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

- **bib** Tim Budd, Gary Levin/**refer** Mike Lesk (1978–82)
- **Scribe** (1976–80) Brian Reid
- **BibTeX** (1984) Oren Patashnik
- **Tib** (1986) Jim Alexander
- **Pro-Cite** (1986)
- **BibIX** (1987) Rick Rodgers
- **EndNote** (1991)
- **Papyrus** (1990s)
- **Bookends** (2000s)
- **amsrefs** (2000, 2004) Michael Downes,David Jones

BIBTEX markup

```
@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}.",
}
```

XML markup

```
<article>
    <tag>Dijkstra:1968:GSC</tag>
    <author>
        <personalname>Edsger</personalname>
        <middlename>Wybe</middlename>
        <familyname>Dijkstra<familyname>
    </author>
    <journal>&jCACM;</journal>    <volume>11</volume>
    <number>3</number>            <pages>147&ndash;148</pages>
    <month>&mar;</month>          <year>1968</year>
    <CODEN>CACMA2</CODEN>         <ISSN>0001-0782</ISSN>
    <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</cite>.
    </note>
</article>
```

BIBTEXML project

News

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

`http://bibtexml.sourceforge.net/`

bib markup

```
%A Edsger Wybe Dijkstra
%T Go to statement considered harmful
%J Comm. ACM
%V 11
%N 3
%P 147–148
%D March 1968
```

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`	→	BibTeX	→	`.bbl`
`.aux`, `.bbl`,				
`.ltx` or `.tex`	→	LaTeX or TeX	→	`.aux`, `.dvi`
`.aux`, `.bib`	→	BibTeX	→	`.bbl`

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

Other typesetters (e.g., **troff**) in principle can be used, since all files are *plain ASCII*.

BibTeX .bbl file output

```
\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`)

Enhanced BIBTEX .bbl file output

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 (\bblnumber{3}):\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}}}
```

Typesetting a bibliography

All 500 bibliographies (419,000 entries) in the 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 **showtags** package, and also include a title-word cross-reference listing.

Master site:

`http://www.math.utah.edu/pub/tex/bib/`

BIBTEX features

Braces protect proper nouns in titles:

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

title = "{Einschlie{\ss}en der L{\"o}sungen von
         Randwertaufgaben}. ({German}) [{Bracketing}
         Solutions to Boundary Value Problems]",

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 = "Erd{\H{o}}s P{\'a}l and Min Guo and Eto Kimio and
  H{\'a}n Th{\^e}\llap{\raise 0.5ex\hbox{\'{\relax}}} Th{\'a}nh
          and Arvind and Juan Garc{\'\i}a 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
bookpages	cross-referenced book page counts
CRclass	*Computing Reviews* classification
CRnumber	*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 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

BIBTEX markup extensions (cont.)

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 =         "",
  OPTvolume =    "",
  OPTnumber =    "",
  OPTpages =     "",
  OPTmonth =     "",
  OPTnote =      "",
  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-keys.el
bibtex-labels.el
bibtex-misc.el
bibtex-mods.el
bibtex-regs.el
bibtex-sort.el
bibtex-support.el
bibtex-tools.el
bibtex-x.el
bibtex.el
bibtools.el
bst.el
btxaccnt.el
filehdr.el
latex.el
ltxaccnt.el
ltxmenu.el
```

emacs accent input

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.

emacs accent languages

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

emacs toolbar items

- update citation label table
- print citation label table
- bibcheck
- bibparse
- check-bbl
- check-page-gaps
- check-page-range
- chkdelim
- find-author-page-matches
- find-braceable-initial-title-words
- find-crossref-year-mismatches
- find-duplicate-author-editor
- find-duplicate-pages
- find-german-titles
- find-hyphenated-title-words
- find-math-prefixes
- find-missing-parbreaks
- find-page-matches
- find-possessive-title-words
- find-superfluous-label-suffixes

Major BIBTEX tools

Programming tools: **awk**, **emacs**, **HTML**, **lex/flex**, **yacc/byacc/bison**, ISO Standard **C**, and **C++** compilers

bibcheck
bibclean
bibdestringify
bibdup
bibextract
bibindex
bibjoin
biblabel
biblex
biblook
biborder
bibparse
bibsearch
bibsort
bibsplit
bibtex
bibunlex
bstpretty
citefind
citesub
citetags
html-pretty
mg
texpretty

NB: **bibclean**, **biblex**, **bibunlex**, and **bibparse** are based on *rigorous* grammar for BIBTEX (Beebe, 1993).

Other tools

awk programs	283 files, 122,000 lines
check-bbl	check for bad downcasing of titles
checksum	file header checksums Robert Solovay
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 BibTeX

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 **awk** programs (136,000 lines) to convert clean HTML to rough BibTeX.

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
```

Converting Web pages (cont.)

The temporary file is further processed in a second pipeline to produce final clean BIBTEX output:

```
biblabel $TMPFILE |
    citesub -f - $TMPFILE |
        bibsort |
            biborder |
                bibclean $BIBCLEANFLAGS |
                    $POSTPOSTFILTER |
                        $COMMENTFILTER
```

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