moved to the local storage, it is archived and stored in the database. The following report focuses on the usage of the Semi-Automated System. The Automation System was developed independently and is described in the Data Movement and Tracking Spring 2007 report by Andrzej Kielbasinski.

1. Data Transfer Process

During the Spring 2007 the Semi-Automated System that had been developed previously was utilized to ensure continuous data transfer process. For the detailed description of the Semi-Automated System and FileZilla usage, please refer to the Web Library: Data Movement Spring 2006 and Fall 2006 reports. Below are the major steps required to perform data transfers:

a) Obtaining and verifying a node list

A node list for a specified crawl is generated by the PHP script that contacts the Internet Archive and retrieves the names of those nodes that contain files for the crawl. Since sometimes Internet Archive nodes are non-responsive, the script must be run several times to ensure integrity. NodesApplication is also utilized – it provides us with the list of nodes and a list of files that each of these nodes contains.

b) Obtaining queue files

Once a node list has been created, another PHP script generates queue files that may be imported into FileZilla. Unfortunately, sometimes Internet Archive nodes are non-responsive, resulting in empty queue files being generated. As of such, it may be necessary to run the PHP script several times for the same node names, until all the queue files are successfully created.

c) Upload queue files

Multiple FileZilla instances are launched to provide concurrent downloads. Queue files that have been created are imported into FileZilla. Download options are

specified, so that nine connecting threads access one node, and original time/date is preserved for the files that are being transferred. When problems with nodes occur, these nodes can be accessed manually using the SmartFTP client. Logging is done manually by keeping and updating the status file.

During the first half of the semester the data was downloaded to the scidata1 machine, because some DP crawl files had already been partially transferred there, and DP crawl was to be finished. Once DP crawl was completed, the data for next crawls was transferred to the weblab machine (weblab.tc.cornell.edu) instead of scidata1 machine (scidata1.tc.cornell.edu). A few problems with Filezilla were observed. However, the weblab machine was clearly more reliable and on average provided a better throughput per connecting thread (400+ KB/s v. 200+ KB/s). Average throughput was about 200 GB/day for the scidata1 machine and 300 GB/day for the weblab machine. While the system capacity allows for higher throughput, manual work involved put a constraint on the amount of data transferred.

2. DP Crawl Status

A substantial portion of DP data had been downloaded during the Fall 2006 semester. A large subset of unprocessed nodes containing DP data was discovered at the end of the Fall 2006 semester. The most probable reason for these nodes being missed earlier in the semester is internal file movement in the Internet Archive.

This semester it was necessary to complete the DP crawl. Since previously downloaded DP crawl files were located on the scidata1 machine, it was decided to continue downloading DP crawl files to the same location and then switch to the weblab machine. DP crawl was finished on March 12, 2007. 161,712 ARC files (14.9 TB) and 161,712 DAT files (0.84 TB) were downloaded to the scidata1 machine.

3. DV Crawl Status

Once DP crawl was finished, DV crawl was to be processed next. In April an additional subset of IA nodes that contain DV data was discovered. A lot of them only stored a single pair of DV files (ARC and DAT). Nevertheless, the amount of DV data

available turned out to be larger than expected originally. The likely reason is that more DV data was moved to the IA nodes after the spreadsheet with crawl data had been constructed. This spreadsheet had been used for reference and updated accordingly.

An observation that I made was that the node ia311136 contained a corresponding pair of DV files (ARC and DAT) on April 6. However, when I accessed this node manually on April 8, after the download from that node had failed, this pair of files was not there any more. It could be an indication that the Internet Archive staff was still moving data I was processing during the semester. This could potentially result in duplication problem.

First, both ARC and DAT files were being transferred concurrently. However, in the middle of the semester there was a change in the data that had to be downloaded. The new priority was to obtain DV DAT files and ED_amzn ARC files (described in the next section), leaving DV ARC files to be downloaded later. DV crawl DAT files were finished on April 25, 2007. 263,259 files were downloaded from 431 nodes. The total size of DV DAT files is 1.89 TB. All the nodes available were processed and the data was stored to the weblab machine. DV ARC files were not finished, although a significant number of them was downloaded. Queue files for all the DV ARC files were generated along with queue files for DV DAT files and are available.

Certain problems were encountered when processing nodes for DV crawl:

a) “No such file or directory” - some files seemed to have been removed or are inaccessible because of disk failure.

Node ia311425.us.archive.org, disk 0; file DV_crawl23.20040318155257.dat.gz was there on April 21, 2007, but was not accessible on April 22, 2007.

Node ia311432.us.archive.org, disk 1; file DV_crawl13.20040220040903.dat.gz was there on April 21, 2007, but was not accessible on April 22, 2007.

Node ia311437.us.archive.org, disk 3; file DV_crawl24.20040318034432.dat.gz was there on April 21, 2007, but was not accessible on April 22, 2007.

Node ia331214.us.archive.org, disk 0; file DV_crawl12.20040208173903.dat.gz was there on April 21, 2007, but was not accessible on April 22, 2007.

Node ia331218.us.archive.org, disk 0; file DV_crawl24.20040213170913.dat.gz was there on April 21, 2007, but was not accessible on April 22, 2007.

Node ia331242.us.archive.org, disk 3; file DV_crawl23.20040316130040.dat.gz was there on April 22, 2007, but was not accessible on April 23, 2007.

Node ia340726.us.archive.org, disk 3; file DV_crawl25.20040318135124.dat.gz was there on April 22, 2007, but was not accessible on April 24, 2007.

Node ia340808.us.archive.org, disk 0; files DV_crawl9.20040326231116.dat.gz and DV_crawl25.20040225122639.dat.gz were there on April 22, 2007, but were not accessible on April 24, 2007.

Node ia340821.us.archive.org, disk 0. The following files were there on April 23, 2007, but were not accessible on April 24, 2007:

DV_crawl23.20040317091732.dat.gz
DV_crawl9.20040325085219.dat.gz
DV_crawl23.20040312150041.dat.gz
DV_crawl22.20040315125651.dat.gz
DV_crawl12.20040217063107.dat.gz
DV_crawl25.20040215225517.dat.gz
DV_crawl11.20040310021827.dat.gz
DV_crawl23.20040212044636.dat.gz
DV_crawl10.20040326034744.dat.gz
DV_crawl25.20040320040500.dat.gz
DV_crawl9.20040325085219.dat.gz"
DV_crawl9.20040322013958.dat.gz"
DV_crawl25.20040320040500.dat.gz
DV_crawl25.20040315135630.dat.gz
DV_crawl25.20040215225517.dat.gz
DV_crawl9.20040303075310.dat.gz
DV_crawl23.20040317091732.dat.gz
DV_crawl23.20040316112528.dat.gz
DV_crawl23.20040312150041.dat.gz
DV_crawl23.20040212044636.dat.gz
DV_crawl12.20040217063107.dat.gz
DV_crawl9.20040303075310.dat.gz

DV_crawl22.20040315125651.dat.gz
DV_crawl11.20040310021827.dat.gz
DV_crawl10.20040326034744.dat.gz
DV_crawl25.20040215225517.dat.gz
DV_crawl9.20040322013958.dat.gz
DV_crawl9.20040325085219.dat.gz
DV_crawl23.20040316112528.dat.gz
DV_crawl9.20040303075310.dat.gz
DV_crawl9.20040322013958.dat.gz
DV_crawl25.20040315135630.dat.gz
DV_crawl11.20040310021827.dat.gz
DV_crawl25.20040320040500.dat.gz
DV_crawl23.20040316112528.dat.gz
DV_crawl12.20040217063107.dat.gz
DV_crawl22.20040315125651.dat.gz
DV_crawl25.20040315135630.dat.gz
DV_crawl10.20040326034744.dat.gz
DV_crawl23.20040312150041.dat.gz
DV_crawl23.20040317091732.dat.gz
DV_crawl23.20040212044636.dat.gz

b) The following nodes contain DV data and are currently connectable via a web browser, but it's impossible to connect to them via an FTP client (Filezilla or SmartFTP):

ia300928.us.archive.org
ia300640.us.archive.org
ia300641.us.archive.org
ia300539.us.archive.org

c) The following nodes are unreachable hosts and may contain DV data because they fall in the sequence of nodes that store DV files:

ia331405.us.archive.org
ia331413.us.archive.org

4. ED Crawl Status (amzn)

Another Web Laboratory team needed a focused subset of files from the Internet Archive – ARC files from ED crawl whose names contain the word “amzn”, such as ED_amzn_crawl31.20050627091044.arc.gz . While downloading the entire ED crawl was not required, the ED_amzn ARC files were needed within a short period of time. To download them, existing Perl scripts and NodesApplication were utilized. Although writing a new script exclusively for this task might have been a better solution, given the

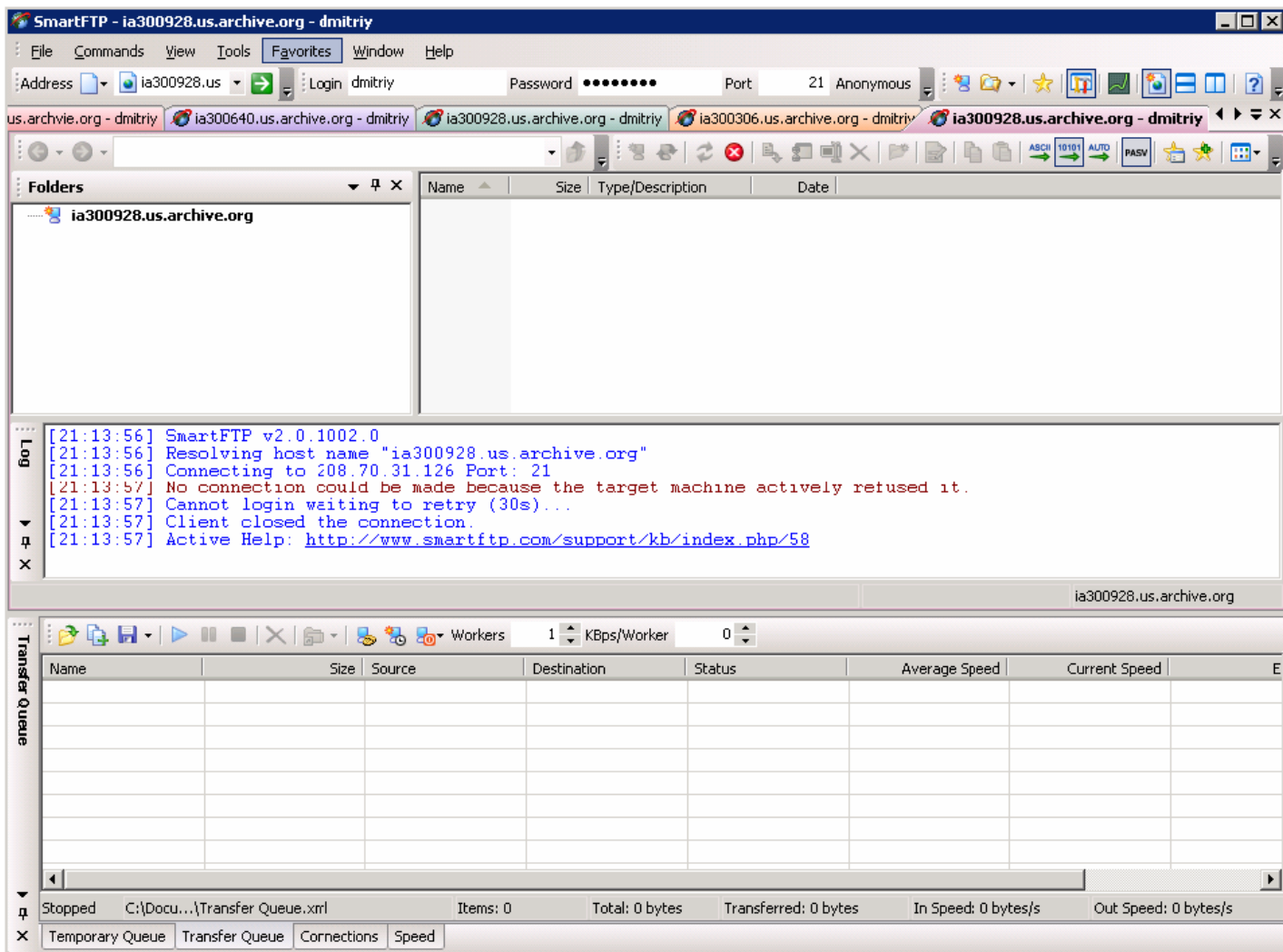
time frame it was decided to use the existing scripts with as little modifications as possible to avoid possible problems and reduce the debug time to minimum. Also, a small Java program was written to complement the script functionality.

A list of nodes which contain ED data was obtained and verified using NodesApplication. Additionally, the PHP script was run several times to ensure that there were no missing nodes. As usually, the list of nodes with ED data was visually inspected for possible incompleteness, as nodes that contain subsets of data are usually named sequentially. Once the list was finalized, another PHP script and the Java program generated queue files only for those IA files whose name starts with “ED_amzn”.

All ED_amzn ARC files that were available were downloaded. 1690 ARC files were transferred to the weblab machine from 96 nodes. ED_amzn crawl (ARC files) was finished on April 30, 2007. The following nodes contain several ED_amzn files, but they were not connectable:

ia300928.us.archive.org
ia300640.us.archive.org
ia300641.us.archive.org

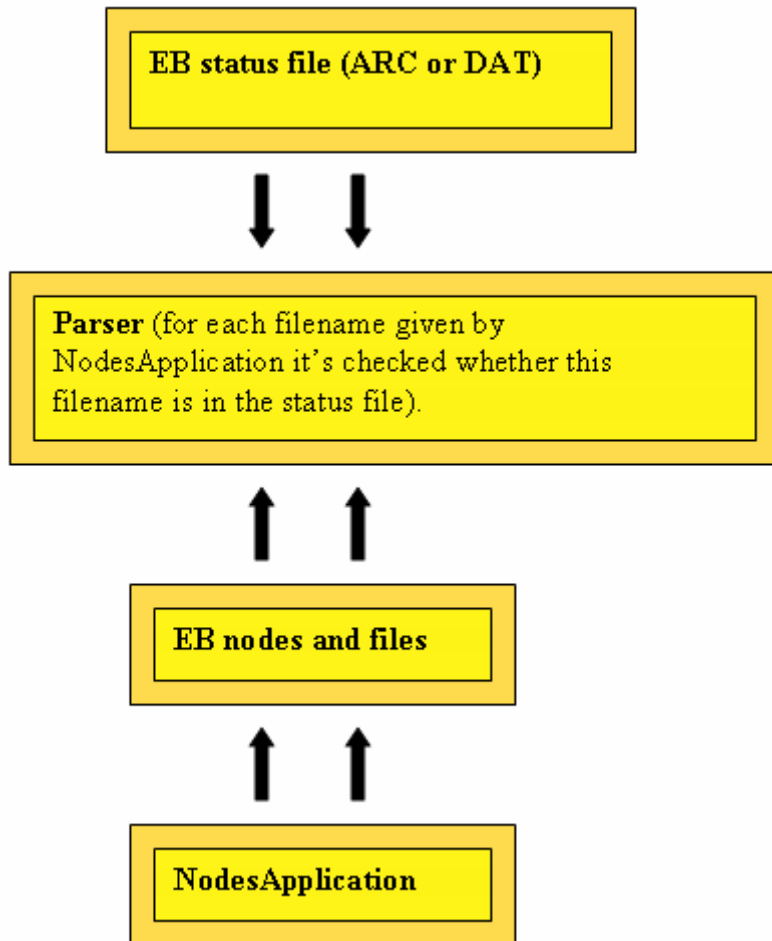
These nodes were accessible via an Internet browser. However, when I tried to access them via an FTP client (either FileZilla or SmartFTP), the server was refusing to connect. Multiple attempts were made, and they all failed. The most likely reason is disk failure in the Internet Archive. The following is the screen shot of a SmartFTP connection attempt:



5. EB Crawl Status (verification)

One of the tasks was to verify EB crawl for completeness. To do so, I utilized NodesApplication and the parser that I wrote. The EB crawl data was already removed from the scidata1 machine. However, there were two status files with lists of EB files that had been downloaded to the scidata1 machine before. Two status files (for ARC and DAT files) located on scidata1 were parsed to check for any missing nodes. NodesApplication generated a list of XML files, each named after an Internet Archive node. Each file contained a list of EB files the corresponding IA node stored. These XML files were used by the parser to determine whether any EB crawl files are missing. The parser was implemented in Java.

The following diagram provides an overview of the steps required to identify missing nodes:



First, NodesApplication is run to obtain XML files. One file is generated for each node that contains EB crawl data. These files contain names of EB files that are stored on the corresponding nodes. The following is a sample file generated by NodesApplication:

```
= <node nodename="ia311104">
<directory>/2/items/EB_binary1_crawl30.20050216082419</directory>
<directory>/2/items/EB_crawl22.20050204085512</directory>
<directory>/2/items/EB_crawl23.20050206142926</directory>
<directory>/2/items/EB_crawl23.20050220122750</directory>
<directory>/2/items/EB_crawl23.20050220150616</directory>
<directory>/2/items/EB_crawl24.20050207235220</directory>
<directory>/2/items/EB_crawl24.20050216013643</directory>
<directory>/2/items/EB_crawl25.20050207213915</directory>
<directory>/2/items/EB_crawl25.20050226043853</directory>
```

```
<directory>/2/items/EB_crawl26.20050219191933</directory>
<directory>/2/items/EB_crawl26.20050226094518</directory>
<directory>/2/items/EB_crawl27.20050206070606</directory>
<directory>/2/items/EB_crawl27.20050222061932</directory>
<directory>/2/items/EB_crawl27.20050302024039</directory>
<directory>/2/items/EB_dad_crawl31.20050223101700</directory>
<directory>/2/items/EB_images_crawl30.20050210054259</directory>
<directory>/2/items/EB_images_crawl30.20050224222032</directory>
</node>
```

Then the parser goes through these XML files and checks whether there is any EB file name that is not in the status file. All the missing filenames and their nodes are recorded. The following EB ARC files were found to be missing:

EB_binary1_crawl30.20050224181445.arc.gz located at node ia300926
EB_crawl24.20050204024821.arc.gz located at node ia300941
EB_images_crawl30.20050223132332.arc.gz located at node ia300941
EB_binary1_crawl30.20050218121805.arc.gz located at node ia300941
EB_crawl23.20050216182611.arc.gz located at node ia300941
EB_crawl23.20050217012404.arc.gz located at node ia300941
EB_crawl25.20050221180455.arc.gz located at node ia300942

No EB DAT files were found to be missing.

6. File Name Patterns

Another task was to identify filename patterns. While ARC and DAT file names usually follow the standard syntax (crawl name followed by “_crawl” followed by a unique identifier), there is a large number of files whose name contains a specific string, such as “slash” or “amzn.” Such files belong to a common source and may be of special interest to researchers. After going through a random selection of Internet Archive nodes, a list of patterns was constructed:

a) Most frequent patterns:

alexa: DI_alexa0.20020106191031

amzn: ED_amzn_crawl31.20050412235151

arc: DD_arc18.20010223050729 (Inside this directory there is a corresponding ARC file DD_arc18.20010223050729.arc.gz and a corresponding DAT file DD_arc18.20010223050729.dat.gz. I have not observed a 'dat' pattern, only an 'arc' pattern).

binary: ED_binary1_crawl30.20050404015418

dad: ED_dad_crawl31.20050801234228

images: ED_images_crawl30.20050405233934

slash: ED_slash_crawl32.20050324025042

b) Rare patterns:

binary_3_0: EE_binary_3_0_crawl25_.20050818022114 (there is a dash between the keyword 'binary' and the following number, unlike previously observed binary names).

binary_8_0: EF_binary_8_0_crawl22_.20051023192028

images_2_0: EE_images_2_0_crawl27_.20050912235733

once_crawl: E02_once_crawl7.20020825164654

1_0: EE_1_0_crawl28_.20050825142544

6_0: EF_6_0_crawl25_.20051026101124

11_0: EG_11_0_crawl22_.20051210050256

1h_0141: E02_1h_0141_crawl3.20021107180004

1w_08: E02_1w_08_crawl8.20020808232418

1w_12: E02_1w_12_crawl8.20020905205239

1w_19: E02_1w_19_crawl5.20021027020040

7. Download Problems Encountered

Occasionally FileZilla instances closed with an exception. In such cases queue files had to be reuploaded and processed again to verify the completion.

Sometimes FileZilla instances could not access a file from the queue and marked it as "too many retries." The status of such queue files had to be reset manually so that the files could be downloaded.

It looks like the Internet Archive staff continued to move data around during the semester. For example, on April 21 node ia311432 contained a relatively large set of DV DAT files - 1417. I set these to be downloaded, and all the downloads were successful except for one. On April 22 I used SmartFTP to manually connect to the node and download the missing file. However, I discovered that this node did not contain DV data any more. Neither did the node ia311425, which had contained 1380 DV DAT files on April 21 (all but one were downloaded).

At the end of the semester the F:\ drive became 100% full. Since all the nodes that are being processed are recorded manually in the status file, this did not result in a serious download problem. Downloads had to be paused for a week.

8. Conclusion

Once again, the Semi-Automated System has proven to be a working solution to the complex problem of data transfer. It provides relatively high throughput and is reliable. However, due to the problem of non-responsive nodes, which was in details described in the Web Library: Data Movement Fall 2006 Report by Dmitriy Shtokman, its usage still involves a substantial amount of manual work. Also, logging is done almost entirely manually. During the next semester it is the goal to utilize the Automation System developed by Andrzej Kielbasinski and described in the Data Movement and Tracking Spring 2007 report and Fall 2006 report.

9. Terminology

ARC file:	An archive file with compressed web page data
DAT file:	An archive file with compressed web page metadata
Crawl:	A snapshot of the web.
FileZilla:	An FTP client for Windows.
Internet Archive:	An organization that maintains an archive of the Web.
Node:	A server at the Internet Archive with crawl data. We are primarily interested in so-called SOLO nodes, because each of them contains data that is not replicated at other Internet Archive nodes.
NodesApplication:	A C# program that creates a collection of XML files for the specified crawl. Each XML file in the collection is named after a node storing data for the specified crawl and contains directories located at that node.
SmartFTP:	An FTP client for Windows.

10. Acknowledgements

The Web Lab team wishes to thank the Internet Archive for their continuing assistance and support. This work is funded in part by National Science Foundation grants CNS-0403340, SES-0537606, and IIS-0634677.

I also would like to deeply thank my advisors Ruth Mitchell and William Arms for their helpful guidance and supervision.

11. References

Andrzej Kielbasinski, *Data Movement and Tracking*. May 2007

Dmitriy Shtokman, *Web Library: Data Movement*. Fall 2006 Report. December 2006.

Andrzej Kielbasinski, *Data Movement and Tracking*. December 2006

Sosa, C. B., Jain, P., Shtokman, D., *Web Library: Data Movement*. Spring 2006 Report. May 2006.

Jain, P., Shtokman, D., Tiwari, H., *Data Movement Research Project*. December 2005.

Kohli, S., Sanghi, L., *Data Monitoring and Tracking*. December 2005.

Internet Archive homepage. <http://www.archive.org>

The Web Lab homepage. <http://www.infosci.cornell.edu/SIN/WebLab/>

Appendix I: Cumulative Transfers to Date

CRAWL NAME	Total # of files available (100%)	Total size (TB)	Cumulative (TB)	Cumulative (%)
DP DAT	161,712	0.84	0.84	100
DP ARC	161,712	14.9	14.9	100
DV DAT*	263,259	1.89	1.89	100
DV ARC*	263,259	25.1**	7.36	29
ED ARC *** (amzn files only)	1,690	0.15	0.15	100

* The following nodes contain DV files but are currently down:

ia300928.us.archive.org (623 ARC and 623 DAT files)

ia300640.us.archive.org (184 ARC and 184 DAT files)

ia300641.us.archive.org (223 ARC and 223 DAT files)

ia300539.us.archive.org (1 ARC and 1 DAT file)

** A typical ARC file is of size 100 MB. As of such, an estimated size of 263,259 files is 25.1 TB.

*** The following nodes contain ED amzn files but are currently down:

ia300928.us.archive.org (1 ARC file and 1 DAT file)

ia300640.us.archive.org (9 ARC files and 9 DAT files)

ia300641.us.archive.org (4 ARC files and 4 DAT files)