

NAME

cattobib – convert Z39.50 library catalog server data to BibTeX markup

SYNOPSIS

```
cattobib [-?] [-author] [-byxxx] [-CODEN] [-debug] [-editor] [-help] [-ISBN] [-ISSN] [-keep-files]
          [-logfile] [-quiet] [-server name-or-path] [-test] [-title] [-version]
          [-volume]
          search-key-1 search-key-2 search-key-3 ... >BIBTeX-file
```

DESCRIPTION

cattobib converts Z39.50 library-catalog server data to BIB_TE_X markup, using one or more search keys provided on the command line. This allows convenient re-use of publication data, and eases the tedious and error-prone task of creating BIB_TE_X entries for books and other cataloged publications.

The library catalog server can be specified by a command-line option, with the default server being the world's largest library catalog, the US Library of Congress.

ANSI/NISO Standard Z39.50-1995 and ISO Standard 23950:1998 “Information and documentation — Information retrieval (Z39.50) — Application service definition and protocol specification” define a library catalog protocol that allows client programs to communicate with library catalog servers around the world, and retrieve data in a small number of different formats, notably USMARC (*United States MACHine-Readable Cataloging*) and SUTRS (*Simple Unstructured Text Record Syntax*).

OPTIONS

Command-line options may be abbreviated to a unique leading prefix, and may begin with either one or two leading hyphens. Uppercase options may also be spelled in lowercase.

- ?** Give a brief help message on *stdout*, and exit immediately with a successful status code.
- author** Restrict search-key matching to author fields.
- by_{xxx}** Pass this sort-order option to **bibsort**(1). The suffix *xxx* is one of *day*, *label*, *number*, *pages*, *seriesvolume*, *volume*, or *year*, or any other suffix supported by future versions of that program.
- CODEN** Restrict search-key matching to CODEN (Chemical Abstracts periodical number) fields.
- debug** Display on *stderr* the commands to be sent to each Z39.50 catalog server immediately before contacting that server. Display the server session as well on *stderr* when possible, and otherwise, log it to a temporary file whose name is reported on *stdout*.
- editor** Restrict search-key matching to editor fields.
- help** Give a help message on *stdout* describing the options and known Z39.50 servers, and exit immediately with a successful status code.
- ISBN** Restrict search-key matching to ISBN (International Standard Book Number) fields.
Because of its frequency of use, this option may be abbreviated to a single letter, even though it shares a two-character prefix with another option.
- ISSN** Restrict search-key matching to ISSN (International Standard Serial Number) fields.
- keep-files** Preserve intermediate scratch files in */tmp* (or wherever the environment variable *TMPDIR* points). They are named *cattobib.bib.nnn* (raw BIB_TE_X data before several cleanup steps), *cattobib.fifo.nnn* (raw data from Z39.50 server for the last item searched), and *cattobib.label.nnn* (citation-label substitution file), where *nnn* is the process number. Normally, these files are not of interest and are deleted on exit, unless this option is specified.
- logfile *filename*** Log the output on the specified filename, which must be a new file, instead of on *stdout*.
- quiet** Suppress search status messages.

—server *names-or-paths*

Specify a list of one or Z39.50 catalog servers. The *names-or-paths* value is either a space-separated list of paths to particular Z39.50 servers, usually in the form *hostname:portnumber/databasename*, or abbreviations for such servers, as shown below. A regularly-updated directory of Z39.50 servers can be found on the Web at

<http://www.indexdata.dk/targettest/>

This option can be specified as many times as needed, and all specified servers are accumulated into a master list that is searched on completion of command-line processing.

In this list, vertical bars separate alternatives, and asterisk matches any word with that prefix:

| | |
|---------------------|--|
| ALL A* | All Z39.50 servers known to cattobib |
| NATIONAL N* | National libraries and union catalogs |
| alberta ab | University of Alberta |
| amherst umass | University of Massachusetts, Amherst |
| amicus ca | National Library of Canada |
| bibsys | Norwegian Union Catalog |
| bne es | National Library of Spain |
| bnp pt | National Library of Portugal |
| boulder co | University of Colorado, Boulder |
| british br | British Library |
| calgary | University of Calgary |
| columbia cu | Columbia University |
| congress lc loc | US Library of Congress |
| copac uk | COPAC (union of 24 research-university catalogs in the UK and Ireland) |
| denmark dk | Royal Library of Denmark |
| dsb dsl | Danish State Library |
| duke | Duke University |
| florida fl | Florida Center for Library Automation |
| gbv | German Union Catalog |
| marriott ut | University of Utah Marriott Library |
| melbourne | University of Melbourne |
| melvyl cal | University of California MELVYL catalog |
| minn* mn | University of Minnesota |
| mit | Massachusetts Institute of Technology |
| newyork ny | New York University |
| nla au | National Library of Australia |
| nlm nih | National Library of Medicine (US) |
| nlz nz | National Library of New Zealand |
| nls | National Library of Scotland |
| norway no | National Library of Norway |
| nsw unsw | University of New South Wales |
| odense sdu | University of Southern Denmark |
| oxford ox* | Oxford University |
| poland pl | National Library of Poland |
| rlg | Research Libraries Group |
| sudoc abes fr | French Union Catalog |
| sweden se | National Library of Sweden |
| texas tx | University of Texas at Austin |
| toronto | University of Toronto |
| usc | University of Southern California |

A later section provides information about representing accented characters in searches of the

- catalog of the National Library of Poland.
- test** Run in test mode: library-catalog data is supplied on *stdin*, instead of being fetched from a Z39.50 catalog server. This option is primarily intended for the installation-time validation test suite, but can be also use for local testing and tuning of the format-conversion software.
 - title** Restrict search-key matching to title fields.
 - version** Display the version number and revision date on *stdout*, and exit immediately with a successful status code.
The option name may be abbreviated to a single letter.
 - volume** Search for a series and volume, where the search key consists of a series name and a volume number, separated by one or more nonalphanumeric, nonhyphen, nonspace characters.

EXAMPLES**Search the default Z39.50 server for a book by its ISBN:**

```
% cattobib 1-57586-011-2
%% Searching [z3950.loc.gov:7090/Voyager] for [1575860112]: flags = [@attr 1=7]
@Book{Knuth:1999:DT,
  author =      "Donald Ervin Knuth",
  title =       "Digital typography",
  volume =      "78",
  publisher =   "CSLI Publications",
  address =     "Stanford, Calif.",
  pages =       "xv + 685",
  year =        "1999",
  ISBN =        "1-57586-011-2 (cloth), 1-57586-010-4 (paperback)",
  ISBN-13 =     "978-1-57586-011-4 (cloth), 978-1-57586-010-7 (paperback)",
  LCCN =        "Z249.3 .K59 1999",
  bibdate =     "Wed Jun 22 18:49:36 2005",
  bibsource =   "z3950.loc.gov:7090/Voyager",
  series =      "CSLI lecture notes",
  URL =         "ftp://uiarchive.cso.uiuc.edu/pub/etext/gutenberg/;
                http://www.loc.gov/catdir/description/cam029/98027331.html;
                http://www.loc.gov/catdir/toc/cam022/98027331.html",
  acknowledgement = ack-nhfb,
  subject =     "Printing; Data processing; Computerized typesetting;
                Computer fonts; TeX (Computer file); METAFONT",
}
```

Remark: The ISBN is a *unique* identifier assigned to books published throughout the world since about 1972. It consists of ten decimal digits, the last of which may also be the letter X, divided into four hyphen- (or rarely, space-) separated parts: country or language, publisher, book number within the publisher, and a final check digit that can be used to detect invalid ISBNs.

Country/language groups 0 and 1 are English, 2 is French, 3 is German, 4 is Japanese, 5 is Russian, and so on. The Republic of Srpska (1996 population about 1.4 million people) is 99938.

Large publishers have small numbers (e.g., Collins is 00, McGraw-Hill is 07, and Prentice-Hall is 13), and small publishers have big numbers (e.g., Peachpit Press is 938151 and Personal T_EX is 9631044).

When a publisher exhausts its range of book numbers, it gets a new publisher number: for example, O'Reilly Media Inc. is assigned publisher numbers 937175, 56592, and 596.

Because the 10-digit ISBN numbers are rapidly being exhausted, effective 1-Jan-2007, they will no longer be issued, and instead are to be replaced by new 13-digit values based on the *European Article Numbering* (EAN) system. The name for the new system is ISBN-13, and such values are also EAN values.

From version 0.02, **cattobib** output includes both ISBN(-10) and ISBN-13 data, since the latter are beginning to appear in some online bookstore and library catalogs, and some publishers now print them both with the back-cover bar code.

ISBN-13 translations of ISBN-10 data are handled automatically by the **biborder(1)** utility, and consist of the prefix 978- followed by the first 9 digits of the ISBN-10 value with the same (optional) hyphenation as before, followed by a hyphen and a new check digit. The latter is computed by a different algorithm, and will not in general match the tenth digit (the check digit) of the ISBN-10 value.

Search the default Z39.50 server for a book by its title:

```
% cattobib --title 'Digital Typography Sourcebook'
%% Searching [z3950.loc.gov:7090/Voyager] for [Digital Typography Sourcebook]: flags = []
@Book{Bryan:1996:DTS,
  author =      "Marvin Bryan",
  title =       "The digital typography sourcebook",
  publisher =    "Wiley",
  address =      "New York",
  pages =        "xxiv + 384, 3",
  year =         "1996",
  ISBN =         "0-471-14811-3 (paper/CD-ROM)",
  ISBN-13 =      "978-0-471-14811-1",
  LCCN =         "Z250.7 .B79 1996",
  bibdate =      "Wed Jun 22 18:49:36 2005",
  bibsource =    "z3950.loc.gov:7090/Voyager",
  URL =          "ftp://uiarchive.cso.uiuc.edu/pub/etext/gutenberg/;
                  http://www.loc.gov/catdir/bios/wiley047/96013161.html;
                  http://www.loc.gov/catdir/description/wiley033/96013161.html;
                  http://www.loc.gov/catdir/toc/onix04/96013161.html",
  acknowledgement = ack-nhfb,
  subject =      "Computer fonts",
}
```

Search the British Library for the same book:

```
% cattobib --server br --title 'Digital Typography Sourcebook'
%% Searching [z3950cat.bl.uk:9909/BLAC] for [Digital Typography Sourcebook]: flags = []
%% IGNORED: Number of hits: 1, setno 1
...
@Book{Bryan:1997:DTS,
  author =      "Marvin Bryan",
  title =       "The digital typography sourcebook",
  publisher =    "Wiley",
  address =      "New York ; Chichester",
  pages =        "xxiv + 384",
  year =         "1997",
  ISBN =         "0-471-14811-3 (paperback)",
  ISBN-13 =      "978-0-471-14811-1",
  bibdate =      "Wed Jun 22 18:49:36 2005",
  acknowledgement = ack-nhfb,
  subject =      "Computer fonts",
}
```

Search the National Library of Australia for two books by ISBN:

```
% cattobib -q --server au --ISBN 0-06-621285-5 0-19-860702-4
@Book{Winchester:2003:KDW,
  author =      "Simon Winchester",
  title =       "Krakatoa: the day the world exploded, 27 August 1883",
}
```

```

publisher = "HarperCollins Publishers",
address = "New York",
pages = "xvi + 416",
year = "2003",
ISBN = "0-06-621285-5",
ISBN-13 = "978-0-06-621285-2",
bibdate = "Wed Jun 22 18:49:36 2005",
bibsource = "catalogue.nla.gov.au:7090/Voyager",
acknowledgement = ack-nhfb,
remark = "Includes bibliographical references and index.",
subject = "Natural disasters; Indonesia; Krakatoa; Social
          aspects; Volcanoes; Indonesia; Krakatoa; Krakatoa
          (Indonesia); Eruption, 1883",
usmarc-019 = "019 1 $a 24669279",
usmarc-043 = "043 $a a-io---",
usmarc-250 = "250 $a 1st U.S. ed.",
usmarc-984 = "984 $a ANL $c YY 551.2109598 W759",
}

@Book{Winchester:2003:MES,
author = "Simon Winchester",
title = "The meaning of everything: the story of the Oxford
        English Dictionary",
publisher = "Oxford University Press",
address = "Oxford",
pages = "xxv + 260",
year = "2003",
ISBN = "0-19-860702-4 (hbk.), 0-19-860702-4 (hbk.)",
ISBN-13 = "978-0-19-860702-1 (hbk.), 978-0-19-860702-1 (hbk.)",
bibdate = "Wed Jun 22 18:49:36 2005",
bibsource = "catalogue.nla.gov.au:7090/Voyager",
price = "No price",
acknowledgement = ack-nhfb,
remark = "Includes ndex.",
subject = "Oxford English dictionary; Lexicology; History",
usmarc-019 = "019 1 $a 25073662",
}

```

SEARCHING THE NATIONAL LIBRARY OF POLAND

The Polish language uses 22 accented letters that are not available in 7-bit ASCII or its 8-bit ISO 8859-1 extension used for most Western European languages. The Z39.50 output from the National Library of Poland uses the ISO 6937-2 character set, which is described in a character map available on the Web at

<ftp://dkuug.dk/i18n/charmmaps/117>

cattobib handles translation to T_EX of just the parts of that character set that are needed for the Polish accented letters. Input search strings are, however, not in any standard encoding, but instead require an awkward and idiosyncratic representation of the 22 accented letters:

| | |
|--------|-----------------------------------|
| {834}A | LATIN CAPITAL LETTER A WITH ACUTE |
| {834}C | LATIN CAPITAL LETTER C WITH ACUTE |
| {834}E | LATIN CAPITAL LETTER E WITH ACUTE |
| {834}N | LATIN CAPITAL LETTER N WITH ACUTE |
| {834}O | LATIN CAPITAL LETTER O WITH ACUTE |
| {834}S | LATIN CAPITAL LETTER S WITH ACUTE |
| {834}Z | LATIN CAPITAL LETTER Z WITH ACUTE |
| {834}a | LATIN SMALL LETTER A WITH ACUTE |

| | |
|--------|---------------------------------------|
| {834}c | LATIN SMALL LETTER C WITH ACUTE |
| {834}e | LATIN SMALL LETTER E WITH ACUTE |
| {834}n | LATIN SMALL LETTER N WITH ACUTE |
| {834}o | LATIN SMALL LETTER O WITH ACUTE |
| {834}s | LATIN SMALL LETTER S WITH ACUTE |
| {834}z | LATIN SMALL LETTER Z WITH ACUTE |
| {839}Z | LATIN CAPITAL LETTER Z WITH DOT ABOVE |
| {839}z | LATIN SMALL LETTER Z WITH DOT ABOVE |
| {846}A | LATIN CAPITAL LETTER A WITH OGONEK |
| {846}E | LATIN CAPITAL LETTER E WITH OGONEK |
| {846}a | LATIN SMALL LETTER A WITH OGONEK |
| {846}e | LATIN SMALL LETTER E WITH OGONEK |
| {888} | LATIN CAPITAL LETTER L WITH STROKE |
| {888} | LATIN SMALL LETTER L WITH STROKE |

Notice that the *same* input encoding is used for both lowercase and uppercase l-with-stroke. The prefix {834} represents the acute accent, {839} the dot accent, and {846} the ogonek (a hook accent attached near the lower right corner of the letter). Pictures of all of these accented letters are available on the Web at <http://www.eki.ee/letter/chardata.cgi?lang=pl+Polish&script=latin>

Thus, to search for the author name represented in T_EX as Bie{\n}, use the command

```
cattobib --server pl --author "Bie{834}n"
```

For the title represented in T_EX as Wi{\k{e}}{\z}niowie Moskwy, use the command

```
cattobib --server pl --title "Wi{846}e{834}zniowie Moskwy"
```

It would of course be much easier for users to allow matches of unaccented letters with accented ones, but that feature is not supported by the library catalog Z39.50 server. Instead, **cattobib** provides a convenient alternative: T_EX markup for the Polish accented letters is silently translated to the form needed for the National Library of Poland. You can then write the sample search commands as:

```
cattobib --server pl --author "Bie{\n}"
```

```
cattobib --server pl --title "Wi{\k{e}}{\z}niowie Moskwy"
```

Outer braces surrounding accented letters may be omitted: both {\k{a}} and \k{a} are recognized, as are {\z}, \z, {\z}, \z, and so on.

BUGS

No matter which server is selected, library-catalog data tends to be rife with errors like these:

- completely wrong author lists;
- duplicated records, sometimes with minor variations;
- faulty title capitalization;
- incomplete, inaccurate, or missing page numbers;
- incorrect author order;
- mangled and missing accents;
- off-by-one copyright years;
- truncated author lists and titles;
- ...

The best advice to the user is to search three or more catalogs for the same data, and then merge the results, using a majority vote to resolve discrepancies.

When multiple catalogs provide the same data, it may indicate that the data are likely to be reliable. However, the user is warned that libraries around the world share cataloging data, so there may not be as much data independence as might appear from geographically-distant catalogs.

While the conversion of USMARC and SUTRS markup to BIB_TE_X works reasonably well, there are many catalog record types that are not converted. When they are known not to be useful in BIB_TE_X entries, they are silently discarded. Otherwise, **cattobib** preserves them as additional key/value pairs, such as the *usmarc-nnn* keys in the BIB_TE_X output in the **EXAMPLES** section, or else complains about them in diagnostic messages.

cattobib produces only BIB_TE_X @Book{...} entries, even for conference proceedings, for which a @Proceedings{...} entry is required. Library catalog information often does not distinguish between these document types, so the user must convert such entries.

A certain amount of manual cleanup of the BIB_TE_X output is almost always necessary.

ENVIRONMENT VARIABLES

TESTSHRLIBDIR Directory where format-conversion software is stored. This variable is primarily intended for the installation-time validation test suite, but can also be used for testing alternate versions of the software.

TMPDIR Directory where temporary files are stored (default: */tmp*).

SEE ALSO

bibclean(1), **biblabeled(1)**, **biborder(1)**, **bibsort(1)**, **bibtex(1)**, **citesub(1)**, **yaz-client(1)**.

AUTHOR

Nelson H. F. Beebe
University of Utah
Department of Mathematics, 110 LCB
155 S 1400 E RM 233
Salt Lake City, UT 84112-0090
Tel: +1 801 581 5254
FAX: +1 801 581 4148
Email: beebe@math.utah.edu, beebe@acm.org, beebe@computer.org
WWW URL: <http://www.math.utah.edu/~beebe>