BibTeX meets relational databases

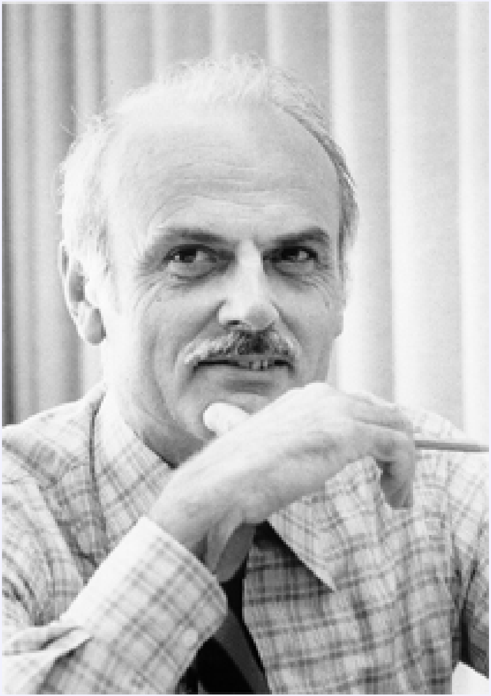
Nelson H. F. Beebe

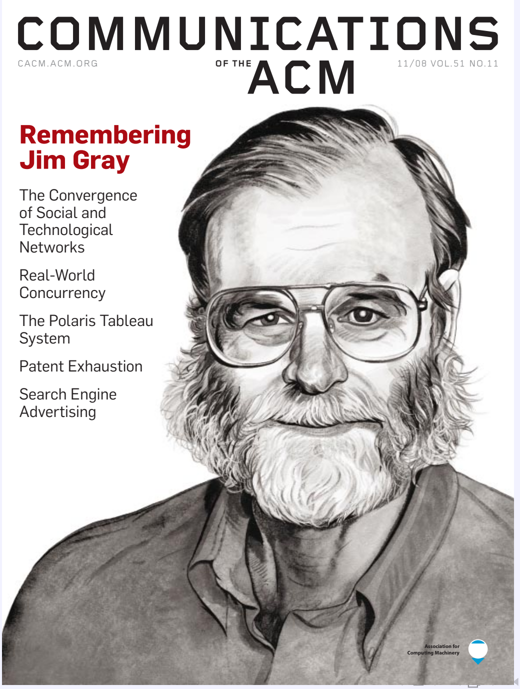
Research Professor
University of Utah
Department of Mathematics, 110 LCB
155 S 1400 E RM 233
Salt Lake City, UT 84112-0090
USA

Email: beebe@math.utah.edu, beebe@acm.org,
beebe@computer.org (Internet)
WWW URL: http://www.math.utah.edu/~beebe
Telephone: +1 801 581 5254
FAX: +1 801 581 4148

29 July 2009

Edgar
Frank
“Ted”
Codd





COMMUNICATIONS OF THE ACM

CACM.ACM.ORG

11/08 VOL.51 NO.11

Remembering Jim Gray

The Convergence of Social and Technological Networks

Real-World Concurrency

The Polaris Tableau System

Patent Exhaustion

Search Engine Advertising

Association for Computing Machinery

BibTeX: a bibliographic database

```
@String{pub-AW      = "Ad{\-d}i{\-s}on-Wes{\-l}ey"}
@String{pub-AW:adr  = "Reading, MA, USA"}
@Book{Graham:1994:CM,
  author =       "Ronald L. Graham and Donald E. Knuth
                 and Oren Patashnik",
  title =        "Concrete Mathematics",
  publisher =    pub-AW,
  address =      pub-AW:adr,
  edition =      "Second",
  pages =        "xiii + 657",
  year =         "1994",
  ISBN =         "0-201-55802-5",
  ISBN-13 =      "978-0-201-55802-9",
  LCCN =         "QA39.2 .G733 1994",
  bibdate =      "Wed Jul 6 14:39:36 1994",
}
```

Relational databases

Reflect BIBTEX entry across its diagonal:

key	**author**	**title**	**year**	**. . .**
`Graham:1994:CM`	Ronald L. Graham and Donald E. Knuth and Oren Patashnik	*Concrete Mathematics*	1994	. . .
. . .				
. . .				

Relational databases: split into key/value tables

key	**author**
`Graham:1994:CM`	Ronald L. Graham and Donald E. Knuth and Oren Patashnik
`Lamport:1994:LDP`	Leslie Lamport
`Knuth:1986:TB`	Donald E. Knuth
...	

key	**title**
`Graham:1994:CM`	*Concrete Mathematics*
`Lamport:1994:LDP`	*LaTeX — A Document Preparation System*
`Knuth:1986:TB`	*The TeXbook*
...	

SQL tables for BibTEX data

A single database can contain multiple tables, and tables can be indexed for rapid access. Tables may be physical data, or logical *views* created from subsets of table data.

For `bibsql`, we have three tables:

`strtab`	BIBTEX `@String{...}` abbreviations
`namtab`	Author/editor names
`bibtab`	BIBTEX fields (`author`, `title`, `year`, ...) and complete entry (`entry`)

Structured Query Language: SQL

S is for Structured, *not* Standard.

Several supported statements, but we often need only `select`:

```
select fieldlist from table
   where     field1 like 'pattern'
         and field2 = 'value2'
         and field3 > 'value3'
   order by field3 desc
   limit n;
```

Sample SQL queries

```
select * from bibtab;

1||9|article|acmturingawards.bib|Perlis:1967:SAS|
Alan J. Perlis|||The Synthesis of Algorithmic Systems||
j-JACM|14||1|||||||19||jan|1|1967|JACOAH|
http://doi.acm.org/10.1145/321371.321372|||00045411
OR 00045411|
||||Mon Dec 05 19:37:58 1994||1994.12.05 19:37:58 ???|
|||||This is the 1966 ACM Turing Award Lecture, and the
first award.||||
@Article{Perlis:1967:SAS,
  author = "Alan J. Perlis",
  title = "The Synthesis of Algorithmic Systems",
  \ldots{}
}|
...
```

Sample SQL queries. . .

```
select year, author, title from bibtab
       where author like ’%Perlis%’ and year = ’1967’;
1967|Alan J. Perlis|The Synthesis of Algorithmic Systems
1967|B. A. Galler and A. J. Perlis|A proposal for definitions

select year, author, title from bibtab
       where author = ’Alan J. Perlis’
       order by year;
1958|Alan J. Perlis|Announcement
1963|Alan J. Perlis|Computation’s development critical to our
1967|Alan J. Perlis|The Synthesis of Algorithmic Systems
...
```

Sample SQL queries. . .

How many variants are there of Guy Steele's name?

```
select count, name from namtab
       where name like '%Steele%'
       order by 1 desc;
15|Guy L. Steele Jr.
3|Guy L. Steele
2|Guy L. Steele, Jr.
1|G. L. Steele, Jr.
1|G. Steele
```

Sample SQL queries. . .

Find five Knuth articles published between 1956 and 1969:

```
select distinct year, author, title from bibtab
       where author like ’%D%Knuth’
       and ’1955’ < year
       and year < ’1970’
       order by year desc
       limit 5;
1969|Donald E. Knuth|Seminumerical Algorithms
1968|Donald E. Knuth|Very magic squares
1967|Donald E. Knuth|The Remaining Trouble Spots in ALGOL 60
1966|Donald E. Knuth|Errata: ‘‘Additional comments on a proble
1966|Donald E. Knuth|Letter to the Editor: Additional comments
```

Sample SQL queries…

What is the percentage of journal articles that have each of one to five authors?

```
select round(100 * count(authorcount) /
       (select count(*) from bibtab
               where authorcount > 0 and
               bibtype = 'article')) || '%',
       authorcount from bibtab
       where authorcount > 0 and bibtype = 'article'
       group by authorcount
       order by count(authorcount) desc
       limit 5;
47.0%|1
29.0%|2
14.0%|3
5.0%|4
1.0%|5
```

Database implementations

- MySQL
- PostgreSQL
- SQLite3
- IBM DB2
- Ingres
- Microsoft SQL Express
- Oracle
- Sybase

All but SQLite3 are client/server databases, and relatively complex to set up and manage. Some are licensed commercial systems ($$$).

SQLite3 requires only one platform independent file, and its software is highly portable and in the public domain.

SQLite3 schemas

```
sqlite> .schema
CREATE TABLE bibtab (
        authorcount  INTEGER,
        editorcount  INTEGER,
        pagecount    INTEGER,
        bibtype      TEXT,
        filename     TEXT,
        label        TEXT,
        author       TEXT,
        ...
        ZMnumber     TEXT,
        entry        TEXT NOT NULL UNIQUE
);
```

SQLite3 schemas . . .

```
CREATE TABLE namtab (
        name            TEXT NOT NULL UNIQUE,
        count           INTEGER
);
CREATE TABLE strtab (
        key             TEXT,
        value           TEXT,
        entry           TEXT NOT NULL UNIQUE
);
CREATE INDEX bibidx on bibtab (author, title, label);
CREATE INDEX bibtimestampidx on bibtab(bibtimestamp);
CREATE INDEX isbn13idx on bibtab (isbn13);
...
```

bibtosql: convert BibTeX entries to database input

```
% bibtosql --help
Usage: /usr/local/bin/bibtosql
       [ --author ]
       [ --create ]
       [ --database dbname ]
       [ --help ]
       [ --version ]
       [ --server ( MySQL | psql | PostgreSQL | SQLite ) ]
       [ -- ]
       BibTeXfiles or <infile
       >outfile

% bibtosql --create *.bib | sqlite3 bibtex.db
```

bibsql: query SQL database

```
% bibsql --help
Usage: /usr/local/bin/bibsql
       [ --author ]
       [ --command ’ command1; command2; ... ’ ]
       [ --database dbname ]
       [ --help ]
       [ --options ’ ... server options ...’ ]
       [ --server ( MySQL | psql | PostgreSQL | SQLite ) ]
       [ --user dbuser ]
       [ --version ]

% bibsql -s psql
psql> ... user input here ...
```

Automating searches

Interfaces to SQL databases are available in common programming and scripting languages.

Sample C code for interfacing to MySQL, PostgreSQL, and SQLite3 is included in the bibsql distribution:

ftp://ftp.math.utah.edu/pub/bibsql/

http://www.math.utah.edu/pub/bibsql/