

**NAME**

pdftosrc – extract source file or stream from PDF file

**SYNOPSIS**

**pdftosrc** *PDF-file* [*stream-object-number*]

**DESCRIPTION**

If only *PDF-file* is given as argument, **pdftosrc** extracts the embedded source file from the first found stream object with /Type /SourceFile within the *PDF-file* and writes it to a file with the name /SourceName as defined in that PDF stream object (see application example below).

If both *PDF-file* and *stream-object-number* are given as arguments, and *stream-object-number* is positive, **pdftosrc** extracts and uncompresses the PDF stream of the object given by its *stream-object-number* from the *PDF-file* and writes it to a file named *PDF-file.stream-object-number* with the ending **.pdf** or **.PDF** stripped from the original *PDF-file* name.

A special case is related to XRef object streams that are part of the PDF standard from PDF-1.5 onward: If *stream-object-number* equals -1, then **pdftosrc** decompresses the XRef stream from the PDF file and writes it in human-readable PDF cross-reference table format to a file named *PDF-file.xref* (these XRef streams cannot be extracted just by giving their object number).

In any case, an existing file with the output file name will be overwritten.

**Notes**

An embedded source file is written unchanged, i.e., it will not be uncompressed.

Only the stream of the object will be written, i.e., not the dictionary of that object.

Knowing which *stream-object-number* to query requires information about the PDF file that has to be gained elsewhere, e.g., by looking into the PDF file with an editor or dumping it with a utility.

The stream extraction capabilities of **pdftosrc** (regarding known PDF versions and filter types, for instance) follow the capabilities of the underlying **xpdf** program version.

Currently the generation number of the stream object is not supported. The default value 0 (zero) is taken.

The wording *stream-object-number* has nothing to do with the ‘object streams’ introduced by the *Adobe PDF Reference*, 5th edition, version 1.6.

**EXAMPLES**

An external file, say *myfile.zip*, can be embedded into a file *foo.pdf* by using pdfTEX primitives, as illustrated by the following example:

```
\immediate\pdfobj
  stream attr {/Type /SourceFile /SourceName (myfile.zip)}
  file{myfile.zip} \pdfcatalog{/SourceObject \the\pdflastobj\space 0 R}
```

Then *myfile.zip* can be extracted from *foo.pdf* by calling "pdftosrc foo.pdf".

**OPTIONS**

None.

**ENVIRONMENT**

None.

## DIAGNOSTICS

If success, the exit code of **pdftosrc** is 0, else 1.

All messages go to stderr. At program invocation, **pdftosrc** issues the current version number of the program **xpdf**, on which **pdftosrc** is based, though it is maintained as part of pdfTEX.

When **pdftosrc** was successful with the output file writing, one of the following messages will be issued:

Source file extracted to *source-file-name*

or

Stream object extracted to *PDF-file.stream-object-number*

or

Cross-reference table extracted to *PDF-file.xref*

When the object given by the *stream-object-number* does not contain a stream, **pdftosrc** issues the following error message:

Not a Stream object

When the *PDF-file* can't be opened, the error message is:

Error: Couldn't open file '*PDF-file*'.

When **pdftosrc** encounters an invalid PDF file, the error message (several lines) is:

Error: May not be a PDF file (continuing anyway)

(more lines)

Invalid PDF file

There are other error messages from **pdftosrc** for various kinds of broken PDF files.

## BUGS

Not all embedded source files will be extracted, only the first one found.

## SEE ALSO

**pdfimages**(1), **pdftex**(1), **pdftotext**(1), **xpdf**(1).

## AUTHORS

**pdftosrc** is part of pdfTEX was written by Hàn The Thành, using **xpdf** functionality from Derek Noonburg. Man page written by Hartmut Henkel.

Public discussion list for pdftosrc: <https://lists.tug.org/pdftex>