A Bibliographer’s Toolbox
Nelson H. F. Beebe
Department of Mathematics
University of Utah
Salt Lake City, UT 84112-0090
USA
A bibliographer’s credo

Bibliographic databases deserve to be widely used, freely shared, and contributed to by many. The time has come to abandon the cryptic reference-list practices of the past that were developed primarily as labor-saving devices, and replace them with accurate, and detailed, reference lists.
# Bibliographic data markup systems

<table>
<thead>
<tr>
<th>System</th>
<th>Version</th>
<th>Developer(s)</th>
<th>Years</th>
</tr>
</thead>
<tbody>
<tr>
<td>bib</td>
<td></td>
<td>Tim Budd, Gary Levin</td>
<td>refer Mike Lesk (1978–82)</td>
</tr>
<tr>
<td>Scribe</td>
<td>(1976–80)</td>
<td>Brian Reid</td>
<td></td>
</tr>
<tr>
<td>BibTeX</td>
<td>(1984)</td>
<td>Oren Patashnik</td>
<td></td>
</tr>
<tr>
<td>Tib</td>
<td>(1986)</td>
<td>Jim Alexander</td>
<td></td>
</tr>
<tr>
<td>Pro-Cite</td>
<td>(1986)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BiblX</td>
<td>(1987)</td>
<td>Rick Rodgers</td>
<td></td>
</tr>
<tr>
<td>EndNote</td>
<td>(1991)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Papyrus</td>
<td>(1990s)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bookends</td>
<td>(2000s)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
@String{j-CACM = "Communications of the ACM"}

@Article{Dijkstra:1968:GSC,
  author = "Edsger Wybe Dijkstra",
  title = "Go to statement considered harmful",
  journal = j-CACM,
  volume = "11",
  number = "3",
  pages = "147--148",
  month = mar,
  year = "1968",
  CODEN = "CACMA2",
  ISSN = "0001-0782",
  note = "This letter inspired scores of others, published mainly in SIGPLAN Notices up to the mid-1980s. The best-known is \cite{Knuth:1974:SPG}.",
}

This letter inspired scores of others, published mainly in SIGPLAN Notices up to the mid-1980s. The best-known is \cite{Knuth:1974:SPG}.
**BIBTEXML project**

**News**

BIBTEXML project at the Swiss Federal Institute of Technology (ETH) in Zürich, Switzerland is back online:

http://bibtexml.sourceforge.net/
Problems: cryptic, deficient, not extensible without major reprogramming
Typesetting process

human → .ltx or .tex
human → .bib

do
  .ltx or .tex → \LaTeX or \TeX → .aux, .dvi
  .aux, .bib → \BIB\TeX → .bbl
  .aux, .bbl, .aux, .bib → \LaTeX or \TeX → .aux, .dvi
  .aux, .bib → \BIB\TeX → .bbl

until (self-consistent (usually 1 to 3 cycles))

Other typesetters (e.g., \texttt{troff}) in principle can be used, since all files are \textit{plain ASCII}.
\bibitem[\protect\citename{Dijkstra, }1968]{Dijkstra:1968:GSC}
Dijkstra, Edsger~Wybe. 1968.
\newblock Go to statement considered harmful.
\newblock \emph{Communications of the ACM},
\textbf{11}(3), 147--148.
\newblock This letter inspired scores of others, published mainly in SIGPLAN Notices up to the mid-1980s. The best-known is
\cite{Knuth:1974:SPG}.

Problem: markup lost (remediable with alternate .bst)
Extended Chicago style: `xchicago.bst, xbbl.sty`

\bibitem\protect\citeauthoryear{Dijkstra}{Dijkstra}{1968}
{Dijkstra:1968:GSC}
% \bblentry{article}
% \bblcite{Dijkstra:1968:GSC}
\bblauthor{Dijkstra, E.~W.} \bblyear{1968}, \bblmonth{March}.
\newblock \bbltitle{Go to statement considered harmful}.
\newblock \em \bbljournal{Communications of the ACM}~/
\bblvolume{11}\penalty0 \bblpages{147--148}.
\newblock \bblnote{This letter inspired scores of others, published mainly in SIGPLAN Notices up to the mid-1980s. The best-known is \cite{Knuth:1974:SPG}.}
\showEXTRA{\showCODEN{\bblCODEN{CACMA2}}
\showISSN{\bblISSN{0001-0782}}}

Practical \TeX{} 2004 talk... – p.10/30
Typesetting a bibliography

All 500 bibliographies (419,000 entries) in the T\TeX\ Users Group and BibNet Project archives are typeset before release:

\documentclass{article}
\begin{document}
\nocite{*}
\bibliographystyle{unsrt}
\bibliography{\jobname}
\end{document}

In practice, I use \texttt{showtags} package, and also include a title-word cross-reference listing.

Master site:

\url{http://www.math.utah.edu/pub/tex/bib/}
BibTeX features

Braces protect proper nouns in titles:

\texttt{title = "The Use of \{Green\} Functions for Modeling Growth of Green Algae",}

\texttt{title = "{\Einschließen der Lösungen von Randwertaufgaben}. (\{German\}) [{Bracketing} Solutions to Boundary Value Problems]",}

\texttt{title = "Instructor’s Manual to Accompany {{\em Physics, by Paul A. Tipler}}"]",}
BibTeX string abbreviations

Consistent string abbreviations for institutions, journals, months, and publishers have many virtues, and can be supplied by software (`publisher.awk`, `journal.awk`).

@String{inst-ANL = "Argonne National Laboratory"}
@String{inst-ANL:adr = "9700 South Cass Avenue, Argonne, IL 60439-4801, USA"}

@String{j-QUEUE = "ACM Queue: Tomorrow’s Computing Today"}

@String{pub-GNU-PRESS = "GNU Press"}
@String{pub-GNU-PRESS:adr = "Boston, MA, USA"}

@Article{label, ..., month = oct, ... }
BibTeX deficiencies

Author/editing naming is more complex than originally planned for:

editor = "Erdős Pál and Min Guo and Eto Kimio and H'án Th'ểnh Th'ân and Arvind and Juan García y Rodriguez",

remark = "Authors listed as: Frank Mittelbach and Michel Goossens

with Johannes Braams, David Carlisle, and Chris Rowley,

and with contributions by Christine Detig and Joachim Schrod."
BibTeX markup extensions

New keys

- abstract: document abstract
- acknowledgement: entry creator credit
- bibdate: date of last change to this entry
- bibsource: bibliographic data source
cross-referenced book page counts
- bookpages
- CRclass: Computing Reviews classification
- CDRnumber: Computing Reviews database number
- CRreviewer: Computing Reviews reviewer name
- CODEN: Chemical Abstracts serial number
- day: publication day
**BibTeX markup extensions (cont.)**

**New keys (cont.)**

- **DOI**
  - Digital Object Identifier

- **ISBN**
  - International Standard Book Number

- **ISSN**
  - International Standard Serial Serial Number

- **LCCN**
  - *U.S. Library of Congress* catalog number

- **MRclass**
  - *Math Reviews* classification

- **MRnumber**
  - *Math Reviews* database number

- **MRreviewer**
  - *Math Reviews* reviewer name

- **price**
  - document price

- **remark**
  - noncitable commentary

- **URL**
  - Uniform Resource Locator
New keys (cont.)

- **ZMclass**: Zentralblatt für Mathematik classification
- **ZMnumber**: Zentralblatt für Mathematik database number
- **ZMreviewer**: Zentralblatt für Mathematik reviewer name

New document types

- `@Periodical{...}`

New styles

- `is-abbrv.bst`, `is-alpha.bst`, `is-plain.bst`, `is-unsrt.bst`, `xchicago.bst`
The bibliographer’s problem

Data and database errors!
emacs templates

Three keystrokes, or selection from a pull-down menu:

@Article{},
  author = "",
  title = "",
  journal = "",
  year = "",
  volume = "",
  number = "",
  pages = "",
  month = "",
  note = "",
  acknowledgement = ack-nhfb,
  bibdate = "Tue Jun 29 11:54:21 2004",
emacs libraries

Emacs Lisp code

- 19,000 lines (1600 in bibtex.el)
- 650 functions
- 120 customization variables

- bibtex-extra.el
- bibtex-sort.el
- bst.el
- bibtex-keys.el
- bibtex-support.el
- btxacctnt.el
- bibtex-labels.el
- bibtex-tools.el
- filehdr.el
- bibtex-misc.el
- bibtex-x.el
- latex.el
- bibtex-mods.el
- bibtex.el
- ltxacctnt.el
- bibtex-regs.el
- bibtools.el
- ltxmenu.el
Easy accent generation

After a base letter, press *single* function key repeatedly until your accent appears:

```
"o" \'o\ .o \=o \H{o}
^o ‘o b{o} c{o} d{o}
~r{o} t{o} u{o} v{o} ~o
```

*Undo* key backtracks in list if you go too far.

The list rotates to put selected item at front for next search.
Easier accent generation

Select language from a menu or command line (\LaTeX entry has full suite):

- Czech
- Danish
- Faroese
- Finnish
- French
- Gaelic
- German
- Greek
- Icelandic
- Irish
- Italian
- \LaTeX
- Latin
- Norwegian
- Polish
- Portuguese
- Romaji
- Romanian
- Spanish
- Swedish
- Turkish
<table>
<thead>
<tr>
<th>emacs toolbar items</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>update citation label table</td>
<td>find-crossref-year-mismatches</td>
</tr>
<tr>
<td>print citation label table</td>
<td>find-duplicate-author-editor</td>
</tr>
<tr>
<td>bibcheck</td>
<td>find-duplicate-pages</td>
</tr>
<tr>
<td>bibparse</td>
<td>find-german-titles</td>
</tr>
<tr>
<td>check-bbl</td>
<td>find-hyphenated-title-words</td>
</tr>
<tr>
<td>check-page-gaps</td>
<td>find-math-prefixes</td>
</tr>
<tr>
<td>check-page-range</td>
<td>find-missing-parbreaks</td>
</tr>
<tr>
<td>chkdelim</td>
<td>find-page-matches</td>
</tr>
<tr>
<td>find-author-page-matches</td>
<td>find-possessive-title-words</td>
</tr>
<tr>
<td>find-braceable-initial-title-words</td>
<td>find-superfluous-label-suffixes</td>
</tr>
</tbody>
</table>
Major \texttt{B Browse\TeX} tools

Programming tools: \texttt{awk}, \texttt{emacs}, HTML, \texttt{lex/flex}, \texttt{yacc/byacc/bison}, ISO Standard C, and C++ compilers

\begin{itemize}
  \item bibcheck
  \item bibjoin \textbf{bibsearch}
  \item citefind
  \item \textbf{bibclean}
  \item biblabel
  \item biblesort
  \item citesub
  \item bibdestingify
  \item biblex
  \item bibsplit
  \item citetags
  \item bibdup
  \item biblook
  \item bibtex \textbf{html-pretty}
  \item bibdestringify
  \item biblex
  \item bibsplit
  \item citetags
  \item bibextract
  \item biborder
  \item bibunlex \textbf{mg}
  \item bibindex
  \item bibparse
  \item bstpretty
  \item texpretty
\end{itemize}

NB: \texttt{bibclean}, \texttt{biblex}, \texttt{bibunlex}, and \texttt{bibparse} are based on \texttt{rigorous} grammar for \texttt{B\Browse\TeX} (Beebe, 1993).
Other tools

**awk** programs  283 files, 122,000 lines
check-bbl  check for bad downcasing of titles
checksum  file header checksums
chkdelim  check for delimiter balance errors
dw  find doubled words
**emacs**  world’s best editor
ispell  GNU spell checker
make  world’s greatest software tool
myspell  NHFB’s spell checker
ref2bib  convert refer files to BibTeX
**spell**  Unix spell checker
Converting Web pages to \TeX

Fetch and store journal Web pages:

    cd nummath
    ./wget.sh
    cd ..
    ./nummath.sh > foo
    emacs foo &

Journal scripts are links to a master 550-line shell script. It handles about 150 journals, with journal or family-specific \texttt{awk} programs (136,000 lines) to convert clean HTML to rough \TeX.

For comparison, \TeX and METAFONT are 20,000 lines of prettyprinted Pascal each.
Converting Web pages (cont.)

The master shell script ends in two Unix pipelines:

```
 eval $PREHTMLFILTER | html-pretty |
 eval $POSTHTMLPRETTYFILTER |
 eval $PREAWKFILTER |
 gawk -f $basename.awk \\
   -v Filename=$f \\
   -v JOURNAL=$JOURNAL \\
   -v Journal=$JOURNAL |
 gawk -f HTML-entity-to-TeX.awk |
 gawk -f iso8859-1-to-TeX.awk |
 $POSTAWKFILTER >$TMPFILE
```
The temporary file is further processed in a second pipeline to produce final clean \texttt{BIB\TeX} output:

\begin{verbatim}
  biblabel \$\$TMPFILE \ |
  citesub -f - \$\$TMPFILE \ |
  bibsort | 
  bibborder | 
  bibclean \$\$BIBCLEANFLAGS \ |
  \$\$POSTPOSTFILTER | 
  \$\$COMMENTFILTER
\end{verbatim}

The 15 tools in these pipelines each do part of the job, do it well, and do it in complete ignorance of all of the others.
Lessons learned

- Write *small* tools that each solve part of the problem
- Do not trust a single source of bibliographic data
- Check, cross-check, test, validate, and then do so again and again
- Share your data
- Grammars, *grammars*, *grammars*
The End

THE BEATLES
JULY/AUGUST 1969