



Generating the Authors index

using PDFMarks

Friday, 4 december 1999

First JACo Workshop on Electronic Publication of Proceedings of Particle Accelerator Conferences

[Jump to first page](#)



Introduction

- Once the complete set of papers has been sorted, numbered and distilled into PDF files...
- Searching for a given paper by a given author through such a huge number of PDF files, without the help of the PDF authors index, would be a pure waste of time...
- But editing this PDF authors index by hand (as it used to be done...) is also a huge amount of work, but it can be done by a computer program...

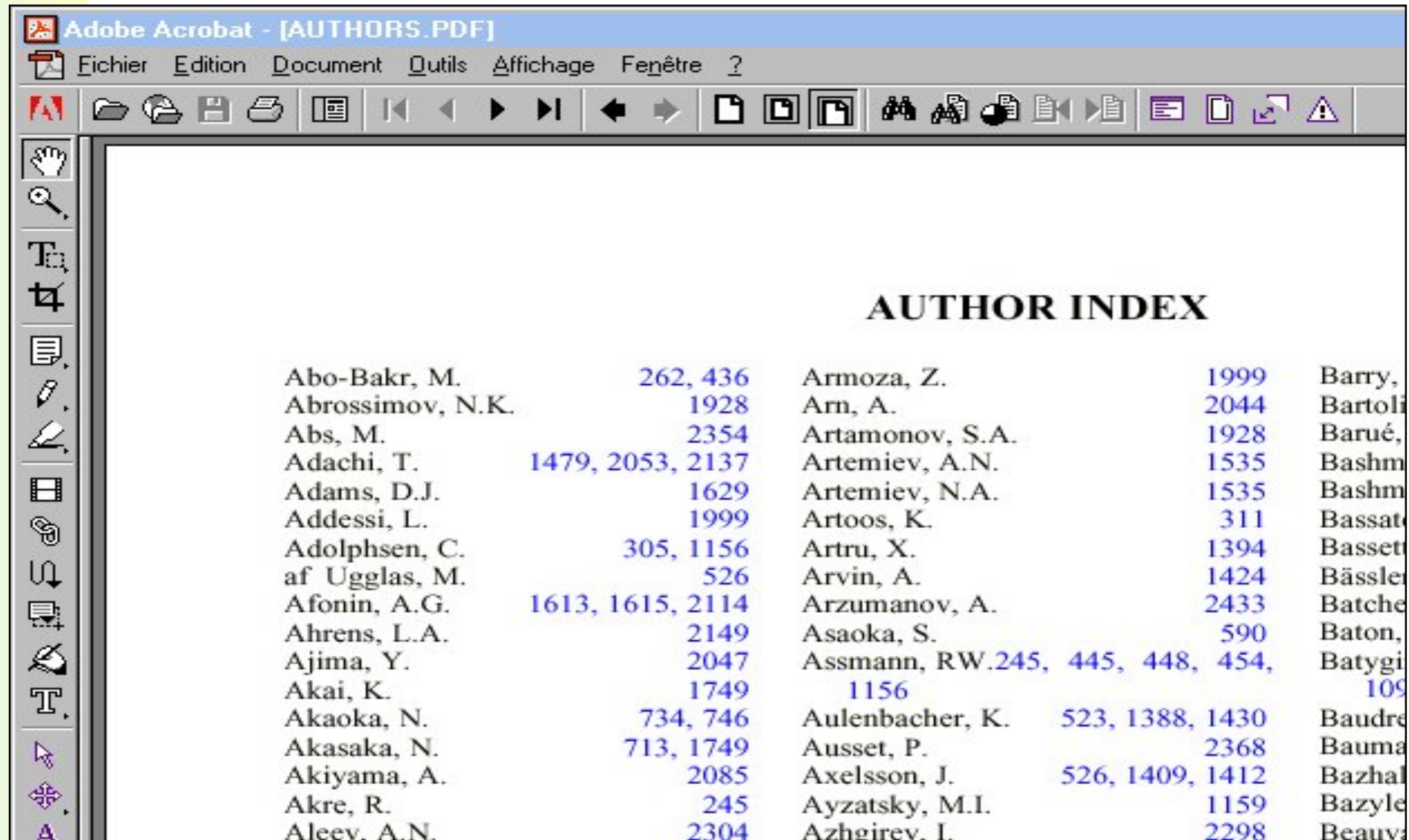


What's a PDF authors index?

- It's a PDF file containing a sorted list of authors followed by a list of “clickable” page numbers corresponding to the papers in which they are involved.
- So, it allows a search by author (sorted alphabetically) to reach papers rapidly by simply clicking on the page numbers.



An example :



The screenshot shows the Adobe Acrobat interface with a document titled 'AUTHORS.PDF'. The main content is an 'AUTHOR INDEX' table of contents. The table lists authors and their corresponding page numbers, arranged in four columns. The authors are listed in alphabetical order across the columns.

AUTHOR INDEX			
Abo-Bakr, M.	262, 436	Armoza, Z.	1999
Abrossimov, N.K.	1928	Arn, A.	2044
Abs, M.	2354	Artamonov, S.A.	1928
Adachi, T.	1479, 2053, 2137	Artemiev, A.N.	1535
Adams, D.J.	1629	Artemiev, N.A.	1535
Addressi, L.	1999	Artoos, K.	311
Adolphsen, C.	305, 1156	Artru, X.	1394
af Ugglas, M.	526	Arvin, A.	1424
Afonin, A.G.	1613, 1615, 2114	Arzumanov, A.	2433
Ahrens, L.A.	2149	Asaoka, S.	590
Ajima, Y.	2047	Assmann, RW.	245, 445, 448, 454, 1156
Akai, K.	1749	Aulenbacher, K.	523, 1388, 1430
Akaoka, N.	734, 746	Ausset, P.	2368
Akasaka, N.	713, 1749	Axelsson, J.	526, 1409, 1412
Akiyama, A.	2085	Ayzatsky, M.I.	1159
Akre, R.	245	Azhgirev, I.	2298
Aleev, A.N.	2304		
			Barry,
			Bartoli
			Barué,
			Bashm
			Bashm
			Bassat
			Basset
			Bässler
			Batche
			Baton,
			Batygi
			109
			Baudre
			Bauma
			Bazhal
			Bazyle
			Beauv



Leif Liljeby's program

- Initially, the authors index used to be made by hand.... What a boring task !
- So, Leif Liljeby had the great idea to implement a short program which performs this job.
- Leif's program makes use of "PDFMarks" for creating the PDF links inside the authors index file.



What's a PDFMark ?

- PDF files are obtained by converting PostScript files using the Acrobat Distiller.
- Nevertheless, PDF can include special features such as links, notes, articles...
which are unknown in the PostScript language.
- So, Adobe invented the PDFMarks operators which can be embedded inside the PostScript to represent these new features.



PDFMark for creating a PDF link

- To create a PDF link, a clickable zone is produced by drawing a rectangle around the word to be linked.
- Then, a linked file also needs to be defined
- To do so, the following lines (the PDFMark operator) must be added to PostScript for each link:

```
[ /Rect [ llx lly urx ury ]  
  /Border [ number number number ]  
  /Action /Launch  
  /File (the path of a file)  
  /SrcPg number  
  /Subtype /Link  
  /ANN pdfmark
```



Explanations

- `[/Rect [llx lly urx ury] :`

Coordinates of the rectangle defining the clickable zone

`llx` : x coordinate of the lower left corner of the rectangle

`lly` : y coordinate of the lower left corner of the rectangle

`urx` : x coordinate of the upper right corner of the rectangle

`ury` : y coordinate of the upper right corner of the rectangle

- `/Border [number number number] :`

Specifies the link's border width

`[0 0 0]` means no border

`[0 0 1]` 1 point width border



- **/Action /Launch :**
Specifies the action type of this PDFMark
/Launch : Launch the application related to the file defined in the /File (the_path_of_a_file) line
- **/File (the path of a file) :**
Specifies the linked file
(eg : /File (CONFERENCE/PAPERS/MOX02A.pdf))
- **/SrcPg number :**
The sequence number of the page on which the link appear. (the first page of a document is page 1, not 0)
If the SrcPg key is present, the PDFMark may be placed anywhere in the PostScript file.



- **/Subtype /Link :**
Must be /Link
- **/ANN pdfmark :**
Specifies the kind of PDFMark.
/ANN is used for notes, and links

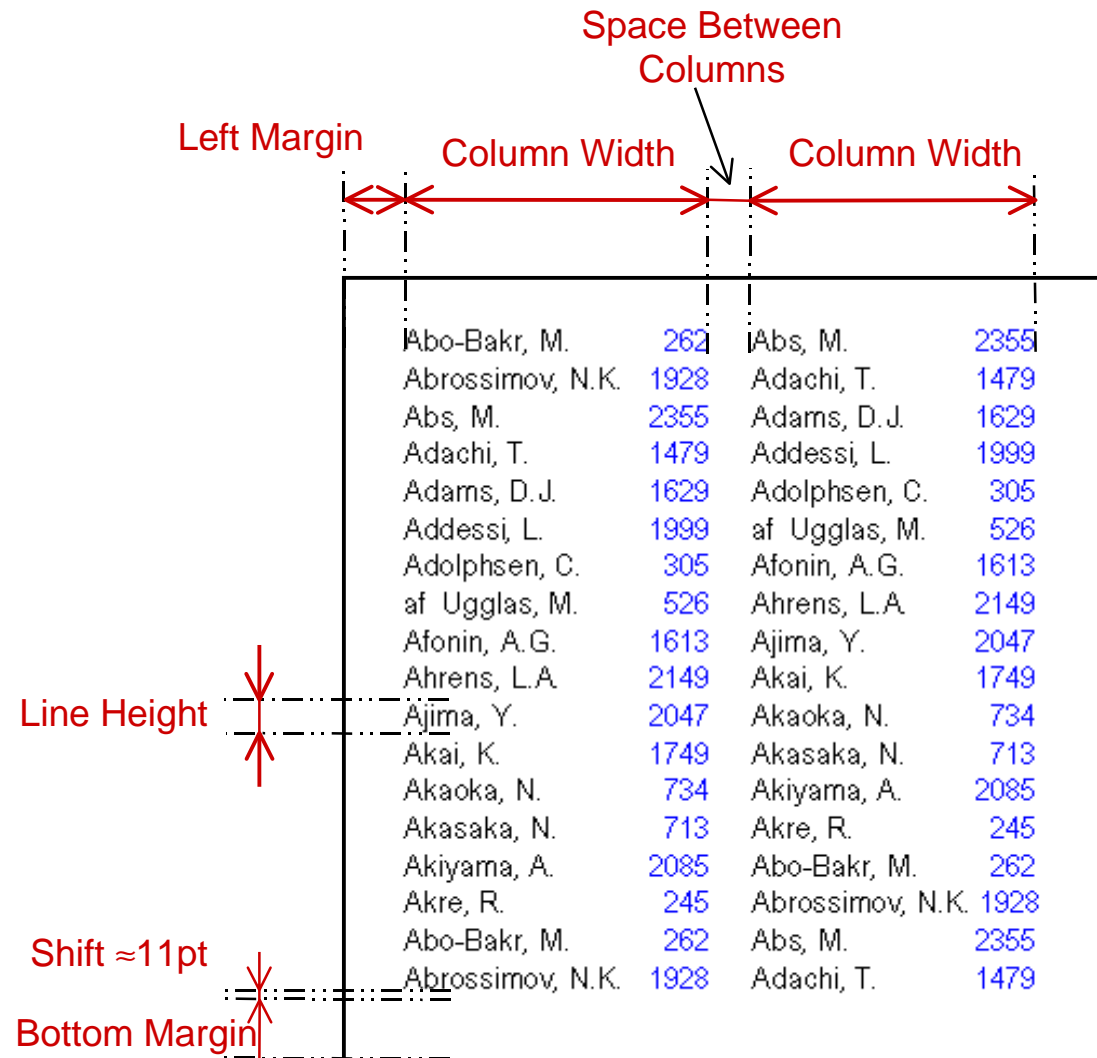


How does Leif's program work ?

- A “layout file” of the index and a “links file” are required as inputs for Leif's program.
These files are generated from the database, once all the papers have been sorted and numbered.
- Then, a few parameters describing the layout file must be set up.



The layout file & the parameters



- Then, this layout file in MS Word format (.doc) is saved as a text file :

Abo-Bakr, M. 262, 436
Abrossimov, N.K. 1928
Abs, M. 2354
Adachi, T. 1479, 2053, 2137
Adams, D.J. 1629
Addessi, L. 1999
Adolphsen, C. 305, 1156
af Ugglas, M. 526
Afonin, A.G. 1613, 1615, 2114
Ahrens, L.A. 2149
Ajima, Y. 2047
...etc



The links file

- The links file looks like this :

```
MOX02A;1  
MOY01A;6  
MOY02A;11  
THX01A;16  
WEX01B;19  
THY01B;24  
WEY02B;29  
...etc
```



The parameters :

The Column number :

Column Number = 2

The number of lines per column :

Line Per Column = 61

The column width :

Column Width = 221 points

The space between two columns :

Space Between Columns = 37 pt

The size of the left margin :

Left Margin = 57 points



The size of the bottom margin :

Bottom Margin = 54 points

The height of a line :

Line Height = 11 points

The width of a character :

Character Width = 5.35 points

Two input files :

Layout File = "layout.txt"

Links File = "link.txt"

One output file :

Output file = "index.out"



The Program's algorithm

- The filenames and page numbers of the links file are extracted

FOR EACH Line OF THE Links File LOOP

 EXTRACT Filename AND Page Number TO PROGRAM

END LOOP



- Then, the treatment of the layout file (.txt) starts :

Line Number = 1

FOR EACH Line OF THE Layout File LOOP

FOR EACH Page Number IN THE Line LOOP

EXTRACT Page Number TO PROGRAM

RETRIEVE Position Of The First Character Of The Page Number IN THE Line TO PROGRAM

RETRIEVE THIS Page Number AMONG Page Numbers EXTRACTED FROM THE LINKS FILE

AND RETRIEVE Filename MATCHING THIS Page Number

- The SrcPg start at 1, and not 0

$\text{SrcPg} = \text{INTEGER PART OF } ((\text{Line Number} - 1) / (\text{Column Number} * \text{Line Per Column})) + 1$

- The column number is between 0 and column_number-1

$\text{Column} = (\text{INTEGER PART OF } ((\text{Line Number} - 1) / (\text{Line Per Column})))$
 $\text{MODULUS Column Number}$

$\text{Line Number In The Column} = ((\text{Line Number} - 1) \text{ MODULUS Line Per Column}) + 1$



Lower Left X =left Margin +(Column * (Column Width + Space Between Columns))
+ Column Width
- ROUND ((Number Of Character in The Line
- Position Of The First Character Of The Page Number) * Character Width)

Lower Left Y = Bottom Margin + (11 + (Line Per Column - Line Number In The Column)
* Line Height)

Upper Right X = Lower left X + ROUND (Number of Character In The Filename
* (Character Width + 0.3))

Upper Right Y = Lower Left Y + 10

```
WRITE TO Output file ( [ /Rect [ Lower Left X Lower Left Y Upper Right X Upper Right Y ] )  
WRITE TO Output file ( /Border [ 0 0 0 ] )  
WRITE TO Output file ( /Action /Launch )  
WRITE TO Output file ( /File "CONFERENCE/PAPERS/" + Filename + ".pdf" )  
WRITE TO Output file ( /SrcPg SrcPg )  
WRITE TO Output file ( /Subtype /Link )  
WRITE TO Output file ( /ANN pdfmark )  
END LOOP  
INCREMENT Line Number  
END LOOP
```



Launching the program...

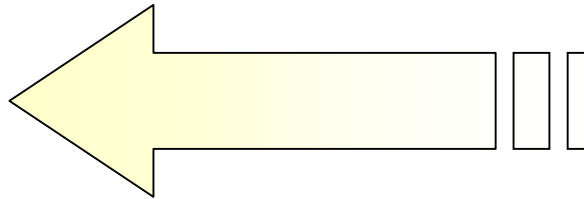
- Once, the two input files are available, and the parameters set up, the program is launched, and after a very short while, the output file is created :

```
[ /Rect [ 235 725 252 735 ]  
  /Border [ 0 0 0 ]  
  /Action /Launch  
  /File (/PAPERS/TUOB03A.pdf)  
  /SrcPg 1  
  /Subtype /Link  
  /ANN pdfmark  
[ /Rect [ 262 725 279 735 ]  
  /Border [ 0 0 0 ]  
  /Action /Launch  
  /File (/PAPERS/WEP30G.pdf)  
  /SrcPg 1  
  /Subtype /Link  
  /ANN pdfmark  
...etc
```



- The MS Word layout file is printed to file so as to obtain its postScript file and the output file is embedded at the end of the postScript file :

```
...
%%Trailer
%%DocumentNeededFonts: Times-Roman
%%DocumentSuppliedFonts:
Pscript_Win_Driver_Incr dup /terminate get exec
savelevel0 restore
%%Pages: 1 [ /Rect [ 113 725 132 735 ]
(%%[ LastPage ]%%) = /Border [ 0 0 0 ]
%%EOF /Action /Launch
- %! /File (pascal/THOA02A.pdf)
/SrcPg 1
/Subtype /Link
/ANN pdfmark
[ /Rect [ 133 725 152 735 ]
/Border [ 0 0 1 ]
/Action /Launch
... ETC
```



- The file is saved, before being distilled.



- And a PDF authors index should be obtained :

Abo-Bakr, M.		262 , 436
Abrossimov, N.K.		1928
Abs, M.		2354
Adachi, T.	1479 , 2053 ,	2137
Adams, D.J.		1629
Addessi, L.		1999
Adolphsen, C.	305 ,	1156
af Ugglas, M.		526
Afonin, A.G.	1613 , 1615 ,	2114
Ahrens, L.A.		2149
Ajima, Y.		2047
Akai, K.		1749
Akaoka, N.	734 ,	746



Known feature (bug !)

- Due to the font used (Times New Roman or any non fixed width font), the links may be shifted to the left and above the page numbers

Abo-Bakr, M. 262, 436
Abrossimov, N.K. 1928





Thank you !

