

Use Python to Combine LST Files into a PDF File with Bookmarks

Hengwei Liu, Hengrui USA, Princeton, New Jersey

ABSTRACT

In a clinical study the programmers may produce hundreds of LST files with SAS® for tables and listings. To make it easier for the study team to review the output, those LST files are often combined into a single file, either in the RTF format with navigation pane or in the PDF format with bookmarks. In this paper we discuss a way to convert those LST files into PDF files and combine the PDF files into a single PDF file with bookmarks. The development is done with Python 3.9.5 in Linux.

INTRODUCTION

When the programmers create the LST files for tables and listings, the LST files need to follow certain format. There is a sample LST file in display 1 in the appendix. It is produced with SAS options LS=132 and PS=52. The table number appears on the fourth line. The requirements about the table numbers are that they are numbers less than 99 separated by a dot; the last number can be followed by one of the letters from a to z. The file name may indicate the table number but may also include some other words suggestive of the contents of the table. Table 1 shows some examples.

Table Number and Title	File Name
Table 14.1.1 Maturity of DoR for Cohort 2 subjects (N=42)	t140101.lst
Table 14.2.1 Maturity of DoR for Cohort 2 subjects with prior platinum therapy (N=38)	table14_02_01.lst
Table 14.4.1 Time to Response (Months) by Independent Central Review Full Analysis Set	t14_04_01.lst
Table 14.4.2 Time to Response (Weeks) by Independent Central Review Full Analysis Set Subjects with prior platinum-based therapy	table14_04_02.lst
Table 14.4.2a Time to Response (Weeks) by Independent Central Review Full Analysis Set Cohort 2 Subjects with prior platinum-based therapy	table14_04_02_a.lst
Table 14.4.11 Time to Response (Weeks) by Independent Central Review Full Analysis Set Cohort 1 Subjects with prior platinum-based therapy	table14_04_11.lst
Table 14.4.12 Treatment-Emergent Adverse Events by System Organ Class, Preferred Term, and Worst CTCAE Grade Safety Analysis Set	table14_04_12.lst

Table 1. the Table Number, Title and File Name

To combine these LST files into a single pdf file with bookmarks, we do it in three steps. In the first step we rename each LST file using the numbering and title of the table. In the second step we produce a pdf file for each LST file. In the third step we combine all the pdf files into a single pdf file with bookmarks.

THE DETAILS

STEP 1

We copy the LST files into a working folder and rename the LST files based on the table numbers and titles. The python program finds out the table numbers and titles by reading the lines from the 4th row to the row above the first horizontal line in the LST file.

We add a new numbering in front of the file name. This new numbering is generated by replacing the dot in the table number by a padding zero if the number after the dot is single digit. For example, the file t140101.lst for Table 14.1.1 Maturity of DoR for Cohort 2 subjects (N=42) is renamed as "140101 Table 14.1.1 Maturity of DoR for Cohort 2 subjects (N=42).lst". This is illustrated in the display 2 in the appendix.

The reason we are doing this is that if we sort the files by table numbers, the output will not be correct. The table

14.4.11 will show up before the table 14.4.2. Adding the numbering with padding zeros such as 140411 and 140402 in front of the table numbers fixes this issue.

```
#rename.py
# rename the LST files based on table number

import glob
import os
import re

pattern1 = re.compile(r"[a-z]", re.IGNORECASE) # Compile a case-insensitive regex
read_files = glob.glob("*.lst")

alp='abcdefghijklmnopqrstuvwxyz'

for f in read_files:
    with open(f, "rt") as infile:
        lines=infile.readlines()
        line=lines[3].strip().split()[1]
        g=' '

        for i in range(10,0, -1):
            if re.search('____', lines[i]) != None:
                title_line=i

        title=lines[3]
        if title_line>4:
            for i in range(4,title_line):
                title = title.strip() + ' ' + lines[i].strip()

        if pattern1.search(line) != None: # If a match is found
            result=re.search(r"[a-zA-Z]", line).start()
            g=line[result:]
            nums=line.split(".")
            lin=' '
            for h in nums:
                h=h.lower()

                h=h.strip(alp)
                if len(h)>1:
                    h2=h
                elif len(h)==1:
                    h2='0'+h
                else:
                    h2=''
                lin += h2
            lin += g
            lin1=lin.strip() + ' ' + title.strip() + '.lst'
            os.rename(f,lin1)
```

STEP 2

We convert each LST file into a pdf file. This is done through the fpdf module in Python. Note that in the pdf.cell function we used the parameters w=250 and h=3.4. We got these two number through some trials and errors. They work fine when the LST file is created with the SAS options PS=52 and LS=132.

One pdf file is created for every LST file. For example, for the file “140101 Table 14.1.1 Maturity of DoR for Cohort 2 subjects (N=42).lst”, a pdf file called “140101 Table 14.1.1 Maturity of DoR for Cohort 2 subjects (N=42).pdf” is created.

This is illustrated in display 3 in appendix.

```
#lst2pdf.py
# convert each LST file to a PDF file
```

```

import glob
import os
from fpdf import FPDF

read_files2 = glob.glob("*.lst")
read_files2.sort()

for f in read_files2:
    lin=os.path.basename(f)
    lin2=lin.replace('lst','pdf')
    pdf = FPDF(orientation='L')
    pdf.add_page()
    pdf.set_font("Courier", size = 8)
    infi=open(f,'r')
    for x in infi:
        pdf.cell(250, 3.4, txt = x, ln = 1, align = 'R')
    infi.close()
    pdf.output(lin2)

```

STEP 3

We combine the multiple PDF files into a single PDF and add bookmarks.

For this step we use a program called pdfmerger.py published in github (Kim and Pashkovsky, reference [1]).

The pdf file names are used as bookmarks. We make a minor modification on the program to remove the table numbering with padding zeros added in STEP 1 and the file extension .pdf from the bookmarks.

The bookmark for the file “140101 Table 14.1.1 Maturity of DoR for Cohort 2 subjects (N=42).pdf” is “Table 14.1.1 Maturity of DoR for Cohort 2 subjects (N=42)”.

In the display 4 in the appendix, it is shown how to run the pdfmerger.py to create a single pdf file called total.pdf.

The bookmarks in the file total.pdf are shown in the display 5 in the appendix.

```

#!/usr/bin/python3
#pdfmerger.py
# Original author Nicholas Kim, modified by Yan Pashkovsky
# New license - GPL v3
import os
import sys
import time
from PyPDF2 import utils, PdfFileReader, PdfFileWriter

def get_cmdline_arguments():
    """Retrieve command line arguments."""

    from optparse import OptionParser

    usage_string = "%prog [-o output_name] file1, file2 [, ...]"

    parser = OptionParser(usage_string)
    parser.add_option(
        "-o", "--output",
        dest="output_filename",
        default=time.strftime("output_%Y%m%d_%H%M%S"),
        help="specify output filename (exclude .pdf extension); default is current date/time stamp"
    )

    options, args = parser.parse_args()
    if len(args) < 2:
        parser.print_help()
        sys.exit(1)

```

```

    return options, args

def main():
    options, filenames = get_cmdline_arguments()
    output_pdf_name = options.output_filename + ".pdf"
    files_to_merge = []

    # get PDF files
    for f in filenames:
        try:
            next_pdf_file = PdfFileReader(open(f, "rb"))
        except(utils.PdfReadError):
            print >>sys.stderr, "%s is not a valid PDF file." % f
            sys.exit(1)
        except(IOError):
            print >>sys.stderr, "%s could not be found." % f
            sys.exit(1)
        else:
            files_to_merge.append(next_pdf_file)

    # merge page by page
    output_pdf_stream = PdfFileWriter()
    j=0
    k=0
    for f in files_to_merge:
        for i in range(f.numPages):
            output_pdf_stream.addPage(f.getPage(i))
            if i==0:
                output_pdf_stream.addBookmark(str(filenames[k]).split(' ',1)[1][::-1].split('.',1)[1][::-1],j)
                j = j + 1
            k += 1

    # create output pdf file
    try:
        output_pdf_file = open(output_pdf_name, "wb")
        output_pdf_stream.write(output_pdf_file)
    finally:
        output_pdf_file.close()

    print ("%s successfully created." % output_pdf_name)

if __name__ == "__main__":
    main()

```

CONCLUSION

Combining multiple LST files into a single pdf file with bookmarks is an important task for programmers. It greatly helps the study team to review the output of a clinical study. The reviewers can open the single pdf file and navigate to each table instead of opening hundreds of tables individually. In this paper we have shown that this can be achieved with Python in three steps.

REFERENCES

[1] Kim and Pashkovsky. Python program pdfmerger.py, available at <https://gist.github.com/dhwajraj/f330ed64915ac98e29ab46966dfc326d>

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Hengwei Liu
Hengrui USA
400 Alexander Park
Princeton, NJ 08540
Email: Hengwei_liu@yahoo.com

Brand and product names are trademarks of their respective companies.

APPENDIX

In the appendix we provide five displays referred to in the paper.

XXXXXXXXXXXXXX, Inc.
 Protocol: XXX-XXX-XXX
 Interim Analysis

Page 1 of 20
 Program Name: demog.sas
 Run datetime: <2019-03-31 12:43>

Table 14.1.2 Demographics and Baseline Characteristics
 Safety Population

	Treatment A (N=xx)	Treatment B (N=xx)
Age		
n	xx	xx
Mean	xx.x	xx.x
Standard Deviation	xx.x	xx.x
Median	xx.x	xx.x
Min, Max	xx.x, xx.x	xx.x, xx.x
Race		
White	xx	xx
Black or African American	xx	xx
Asian	xx	xx
Other	xx	xx
Sex		
Male	xx	xx
Female	xx	xx

Note: xxx
 Program: /xxx/xxx/xxx/xxx/demog.sas

Display 1. Sample LST File

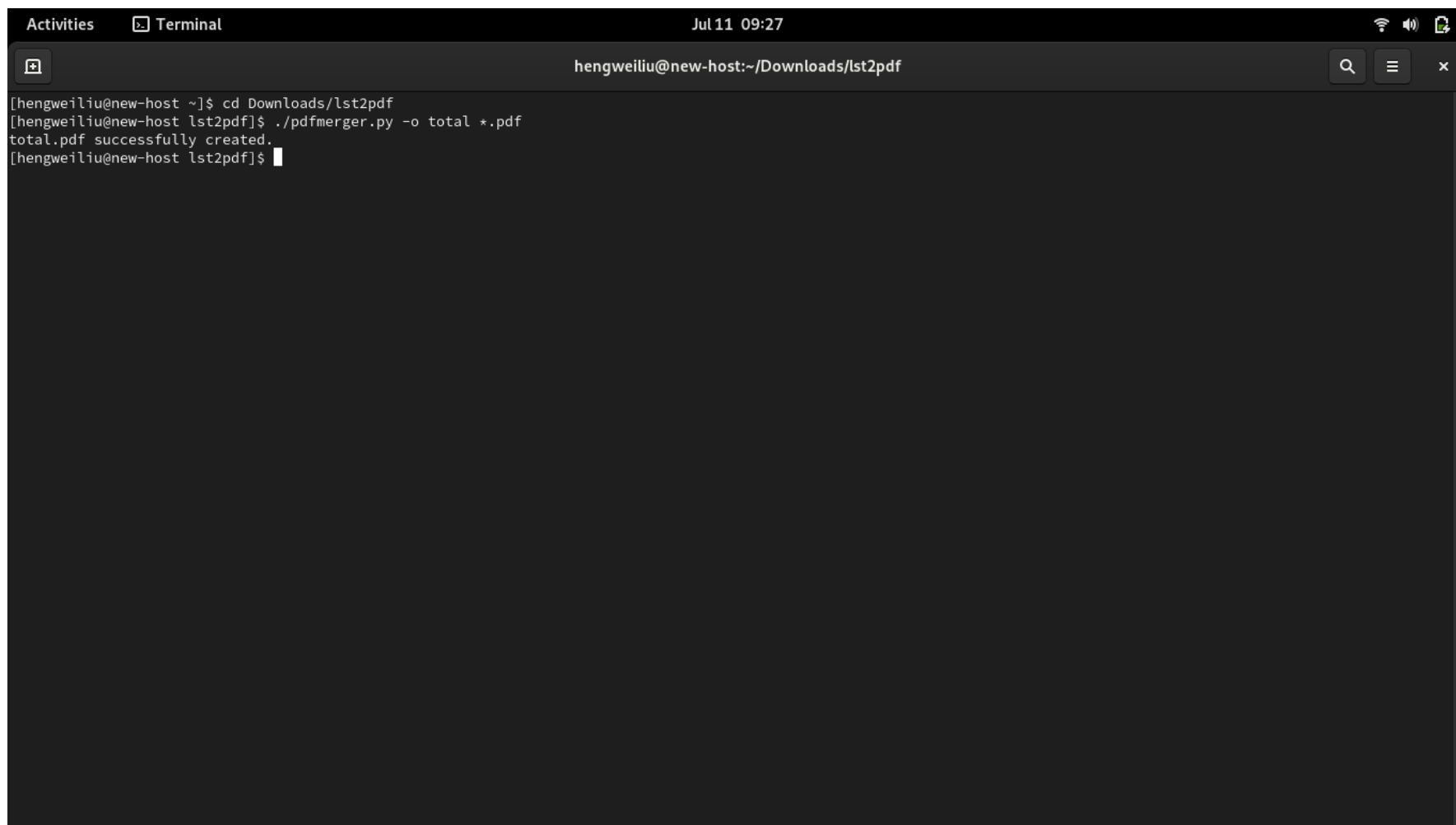
```
Activities Terminal Jul 11 09:24
hengweiliu@new-host:~/Downloads/lst2pdf

[hengweiliu@new-host ~]$ cd Downloads/lst2pdf
[hengweiliu@new-host lst2pdf]$ ls -l
total 596
-rw-rw-r--. 1 hengweiliu hengweiliu 485 Jun 4 23:56 lst2pdf.py
-rwxrw-r--. 1 hengweiliu hengweiliu 1971 Jun 4 23:56 pdfmerger.py
-rw-rw-r--. 1 hengweiliu hengweiliu 1174 Jul 11 09:18 rename.py
-rw-rw-r--. 1 hengweiliu hengweiliu 2563 Jul 11 09:23 t140101.lst
-rw-rw-r--. 1 hengweiliu hengweiliu 3225 Jul 11 09:23 t14_04_01.lst
-rw-rw-r--. 1 hengweiliu hengweiliu 2563 Jul 11 09:23 table14_02_01.lst
-rw-rw-r--. 1 hengweiliu hengweiliu 3367 Jul 11 09:23 table14_04_02_a.lst
-rw-rw-r--. 1 hengweiliu hengweiliu 3357 Jul 11 09:23 table14_04_02.lst
-rw-rw-r--. 1 hengweiliu hengweiliu 3367 Jul 11 09:23 table14_04_11.lst
-rw-r--r--. 1 hengweiliu hengweiliu 572662 Jul 11 09:23 table14_04_12.lst
[hengweiliu@new-host lst2pdf]$ python3 rename.py
[hengweiliu@new-host lst2pdf]$ ls -l
total 596
-rw-rw-r--. 1 hengweiliu hengweiliu 2563 Jul 11 09:23 '140101 Table 14.1.1 Maturity of DoR for Cohort 2 subjects (N=42).lst'
-rw-rw-r--. 1 hengweiliu hengweiliu 2563 Jul 11 09:23 '140201 Table 14.2.1 Maturity of DoR for Cohort 2 subjects with prior platinum therapy (N=38).lst'
-rw-rw-r--. 1 hengweiliu hengweiliu 3225 Jul 11 09:23 '140401 Table 14.4.1 Time to Response (Months) by Independent Central Review Full Analysis Set.lst'
-rw-rw-r--. 1 hengweiliu hengweiliu 3367 Jul 11 09:23 '140402a Table 14.4.2a Time to Response (Weeks) by Independent Central Review Full Analysis Set Cohort 2 Subject
s with prior platinum-based therapy.lst'
-rw-rw-r--. 1 hengweiliu hengweiliu 3357 Jul 11 09:23 '140402 Table 14.4.2 Time to Response (Weeks) by Independent Central Review Full Analysis Set Subjects with prio
r platinum-based therapy.lst'
-rw-rw-r--. 1 hengweiliu hengweiliu 3367 Jul 11 09:23 '140411 Table 14.4.11 Time to Response (Months) by Independent Central Review Full Analysis Set Cohort 1 Subject
s with prior platinum-based therapy.lst'
-rw-r--r--. 1 hengweiliu hengweiliu 572662 Jul 11 09:23 '140412 Table 14.4.12 Treatment-Emergent Adverse Events by System Organ Class, Preferred Term, and Worst CTCAE G
rade Safety Analysis Set.lst'
-rw-rw-r--. 1 hengweiliu hengweiliu 485 Jun 4 23:56 lst2pdf.py
-rwxrw-r--. 1 hengweiliu hengweiliu 1971 Jun 4 23:56 pdfmerger.py
-rw-rw-r--. 1 hengweiliu hengweiliu 1174 Jul 11 09:18 rename.py
[hengweiliu@new-host lst2pdf]$
```

Display 2. The File Names Before and After Running rename.py

```
Activities Terminal Jul 11 09:25
hengweiliu@new-host:~/Downloads/lst2pdf
[hengweiliu@new-host ~]$ cd Downloads/lst2pdf
[hengweiliu@new-host lst2pdf]$ python3 lst2pdf.py
[hengweiliu@new-host lst2pdf]$ ls -l
total 756
-rw-rw-r--. 1 hengweiliu hengweiliu 2563 Jul 11 09:23 '140101 Table 14.1.1 Maturity of DoR for Cohort 2 subjects (N=42).lst'
-rw-rw-r--. 1 hengweiliu hengweiliu 1719 Jul 11 09:25 '140101 Table 14.1.1 Maturity of DoR for Cohort 2 subjects (N=42).pdf'
-rw-rw-r--. 1 hengweiliu hengweiliu 2563 Jul 11 09:23 '140201 Table 14.2.1 Maturity of DoR for Cohort 2 subjects with prior platinum therapy (N=38).lst'
-rw-rw-r--. 1 hengweiliu hengweiliu 1737 Jul 11 09:25 '140201 Table 14.2.1 Maturity of DoR for Cohort 2 subjects with prior platinum therapy (N=38).pdf'
-rw-rw-r--. 1 hengweiliu hengweiliu 3225 Jul 11 09:23 '140401 Table 14.4.1 Time to Response (Months) by Independent Central Review Full Analysis Set.lst'
-rw-rw-r--. 1 hengweiliu hengweiliu 1886 Jul 11 09:25 '140401 Table 14.4.1 Time to Response (Months) by Independent Central Review Full Analysis Set.pdf'
-rw-rw-r--. 1 hengweiliu hengweiliu 3367 Jul 11 09:23 '140402a Table 14.4.2a Time to Response (Weeks) by Independent Central Review Full Analysis Set Cohort 2 Subject
s with prior platinum-based therapy.lst'
-rw-rw-r--. 1 hengweiliu hengweiliu 1921 Jul 11 09:25 '140402a Table 14.4.2a Time to Response (Weeks) by Independent Central Review Full Analysis Set Cohort 2 Subject
s with prior platinum-based therapy.pdf'
-rw-rw-r--. 1 hengweiliu hengweiliu 3357 Jul 11 09:23 '140402 Table 14.4.2 Time to Response (Weeks) by Independent Central Review Full Analysis Set Subjects with prio
r platinum-based therapy.lst'
-rw-rw-r--. 1 hengweiliu hengweiliu 1910 Jul 11 09:25 '140402 Table 14.4.2 Time to Response (Weeks) by Independent Central Review Full Analysis Set Subjects with prio
r platinum-based therapy.pdf'
-rw-rw-r--. 1 hengweiliu hengweiliu 3367 Jul 11 09:23 '140411 Table 14.4.11 Time to Response (Months) by Independent Central Review Full Analysis Set Cohort 1 Subject
s with prior platinum-based therapy.lst'
-rw-rw-r--. 1 hengweiliu hengweiliu 1925 Jul 11 09:25 '140411 Table 14.4.11 Time to Response (Months) by Independent Central Review Full Analysis Set Cohort 1 Subject
s with prior platinum-based therapy.pdf'
-rw-r--r--. 1 hengweiliu hengweiliu 572662 Jul 11 09:23 '140412 Table 14.4.12 Treatment-Emergent Adverse Events by System Organ Class, Preferred Term, and Worst CTCAE G
rade Safety Analysis Set.lst'
-rw-rw-r--. 1 hengweiliu hengweiliu 136554 Jul 11 09:25 '140412 Table 14.4.12 Treatment-Emergent Adverse Events by System Organ Class, Preferred Term, and Worst CTCAE G
rade Safety Analysis Set.pdf'
-rw-rw-r--. 1 hengweiliu hengweiliu 466 Jul 11 09:24 lst2pdf.py
-rwxrwxr--. 1 hengweiliu hengweiliu 1971 Jun 4 23:56 pdfmerger.py
-rw-rw-r--. 1 hengweiliu hengweiliu 1174 Jul 11 09:18 rename.py
[hengweiliu@new-host lst2pdf]$
```

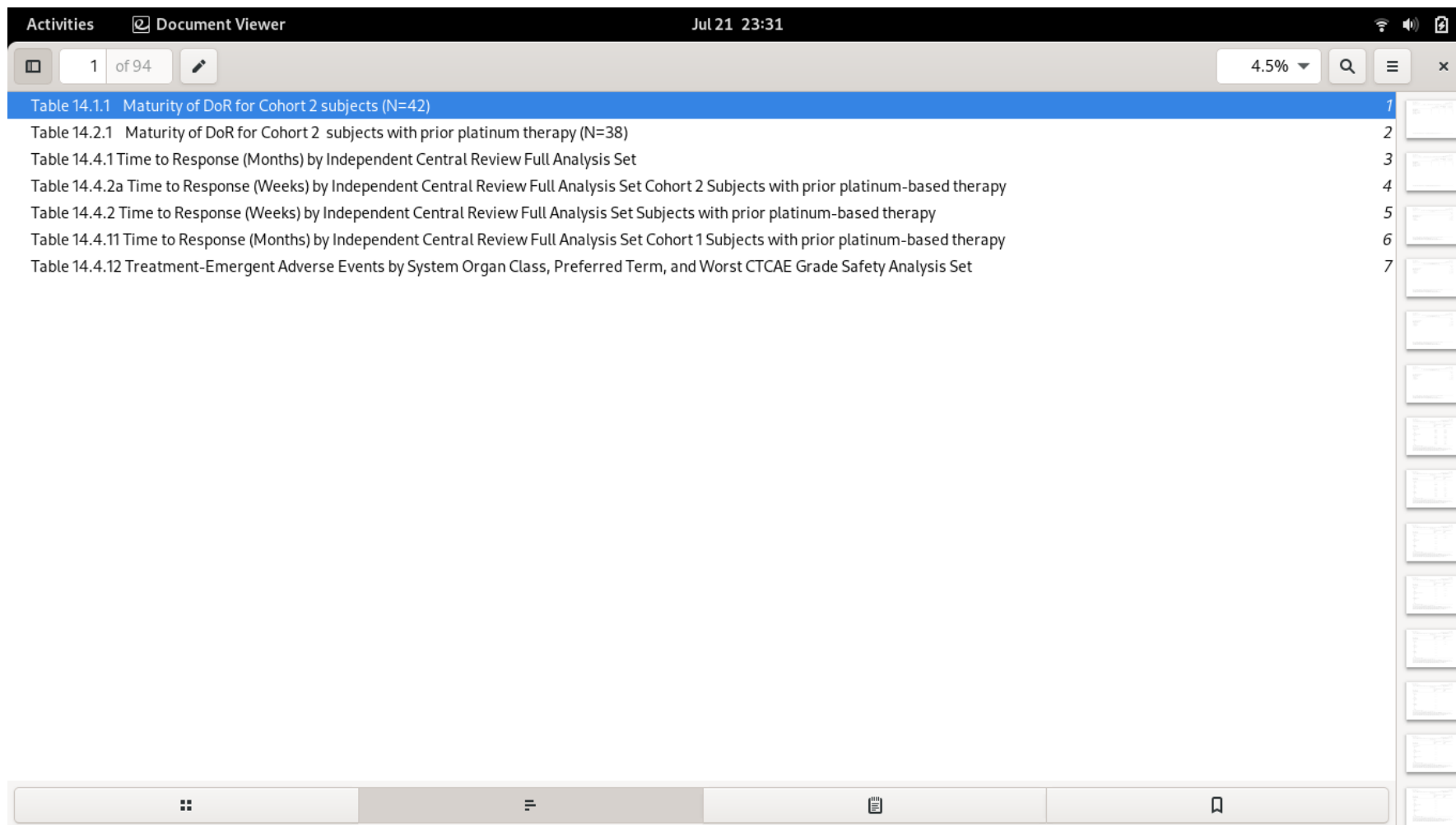
Display 3. Files Names after Running lst2pdf.py



A terminal window titled "Terminal" with a timestamp of "Jul 11 09:27". The window shows a user named "hengweiliu" at a host named "new-host" in the directory "~/Downloads/lst2pdf". The user enters the command `cd Downloads/lst2pdf`, followed by `./pdfmerger.py -o total *.pdf`. The terminal output shows "total.pdf successfully created." and the prompt returns to `[hengweiliu@new-host lst2pdf]$`.

```
[hengweiliu@new-host ~]$ cd Downloads/lst2pdf
[hengweiliu@new-host lst2pdf]$ ./pdfmerger.py -o total *.pdf
total.pdf successfully created.
[hengweiliu@new-host lst2pdf]$
```

Display 4. Run pdfmerger.py to Create the File total.pdf



Display 5. The bookmarks in the File total.pdf