NAME

biblex − lexically analyze BibTeX bibliography data base files

SYNOPSIS

biblex <*infile* >*outfile*
or
biblex *bibfile1 bibfile2 bibfile3 . . .* >*outfile*

DESCRIPTION

biblex converts one or more bibliography data base files in BIBTEX format to a lexical token stream that is convenient for processing by other tools.

The companion **bibunlex**(1) program can be used to recombine such a token stream back into a BIBTEX file.

SCRIBE-format bibliography files can be handled as well if they are first converted to BIBTEX form by **bibclean**(1).

Only minimal checks are made on the correctness of the input stream, and **biblex** will happily carry out a lexical analysis of nonsensical input, without issuing warnings or errors of any kind, other than possible internal string buffer overflow. To verify that **biblex**'s output token stream is meaningful, the input files can be given to **bibparse**(1) for parsing analysis according to a proposed grammar for BIBTEX.

LEXICAL ANALYSIS

biblex produces output in lines of the form

<token-number><tab><token-name><tab>"<token-value>"

Each output line contains a single complete token, identified by a small integer number for use by a computer program, a token type name for human readers, and a string value in quotes.

Special characters in the token value string are represented with ANSI/ISO Standard C escape sequences, so all characters other than NUL are representable, and multi-line values can be represented in a single line.

Here are the token numbers and token type names that can appear in the output:

0 UNKNOWN
1 ABBREV
2 AT
3 COMMA
4 COMMENT
5 ENTRY
6 EQUALS
7 FIELD
8 INCLUDE
9 INLINE
10 KEY
11 LBRACE
12 LITERAL
13 NEWLINE
14 PREAMBLE
15 RBRACE
16 SHARP
17 SPACE
18 STRING
19 VALUE

Programs that parse such output should also be prepared for lines beginning with the warning prefix, %%, or the error prefix, ??, and for ANSI/ISO Standard C line number directives of the form

line 273 "texbook1.bib"

which record the line number and file name of the current input file.

As an example of the use of **biblex**, the UNIX command pipeline

```
biblex mylib.bib | \
  awk '$2 == "KEY" {print $3}' | \
  sed -e 's/"//g' | \
  sort
```

will extract a sorted list of all citation keys in the file *mylib.bib*.

The LITERAL token type is used for arbitrary text that **biblex** does not examine further, such as the contents of a @Preamble{...} or a @Comment{...}.

The UNKNOWN token type should never appear in the output stream. It is used internally to initialize token type variables.

BUGS

Limitations of the **lex**(1) lexical analyzer generator used to construct **biblex** prevent handling of files containing ASCII NUL; that character will be interpreted as an end-of-file condition.

Older versions of **lex**(1) are not *8-bit clean*; they will not reliably handle characters 128–255. This latter deficiency is being remedied by the X/Open Consortium activities to internationalize and standard UNIX applications.

SEE ALSO

bibcheck(1), **bibclean**(1), **bibdup**(1), **bibextract**(1), **bibjoin**(1), **biblabel**(1), **biborder**(1), **bibparse**(1), **bibsearch**(1), **bibsort**(1), **bibtex**(1), **bibunlex**(1), **citefind**(1), **citesub**(1), **citetags**(1), **latex**(1), **scribe**(1), **tex**(1).

X/Open Company, Ltd., *X/Open Portability Guide, XSI Commands and Utilities*, volume 1. Prentice-Hall, Englewood Cliffs, NJ 07632, USA, 1989. ISBN 0-13-685835-X.

AUTHOR

Nelson H. F. Beebe
Center for Scientific Computing
University of Utah
Department of Mathematics, 322 INSCC
155 S 1400 E RM 233
Salt Lake City, UT 84112-0090
USA
Email: `beebe@math.utah.edu`, `beebe@acm.org`, `beebe@ieee.org` (Internet)
WWW URL: `http://www.math.utah.edu/~beebe`
Telephone: +1 801 581 5254
FAX: +1 801 585 1640, +1 801 581 4148