NAME

catdvi – a DVI to plain text converter

SYNOPSIS

catdvi [**–d** *debuglevel*, **––debug=***debuglevel*] [**–e** *outenc*, **––output-encoding=***outenc*] [**–p** *pagespec*, **––first-page=***pagespec*] [**–l** *pagespec*, **––last-page=***pagespec*] [**–N, ––list-page-numbers**] [**–s, ––sequential**] [**–U, ––show-unknown-glyphs**] [**–h, ––help**] [**––version**] [**––copyright**] [*dvi-file*]

DESCRIPTION

This manual page documents **catdvi** version 0.14

catdvi reads the DVI (typesetter DeVice Independent) file *dvi-file* and dumps a plain text approximation of the document it describes to stdout. If the argument *dvi-file* is omitted or a dash ('–'), **catdvi** will read from stdin. Several *output encodings* (different character sets of the plain text output) are supported, most notably UTF-8.

The current version of **catdvi** is a work in progress; it may not be robust enough for production use, but already works fine with linear english text. Many mathematical symbols (e.g. the uppercase greek letters) and moderately complex formulae also come out right.

The program needs to read the TFM (Tex Font Metric) files corresponding to the fonts used in the DVI file. These are searched (and, if necessary and possible, created on the fly) through the *Kpathsea* library.

In order to correctly translate a DVI file to text, the *input encoding* of the fonts used in it (i.e. a meaning-preserving mapping from font code points to Unicode) must be known. There are a lot of different font encodings in use. At the time of writing, **catdvi** understands the following input encodings:

'TEX TEXT'
: Knuth's original font encoding, also known as OT1.

'TEX TEXT WITHOUT F-LIGATURES'
: A variant of the above.

'EXTENDED TEX FONT ENCODING – LATIN'
: The Cork encoding, also known as T1.

'TEX MATH ITALIC'
: The encoding of Knuth's math italic fonts, also known as OML.

'TEX MATH SYMBOLS'
: The encoding of Knuth's math symbol fonts, also known as OMS.

'TEX MATH EXTENSION' (most of it)
: The encoding of Knuth's math extension fonts (big operators, brackets, etc.), also known as OMX.

'TEX TYPEWRITER TEXT'
: The encoding of Knuth's typewriter type fonts.

'LATEX SYMBOLS'
: The encoding of the lasy fonts.

Henrik Theilings European currency symbol ('eurosym') font.

'TEX TEXT COMPANION SYMBOLS 1---TS1' (almost everything)
: The encoding of the text companion fonts.

Martin Vogels symbol ('MarVoSym') font.
: Both the 1998 and the 2000 version are supported as far as possible -- about half of the symbols are not representable in Unicode.

'BLACKBOARD'
: The encoding of the blackboard bold math ('bbm') fonts.

All AMS fonts except the Cyrillic ones.
: This includes the AMS math symbols group A and group B, Euler fraktur, Euler cursive, Euler script and Euler compatible extension fonts.

It is impossible to do perfect translation from unmarked-up DVI to plain text, since the former does only describe the layout of a page, and a translator such as this should really know where words and paragraphs end, and more importantly, which glyphs should be aligned vertically and which shouldn't. The current alignment algorithm tries to preserve the relative horizontal positions of word beginnings; this works well in most cases. Word breaks are detected using simple heuristics; paragraphs are not detected at all (and no paragraph fill is attempted).

The price of alignment is that the output will likely be more than 80 columns wide, even though **catdvi** tries very hard not to use more columns than strictly necessary. Output is usually less than 120 columns, almost always less than 132 columns wide. It may be a good idea to switch your terminal to one of these modes if possible.

OPTIONS

The program follows the usual GNU command line syntax, with long options starting with two dashes.

−d *debuglevel*, **−−debug=***debuglevel*

Set the debug output level to *debuglevel* (default is 10). Large values will result in lots of debug output, 0 in none at all. The maximal debug output level currently used is 150.

−e *outenc*, **−−output-encoding=***outenc*

Specify the encoding of the output character set. *outenc* can be one of the numbers or names from the table below. Names are case insensitive. The following output encodings should be available:

0: UTF-8
1: US-ASCII
2: ISO-8859-1
3: ISO-8859-15

The command **catdvi −−help** (see below) will give a more up-to-date list of all compiled-in output encodings. The default encoding is 1.

−p *pagespec*, **−−first-page=***pagespec*

Do not output pages before page *pagespec*. Pages can be specified in three different ways; the first two are exactly the same as for **dvips**(1).

A (possibly negative) number *num* specifies a TeX page number, which is stored as the so-called *count0* value in the DVI file for every page. Plain TeX uses negative page numbers for roman-numbered frontmatter (title page, preface, TOC, etc.) so the *count0* values compare as

-1 < -2 < -3 < ... < 1 < 2 < 3 < ...

There may be several pages with the same *count0* value in a single DVI file. This usually happens in documents with a per-chapter page numbering scheme.

A number prefixed by an equals sign ('=*num*') specifies a physical page, i.e. the *num*-th page appearing in the DVI file. Numbering starts with 1. Note that with the long form of the option you actually need *two* equals signs, one as part of the long option and one as part of the page specification. Example:

catdvi −−first-page==5 foo.dvi

The third form of a page specification, two numbers separated by a colon ('*num1*:*num2*'), is useful for documents with separately-numbered parts, e.g. chapters. It refers to the page with *count0* value equal to *num2* that **catdvi** believes to be in part *num1*. Since those part numbers are not stored in the DVI file, the program has to guess them: an internal *chapter* counter is increased by one every time the *count0* value of the current page is not greater (in above ordering) than that of the previous page. The counter is initialized to 1 if the first page has negative *count0* value and to 0 otherwise. (A document with separately numbered parts will probably have separately numbered frontmatter as well, and then this rule keeps the internal counter equal to real world part numbers.)

−l *pagespec*, **−−last-page=***pagespec*
: Do not output pages after page *pagespec*. Pages are specified exactly as for the **−−first-page** option above.

−N, −−list-page-numbers
: Instead of the contents of pages, output their physical page count, *count0* value and *chapter* count (see the **−−first-page** option above for a definition of these).

−s, −−sequential
: Do not attempt to reproduce the page layout; output glyphs in the order they appear in the DVI file. This may be useful with e.g. multi-column page layouts.

−U, −−show-unknown-glyphs
: Show the Unicode number of unknown glyphs instead of '?'.

−h, −−help
: Show usage information and a list of available output encodings, then exit.

−−version
: Show version information and exit.

−−copyright
: Show copyright information and exit.

ENVIRONMENT

The usual environment variables TFMFONTS, TEXFONTS, etc. for *Kpathsea* font search and creation apply. Refer to the *Kpathsea* documentation for details.

SEE ALSO

xdvi(1), **dvips**(1), **tex**(1), **mktextfm**(1), the *Kpathsea* texinfo documentation, **utf-8**(7).

BUGS

These things do not work (yet):

- No rules are converted.
- Extensible recipes (very large brackets, braces, etc. built out of several smaller pieces) are not properly handled.
- Complicated math formulae are sometimes misaligned (mostly due to lack of appropriate word break heuristics).
- Some fonts and font encodings are not recognised yet.
- Most mathematical symbols have no representation in the available output character sets except Unicode, and hence show up as '?' unless UTF-8 output encoding is selected. A textual transcription would be desirable.

Watch out for these:

- If there is a space where it does not belong or if there is no space where there should be one, report this as a bug (send the DVI file to the **catdvi** maintainer, stating where in the file the bug is seen).

AUTHORS

catdvi was written by Antti-Juhani Kaijanaho <gaia@iki.fi>, based on a skeletal version by J.H.M. Dassen (Ray). Bjoern Brill <brill@fs.math.uni-frankfurt.de> did further improvements and currently maintains the program.

The manual page was compiled by Bjoern Brill, using material written by the first two program authors.