Friday January 9, 17:21:05 1998

To the SIGSAM Editors:

I'm writing this letter to draw SIGSAM reader attention to two active and growing bibliography projects, the TeX User Group bibliography archive at

ftp://ftp.math.utah.edu/pub/tex/bib
http://www.math.utah.edu/pub/tex/bib/index.html

and the BibNet Project bibliography archive at

ftp://ftp.math.utah.edu/pub/bibnet
http://www.math.utah.edu/pub/bibnet/index.html

Much of the software that supports the bibliography archive is available at the first two locations above, and important related software can be found at

ftp://ftp.math.utah.edu/pub/mg/mg-1.3x
http://www.math.utah.edu/pub/mg/mg-1.3x/index.html

ftp://ftp.math.utah.edu/pub/sgml
http://www.math.utah.edu/pub/sgml/index.html

These collections are mirrored nightly to several other sites around the world, and traffic in the bibliography archives alone has averaged 111MB daily for the past half year.

The bibliographies in the archives cover recent literature in linear algebra, numerical analysis, symbolic algebra, electronic document production, fonts, typography, and major parts of computer science; some are subject-specific, some are journal-specific, and in the BibNet Project archive, some are author-specific.

For SIGSAM readers, the most noteworthy members of the TUG archive are:

axiom	AXIOM system
issac	International Symposia on Symbolic and Algebraic Computation
jsymcomp	Journal of Symbolic Computation
macsyma	MACSYMA and VAXIMA
maple-extract	Maple system
maple-tech	Maple Newsletter and Maple Technical Newsletter
mathematica	Mathematica system
red-a-f	Part 1 of REDUCE bibliography
red-g-l	Part 2 of REDUCE bibliography
red-m-z	Part 3 of REDUCE bibliography
redbooks	REDUCE books
sigsam	SIGSAM Bulletin

As of early January 1998, the two archive collections contain about 180,000 bibliographic entries from more than 300 bibliography files occupying about 140MB of disk space.

The actual disk space use is significantly more, because these bibliographies are offered in several forms: BibTeX, TeX DVI, HTML, gzipped Portable Document Format (PDF), and gzipped PostScript.

Each bibliography also includes a LaTeX wrapper for typesetting the complete bibliography, including a comprehensive title-word index, and a customized spelling exception dictionary.

For journals, there are also compact tables-of-contents files at

ftp://ftp.math.utah.edu/pub/tex/bib/toc
http://www.math.utah.edu/pub/tex/bib/toc/index.html

The master form of each bibliography is always the BibTeX file; the other forms are derived from it completely automatically.

BibTeX format was chosen because it can be defined by a rigorous grammar, is extensible to new key/value pairs and document types, is readily understood by humans, as well as by simple computer programs, and is automatically reusable in a wide variety of styles. I find it superior to any bibliography format that I've encountered.

It is also extremely important that the BibTeX processor is non-proprietary. BibTeX is defined as a literate program, and the software is freely available on all major computing platforms. Since BibTeX is style-file based, it is possible to write new styles to satisfy new formatting requirements, or to convert BibTeX data to a format expected by another bibliographic software system.

Each file contains a comment header, with author, address, date, version number, checksum, keywords, documentation, and other metadata, all in a standard format. The architecture-independent checksum makes it possible to detect corruption after files have been transferred to other sites.

All of the data files have been under the GNU Revision Control System (RCS) since the projects began, so I have a reliable record of changes, and can retrieve any previously-released version.

The data in these archives has been derived from many sources and databases; the merging process, partly manual and partly automated, increases confidence in the correctness of the data when multiple sources are found to agree. When they do not, best-2-out-of-3 voting for data from major databases, or manual library search of original volumes, is needed to resolve differences. Consequently, I believe that the quality of these bibliographies is notably higher than most other bibliography archives on the Internet.

Uncertain data in the bibliographies is uniformly identified by doubled queries in string values:

```
pages = "123--??",
```

or by BibTeX keys beginning with xx;

```
xxnote = "Check math in title.",
```

All of the bibliographies have been processed many times by software that I've developed, the most important of which is bibclean, which serves as a prettyprinter, to standardize the data format, and as a string value checker. You'll find no instances of publications from the year 190 in these archives!

Because each bibliography is typeset completely before each new version is released, one can be confident that entries are free of TeX errors. No other bibliography archive that I'm aware of has this important property.

That is the announcement, and now on to some comments...

The problem with the bibliographic software used to convert the output of library catalog and database searches to BibTeX is that it is impossible to make it completely automatic and foolproof. This makes it hard to turn it loose on humans who are not prepared to spend some time learning how to use it, and how to recognize its limitations. Natural language processing is frankly a horrid job — I'd much rather parse a rigorously-defined computer language any day!

Current journal databases vary substantially in quality and coverage, but they all suffer from a major problem: irregular to non-existent markup, (sometimes extensive) typographical errors, and missing data.

Most have painful user interfaces that are entirely unsuited to batch processing or automated searching, and none has adequate search limiting capabilities, such as "show me the papers published on Groebner bases in July 1997 by an author whose name starts with P..., or maybe R... [but I cannot remember it completely.]". With the trend to offering flashy Web interfaces, rather than simple telnet access, this situation is getting worse, not better!

Thus, even with software that automatically converts search results from major journal databases to BibTeX form, there is a considerable amount of hand editing to be done.

Still, I'm sometimes able to make considerable strides: the most recent addition to the archive, for the Journal of Mathematical Physics, added 12,064 new entries, and was completed in about 3 days. The fact that the MathSciNet database has very good mathematical markup helped a great deal; the task would have otherwise taken weeks, and therefore, probably not been attempted. Unfortunately, every other database that I have access to completely botches mathematical material, and manual intervention is required to repair their deficiencies.

Another thing that needs to be done, and is impossible with my limited resources, is to fill in the holes in the bibliographies. About two dozen of the bibliography files offer complete coverage of their journals, including ACM Transactions on Mathematical Software, the IBM Systems Journal, and the Journal of the ACM. In the first case, the bibliography even includes keywords, classification codes, abstracts, and URLs for a substantial portion of the entries.

However, for the ACM Special Interest Group publications, coverage by journal databases is sparse to nonexistent, and there are consequently many holes in the bibliography coverage. For that reason, all of the bibliography files contain a section like this one from sigsam.bib:

```
At version 1.28, the year coverage looked like this:
 1972 ( 4)    1981 ( 16)   1990 ( 12)
 1973 ( 1)    1982 ( 15)   1991 ( 29)
 1974 ( 2)    1983 ( 21)   1992 ( 16)
 1975 ( 1)    1984 ( 35)   1993 ( 17)
 1976 ( 2)    1985 ( 26)   1994 ( 7)
 1977 ( 2)    1986 ( 8)    1995 ( 18)
 1978 ( 2)    1987 ( 24)   1996 ( 42)
 1979 ( 3)    1988 ( 16)   1997 ( 55)
 1980 ( 11)   1989 ( 17)
Article: 392
Proceedings: 9
Total entries: 401
```

Most journal databases' coverage tends to start in the 1980s or 1990s, so bibliographyic data for older publications is regrettably sparse. A single individual like me cannot remedy this. However, I believe that the professional organizations responsible for the journal publications can, and should, and in some cases, are doing so. I received word last year that IEEE completed a project to scan and OCR about two million journal pages; I'm now waiting to hear an official announcement that this material is available.

Journal editors can help by initiating their own projects to record correct and comprehensive bibliographic data for their own journals, and make it freely available on the World-Wide Web, preferably months before the printed journal appears.

Indexing of journals is something that has been done poorly, if at all. Many journals lack yearly indexes, and some don't even manage decennial indexes. The Journal of Mathematical Physics now publishes more than 7000 pages a year in thick monthly issues; good luck in finding something there that you saw several months ago!

For the American Mathematical Monthly, which is the journal that spawned my recent indexing work, the last cumulative index was done in 1974: it was a book that indexed the first 80 years, 1894–1974, of the journal. Since then, the editorial staff has spent two to three weeks a year preparing an author index, and a title index (i.e., you can find something, provided that you know the first word of the title). I was able to prepare a comprehensive decennial index in 26 sec of CPU time on my Sun UltraSPARC 170 workstation, and the first cut at that index took only a couple of days of software preparation. More importantly, since the software is independent of the journal, it can be applied immediately to all of the other journals covered by the archives.

The journal indexes and indexing software are still under development, so I'm not announcing their locations here, but the index.html file(s) will eventually contain pointers to the indexes, so drop in for a Web visit from time to time!

Sincerely,

Nelson H. F. Beebe, Center for Scientific Computing
University of Utah, Department of Mathematics,
105 JWB 155 S 1400 E RM 233
Salt Lake City, UT 84112-0090 USA

Email: beebe@{math.utah.edu,{acm,ieee}.org}
http://www.math.utah.edu/~beebe
Tel: +1 801 581-5254
FAX: +1 801 581-4148