# EXPLORING AN INTEGRATED DATA BASE STRUCTURE FOR BUILDING ENERGY MONITORING DATA

Jeff Haberl, Vandana Jagannathan, Robert López,
Robert Sparks, Kelly Kissock, Dean Willis, David Claridge
Mechanical Engineering, Texas A&M University
College Station, Texas 77843-3123

**ABSTRACT**

One of the inherent problems with monitoring hourly energy use and environmental conditions in commercial buildings is efficiently processing the "sea" of data that accumulates into an easily understood form. Even when the data exist, building energy analysts generally rely on multiple "flat" ASCII files for storing and retrieving their data only to find that it can take several hours to perform a simple task such as creating a 2-D time series plot of energy use using data from several monitored channels. Integrated data base structures such as relational data bases, if carefully designed, may offer some relief because they can provide the user with an easier access to the data that automatically keeps track of where data are and how to assemble them to satisfy a particular request. Ibis paper presents a brief review of the different types of data required for a large building monitoring project, and the methods that have been developed for acquiring, archiving and retrieving data for the Texas LoanSTAR program, an eight year, $98.6 million revolving loan program for energy conservation retrofits in Texas state, local government and school buildings.

## INTRODUCTION

### The Texas LoanSTAR Program

The Texas LoanSTAR program is an eight year, $98 million revolving loan program for energy conservation retrofits in Texas state, local government and school buildings funded by oil overcharge dollars. The program began in 1988. Public sector institutions participating in the program must repay the loans according to estimated energy savings in four years or less.

As part of this program, a state-wide energy Monitoring and Analysis Program (MAP) was established in 1989. The major objectives of the LoanSTAR MAP are to: 1) verify energy and dollar savings of the retrofits, 2) reduce energy costs by identifying operational and maintenance improvements, 3) improve retrofit selection in future rounds of the LoanSTAR program, and 4) initiate a data base of energy use in institutional and commercial buildings in Texas.

Currently, the program is monitoring 1500+ channels of hourly data from over three dozen buildings, and seventy-five weather stations, using public domain polling procedures that collect information from microcomputer-based field data recorders supplied by several manufacturers. Additional information concerning the program can be found in Verdict et al. (1990), Turner (1990), Nutter et al. (1990), O'Neal et al. (1990), Haberl et al. (1990a), Claridge et al. (1990), Ruch et al. (1991), Ruch and Claridge (1991), and Katipamula and Haberl (1991).

## THE CURRENT LOANSTAR BUILDING ENERGY DATA BASE

### Acquiring and storing the LoanSTAR Data

The monitoring of energy use and environmental conditions in commercial buildings depends on the purpose of the data collection and analysis, the intended use of the data, the type of analysis to be performed, the experiment design, the budget available, and the extent of monitoring and data gathering required (Haberl et al. 1990b).

Because a primary goal of the LoanSTAR program is to verify energy and dollar savings from retrofits, hourly pre-retrofit and post-retrofit energy use data are collected at agencies where the size of the retrofit justifies the cost of the monitoring. Data from each site include whole building energy use and measurements of the primary influencing parameters; submetered data are often collected. Information being gathered for the agencies participating in the LoanSTAR program is represented by three primary types of information as shown in Table 1: hourly data, site description data, and utility billing data. Utility billing data is available for only a few of the sites currently monitored.

Figure 1 illustrates the three primary data paths for the LoanSTAR agencies. Data from remote microprocessor-based data recorders are collected electronically via existing phone lines; data from utility billing records are entered manually for verifying monitored data; and data existing on other computer networks (i.e., the National Weather Service [NWS], and utility load research data) are transferred either via the campus Ethernet or by electronic media.

Data are retrieved from the field recorders by polling with a modem over the phone lines. This process is performed weekly, since the recorders store a fixed amount of data before exhausting their memory and overwriting previously recorded information. The resulting data files, which vary in format, are converted to a standard format and archived into permanent storage on a UNIX-based server as shown in Figure 4. The primary software packages involved in this processing include the Free Software Foundation's GAWK columnar processing software (FSF 1989), Princeton's ARCHIVE program (Feuerman and Kempton 1987), and various statistical routines performed with PRISM (Fels 1986), and SAS (SAS 1990).

Analysis workstations access the data from the server for testing and developing energy use models. The primary products of the analysis consist of weekly verification plots (Figure 2), and six-page monthly agency reports (Figure 3), which include graphical consumption data and a tabular accounting of the energy retrofit savings. Software packages used for graphical reporting include Surfer and Grapher (Golden Software 1990). 3-D plots are produced using SAS and Intex Solution's Lotus add-on product (Intex Solutions 1990; Lotus 1985). The final layout of reports is controlled using TeX (TeX 1990). Browsing of the data is accomplished with Voyager (Lantern 1990).

Weekly polling, archiving and processing of data from over three dozen sites and 75+ weather stations produces 103+ raw files, 500+ verification graphs and over 2 Mbytes of data that must be checked for errors, converted into a standard format and placed on the Unix server. Although more tedious than complex, this task occupies a major portion of the work load for several of the full-time staff.

The current processing scheme relies on multiple, fixed format, "flat" ASCII files and custom batch files written for each site. The batch file shown in Table 2 uses the custom batch files to produce the upper graph shown in Figure 3. All graphic output is produced in a similar fashion. Data from each site is kept in multiple files housed in separate directories on the Unix server. Access is closely controlled by the data base administrator to assure data integrity.

## WHY USE AN INTEGRATED DATA BASE?

Placing the data under the control of an integrated data base system greatly reduces the amount of data handling that would otherwise be performed by a person. This automation allows scientists to spend their time more productively and protects the data from human error. Furthermore, integrated systems often provide manipulation and summary tools that would otherwise have to be created on an ad-hoc basis.

Luckily, for building scientists, several exciting prototypes have been developed that demonstrate the true power of an integrated system and, in our minds, justify the work required to design and develop such a system. One example is the Princeton Boiler Plant Electronic Logbook, a hyper-media platform developed to instantly process and graphically display manual readings from the campus boilers (Haberl et al. 1989; Englander et al. 1990). Another is the Voyager data exploration software developed by Lantern Corporation (Lantern 1990).

The Princeton Logbook successfully combines a graphical user interface with a relational data base management system to provide icon-driven data entry, automatic error checking, data analysis, report generation, and data browsing for daily, monthly and annual boiler data. The Logbook was also designed for use by the boiler plant foreman who had limited previous experience with computers.

The Voyager data exploration software was originally developed to help organize meteorological data, but has proven to have extended capabilities for most kinds of time series and geographically distributed data. Voyager provides multiple window viewing capabilities, and zoom-in capabilities, by utilizing a cross-indexed, compiled data base structure. Interval data in columnar format can be compiled and browsed with this software in a straightforward fashion.

Most of these systems were designed to handle homogeneous data, storing the data from different buildings in different files. Developing a data base for the heterogeneous monitored data from more than three dozen buildings is not as simple. This problem led us to investigate using the *relational model* (Ullman 1988, p.43) for managing the data.

## PROTOTYPE OF A RELATIONAL DATA BASE STRUCTURE FOR HOURLY MONITORED CONSUMPTION DATA.

"A relational data base management system ... is a system that allows users ... to store data in, and retrieve data from, data bases that are perceived as collections of *relations* or tables." (Date 1989, p. 3). Relational data base management systems (RDBMS) can provide an advantage over flat-file ASCII data base systems because the access to and information about the data are carefully controlled by the data base management system. In a carefully implemented system this can allow the user to concentrate on data analysis and data presentation. Complex data queries can be significantly reduced to a series of simple input commands by using a RDBMS. For example, retrieving electricity consumption for air-handling units in buildings of 20,000 square feet and larger for the month of February when temperatures are 65F and greater can be reduced to something as simple as:

```
SELECT BLDG_ID, AHU1, AHU2, AHU3
FROM BLDG_DATA, BLDG_SPEC
WHERE BLDG_SPEC.BLDG_SIZE >= 20000
  AND BLDG_DATA.MONTH = 2
  AND BLDG_DATA.TEMP >=65
```

There are many commercially available data base management systems operating in just about any computing environment. In order to investigate the use of a RDBMS for the LoanSTAR data we are constructing prototypes for a pilot building's data in both the DOS (Borland 1990), and Unix environments (SAS 1990).

One of the first things that needed to be considered was the structure of the individual records in the data base. Table 3 illustrates the different structures that we decided to test. In the upper half of Table 3 is the untrimmed ASCII data structure that we currently use for data storage. Data for all channels are kept in a columnar format which is produced by the ARCHIVE program (Feuerman & Kempton 1987). Each record has a site (building) number, month, day, year, Julian day, decimal date, and hour time stamp followed by the actual channel values. For the month of February, 672 of these records are gathered for each site. For the test site, our site ID is 001, followed by six calendar-related channels and 39 data channels.

The second section in Table 3 is a trimmed traditional RDBMS structure where each hourly value occupied one record, or 26,208 records for February. Each record contains the site ID, start time, stop time, channel number, and value. There is considerably more overhead with a traditional RDBMS (368%). Hence, one must weigh carefully the difference between simplified queries, disk storage space and processing time.

The third section in Table 3 is the trimmed hybrid RDBMS file structure. This table also has 672 records. Each record has a site number, start time, and stop time followed by the 39 data values. The major difference between this record structure and the ASCII record structure is the replacement of the 6 previous calendar stamps with two start-stop stamps. As can be seen in Table 4, the compiled hybrid RDBMS was very similar in size to the flat ASCII file (97%).

The initial tests revealed that a hybrid RDBMS can be constructed which stores hourly building energy monitoring data in about the same space as a flat ASCII file and actually can process the intended graph in less time. A traditional RDBMS will use considerably more disk space to perform the same function in about the same time.

Because we are really interested in storing data for more than one site, file size and processing speed are not the only criteria that must be considered when prototyping a RDBMS for hourly data. Query complexity and data base flexibility must also be considered. For one building, the hybrid RDBMS seems to have the advantage over the traditional RDBMS in file size and processing time. As other buildings are added, the queries using the hybrid RDBMS will become exceedingly complex because each channel must be individually located within the table. The traditional model should not exhibit this complexity.

The hybrid RDBMS structure will also have problems with flexibility. When a new channel is added to a site some time after the data base was assembled, the entire hybrid RDBMS must be reformatted (even prior to the event) to include it. The traditional RDBMS does not suffer from this problem.

## DISCUSSION

This paper has given a brief introduction to the Texas LoanSTAR Monitoring and Analysis Program and a discussion about some preliminary results from the explorations with a RDBMS to capture and store 1,500+ channels of hourly data from 3 dozen sites.

The investigations concerning the design of the RDBMS are far from complete. There appear to be significant differences in files sizes, processing time, query complexity and data base flexibility when one compares a hybrid data base with a traditional data base. One thing that has been learned is that this type of prototyping process must be performed using the actual RDBMS software that is intended for use.

## ACKNOWLEDGEMENTS

This project was funded and supported by the State of Texas, Governor's Energy Office, as part of Texas A&M's LoanSTAR Monitoring and Analysis contract using oil overcharge funds.

## REFERENCES

Borland 1990, *Paradox*, Borland International, Scotts Valley, California.

Claridge, D., Haberl, J., Katipamula, S., O'Neal, D., Ruch, D., Chen, L., Heneghan, T., Hinchey, S., Kissock, K., Wang, J. 1990. "Analysis of Texas LoanSTAR Data", *Proceedings of the Seventh Annual Symposium on Improving Building Systems in Hot and Humid Climates*, Texas A&M University, College Station, Texas, October.

Date, C. 1989. *A Guide to SQL/DS*, Addison Wesley Publishing Co., Reading Massachusetts.

Englander, S., Reynolds, C., Haberl, J. 1990. "The Princeton Boiler Plant Electronic Logbook Project", *Proceedings of the ACEEE 1990 Summer Study on Energy Efficiency in Buildings*, ACEEE, Washington, D.C., August.

Fels, M. (ed.) 1986. "Special Issue Devoted to Measuring Energy Savings, The Princeton Scorekeeping Method (PRISM)", *Energy and Buildings*, Vol. 9, Nos. 1 and 2.

Feuermann, D., Kempton, W. 1987, "ARCHIVE: Software for the Management of Field Data", *Center for Energy and Environmental Studies Report No.216*, (This also includes Tony's Tools, and Art's Tools which are useful columnar data processing tools), Princeton University.

FSF 1989. *GAWK*, Free Software Foundation (PC version of the Unix-based AWK toolkit), 675 Massachusetts Ave., Cambridge, Massachusetts 02139.

Golden 1990. *Grapher* and *Surfer*, Golden Software, 809 14th Street, P.O. Box 281, Golden, Colorado, 80402-0281.

Haberl, J., Englander, S., Reynolds, C., Nyquist, T., McKay, M. 1989. "Whole Campus Performance Analysis Methods", *Sixth Annual Symposium on Improving Building Systems in Hot and Humid Climates*, Texas A&M University, College Station, Texas, October.

Haberl, J., Katipamula, S., Willis, D., Weber, K., Matson, J., Rayaprolu, M., Subramanian, U. 1990a. "The Texas LoanSTAR Program: Acquiring and Archiving LoanSTAR Data", *Proceedings of the Seventh Annual Symposium on Improving Building Systems in Hot and Humid Climates*, Texas A&M University, College Station, Texas, October.

Haberl, J., Claridge, D., Harrje, D. 1990b. "The Design of Field Experiments and Demonstrations", *Proceedings of the IEA Field Monitoring Workshop*, Gothenburg, Sweden, April.

Intex 1990. *3-D Graphics*, Intex Solutions, 161 Highland Ave., Needham, Massachusetts 02194 (requires Lotus 123).

Katipamula, S., K., and Haberl, J., 1991. "A Methodology to Identify Diurnal Load Shapes for Non-Weather Dependent Electric End-Uses", *Solar Engineering, 1991: Proceedings of the ASME-JSES -JSME International Solar Energy Conference*, pp. 457-467, Reno, Nevada, March.

Lantern 1990. *Voyager: Data Exploration,* Lantern Corporation, 63 Ridgemont Drive, Clayton, Missouri 63105 (requires Microsoft Windows).

Lotus 1985, *Lotus 1-2-3 Spreadsheet,* Lotus Development Corp., 55 Cambridge Parkway, Cambridge, Massachusetts 02142.

Nutter, D., Britton, A., Muraya, N., Heffington, W. 1990. "LoanSTAR Energy Conservation Audits: January 1989 - August 1990", *Proceedings of the Seventh Annual Symposium on Improving Building Systems in Hot and Humid Climates,* Texas A&M University, College Station, Texas, October.

O'Neal, D., Bryant, J., Turner, W., Glass, M. 1990. "Metering and Calibration in LoanSTAR Buildings", *Proceedings of the Seventh Annual Symposium on Improving Building Systems in Hot and Humid Climates,* Texas A&M University, College Station, Texas, October.

Ruch, D., Chen, L, Haberl, J., Claridge, D. 1991. "A Change-Point Principal Component Analysis (CP/PCA) Method for Predicting Energy Usage in Commercial Buildings: The PCA Model", *Solar Engineering 1991: Proceedings of the ASME-JSES-JSME International Solar Energy Conference,* pp.441-448, Reno, Nevada, March.

Ruch, D., Claridge, D. 1991. "A Four Parameter Change-Point Model for Predicting Energy Consumption in Commercial Buildings", *Solar Engineering 1991: Proceedings of the ASME-JSES -JSME International Solar Energy Conference,* Reno, Nevada, March.

SAS 1990, *Statistical Analysis Software,* SAS Institute, SAS Circle, Box 8000, Cary, North Carolina.

TeX 1986. *The TEXbook,* Donald Knuth, The American Mathematical Society and the Addison Wesley Publishing Company, Reading, Massachusetts.

Turner, W., D. 1990. "Overview of the Texas LoanSTAR Monitoring Program", *Proceedings of the Seventh Annual Symposium on Improving Building Systems in Hot and Humid Climates,* Texas A&M University, College Station, Texas, October.

Ullman, J., 1988. *Principles of Database and Knowledge-base systems,* Computer Science Press, Rockville Maryland.

Verdict, M., Haberl, J., Claridge, D., O'Neal. D., Heffington, W., Turner, D. 1990. "Monitoring $98 Million in Energy Efficiency Retrofits: The Texas LoanSTAR Program", *Proceedings of the ACEEE 1990 Summer Study on Energy Efficiency in Buildings,* ACEEE, Washington, D.C., August.

***TABLE 1:*** *Typical Flat Files required for LoanSTAR Agencies.* Information being gathered for the agencies participating in the LoanSTAR program is represented by three primary types of information; hourly data, site description data and utility billing data. A listing of the different types of information is given in this table.

| |
|---|
| **HOURLY DATA** |
| 1. Hourly whole-building data |
| 2. Hourly sub-metered data |
| 3. Hourly weather data |
| **SITE DESCRIPTION DATA** |
| 1. General site information |
| 2. Envelope characteristics |
| 3. Equipment inventory |
| 4. Zoning information |
| 5. Scheduling information |
| 6. Equipment control information |
| 7. Site survey and interviews |
| 8. ECRM summary information |
| **UTILITY BILLING DATA** |
| 1. Monthly energy usage |
| 2. Monthly demand data |
| 3. Usage cost information |
| 4. Demand costs |
| 5. Service charges |
| 6. Other charges |

***TABLE 2:*** *Typical ASCII batch processing instructions.* This table contains an example of the ASCII batch processing instructions that are used to produce the electrical consumption plot shown in the upper half of Figure 3. These graphs are produced each month for the agencies participating in the program. Batch files similar to these are assembled for each site.

```
@echo off

if "%1"=="" goto error
if "%2"=="" goto error
if "%3"=="" goto error
if "%4"=="" goto error
if "%5"=="" goto error
if "%6"=="" goto error
if "%7"=="" goto error
if "%8"=="" goto error
if "%9"=="" goto error

rem ***
rem * Put grapher into portrait mode
rem ***
copy \grp\install.por \grp\install.cnf

rem ***
rem * concatenate six weeks of data together for monthly extraction
rem ***
echo Pulling correct information from ACS files
copy \site\%1\acs\%1%4.acs + \site\%1\acs\%1%5.acs + \site\%1\acs\%1%6.acs
%1%2.m1 > null
copy %1%2.m1 + \site\%1\acs\%1%7.acs + \site\%1\acs\%1%8.acs +
\site\%1\acs\%1%9.acs %1%2.m2 >null

rem ***
rem * Fill in missing data with a flag
rem ***
gawk -v site=%1 -f \syu\missing.awk < %1%2.m2 > %1%2.m3

rem ***
rem * Use the SITE SPECIFIC extraction routine to get the correct channels
rem ***
gawk -v mtag=%2 -f %1mext.awk < %1%2.m3 > %1mon.dat
del %1%2.m?

rem ***
rem * Adjust the grapher file to use the correct date stamp
rem ***
echo Configuring grapher files
gawk -v d="%3 1991" -v s=%1 -f makeele.awk < xele.grf > %1mele.grf

rem ***
rem * create the grapher plot file
rem ***
echo Creating plot files
echo ------------------------------------------------------------------
grapher %1mele
echo Complete
echo ------------------------------------------------------------------

echo Creating electrical consumption Postscript file (Sheet 4a)
plot %1mele /s=.8 /b > null

rem echo Determining summary values for Sheet 1
rem gawk -v site=%1 -f sheet1.awk < %1mon.dat > %1sh1.dat
echo Cleaning...
del %1dew.*

goto done

:error
echo                #1  #2      #3     #4     #5     #6     #7     #8     #9
echo Usage: psmon site Mon  Mon Name file#1 file#2 file#3 file#4 file#5 file#6

:done

rem ***
rem * Put grapher back into landscape mode
rem ***
copy \grp\install.lan \grp\install.cnf
```

***TABLE 3:*** *ASCII, traditional and hybrid RDBMS files structures.* This figure contains examples of three different files types: an untrimmed ASCII file (currently in use), a trimmed traditional RDBMS file structure, and a trimmed hybrid RDBMS files structure.

**Untrimmed ASCII file**

```
001 2 1 91 91032 4049.0000  0  356.629  367.680  385.260  80.920  82.627  76.499  39.430  41.439  36.441
47.115  43.599  38.375  9.562  9.864  9.424  0.009  48.089  48.615  -8.928  0.332  38.000  41.000
118.832  53.964  56.427  80.496  20.872  73.975  26.681  503.238  -6.099  0.000  0.000  1.000  0.000
0.000  0.000 4500.000  304.000

001 2 1 91 91032 4049.0417  100  352.109  362.657  380.488  80.116  81.974  75.796  39.405  41.439
36.391  47.492  43.675  38.903  9.562  9.876  9.418  0.009  47.208  50.468  -9.240  0.346  37.000
43.000  118.945  53.795  56.828  80.158  21.122  73.806  26.831  494.620  -6.109  0.000  0.000  1.000
0.000  0.000  0.000 4580.000  303.000

001 2 1 91 91032 4049.0833  200  347.337  357.885  375.465  80.869  82.577  76.449  39.505  41.490
36.492  46.563  43.172  38.451  9.562  9.851  9.418  0.009  44.964  55.776  -9.240  0.346  36.000
44.000  120.294  53.514  56.878  79.821  21.147  74.088  26.831  496.775  -6.128  0.000  0.000  1.000
0.000  0.000  0.000 4644.000  305.000
```

**Trimmed Traditional RDBMS File Structure**

| Site ID | Start Time | Stop Time | Channel Number | Value |
|---|---|---|---|---|
| 001 | 4049.0000 | 4049.0417 | 001 | 356.629 |
| 001 | 4049.0000 | 4049.0417 | 002 | 367.680 |

**Trimmed Hybrid RDBMS File Structure**

| Site ID | Start Time | Stop Time | Channel 1 | .... | Channel 39 |
|---|---|---|---|---|---|
| 001 | 4049.0000 | 4049.0417 | 356.629 | .... | 304.000 |

***TABLE 4:*** *Comparisons of file size and runtime for RDBMS and ASCII files.* This table contains comparisons of file sizes (39 channels, 1 month, 1 site) and runtimes for the current ASCII files, hybrid RDBMS, traditional RDBMS and trimmed RDBMS files. Values in this table were obtained on a 25 Mhz, 386-class PC with a math co-processor. The graph produced in each test is similar to the 2-D electricity graph shown in Figure 3.

| | Size of Text File (in bytes) | Size of Database file (in bytes) (% of ASCII) | Time to Produce the Graph | Number of Records in the Database File |
|---|---|---|---|---|
| ASCII file | 286,279 (100%) | 286,279 (100%) | 3 min 30 sec | 672 |
| Hybrid RDBMS | 204,216 (71%) | 278,528 (97%) | 1 min 00 sec | 672 |
| Traditional RDBMS | 1,124,041 (393%) | 1,918,976 (670%) | 3 min 25 sec | 26,208 |
| Trimmed ASCII file | 253,344 (88%) | 253,344 (88%) | Note 1 | 672 |
| Trimmed Hybrid RDBMS | 253,344 (88%) | 231,424 (81%) | Note 1 | 672 |
| Trimmed Traditional RDBMS | 1,067,820 (373%) | 1,054,720 (368%) | Note 1 | 26208 |

Note 1: Timed tests were not conducted on these files.

***FIGURE 1:*** *Data Paths for the Texas LoanSTAR Program.* The three primary data paths for the LoanSTAR agencies are shown in this figure. Data can be polled electronically, manually entered, or transferred to/from external data bases via the campus Ethernet or electronic media.

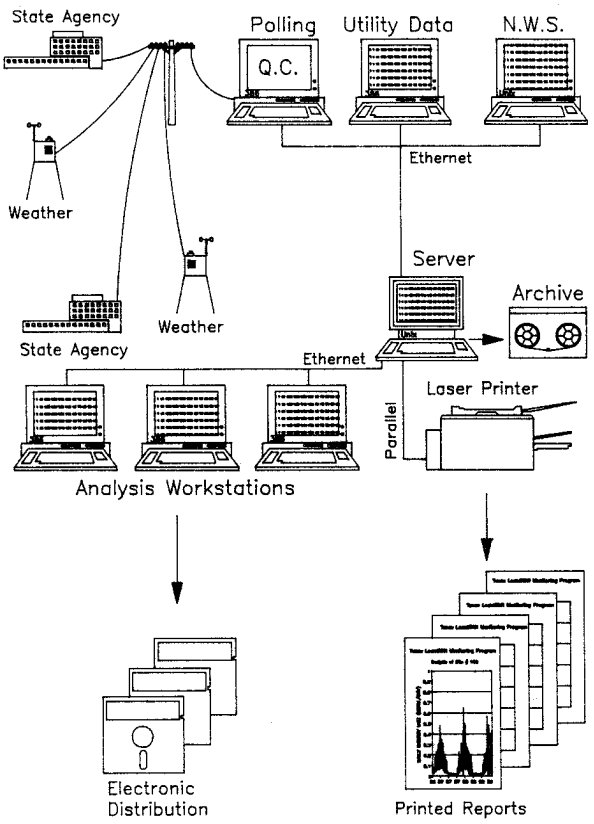


***FIGURE 2:****Typical Weekly Verification Plots for a LoanSTAR Agency.* This figure is an example of the weekly verification plots for a building located at the University of Texas at Arlington. Each graph represents one or more channels of information plotted in time-series or versus another channel of information.

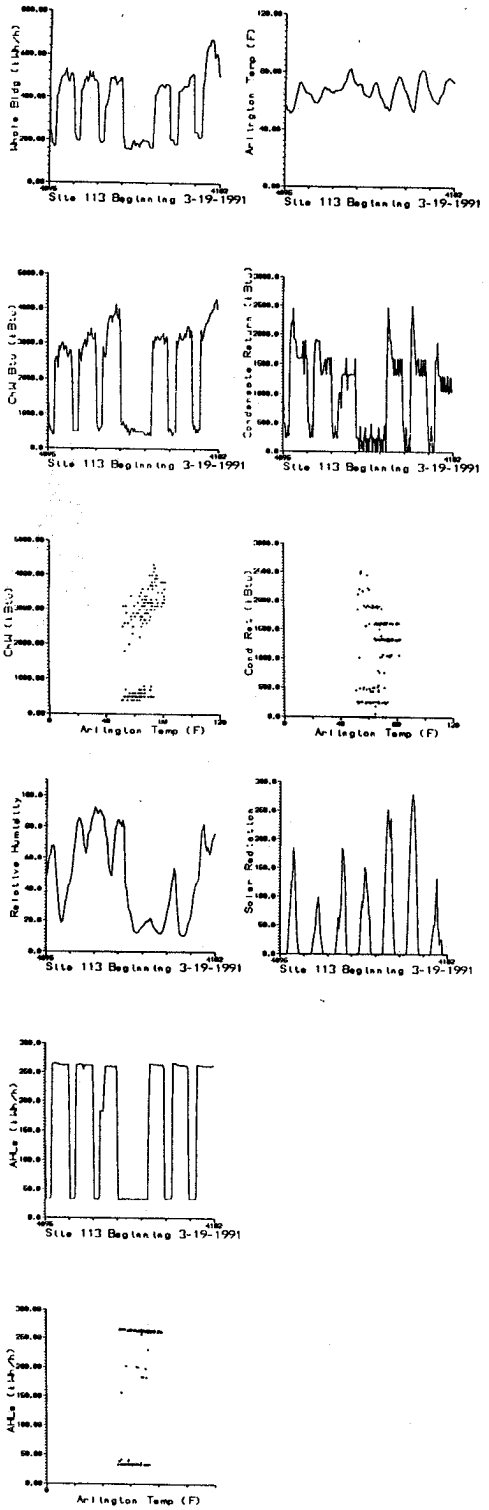

***FIGURE 3:*** *Typical Monthly Agency Report.* This figure is an example page of the monthly agency report for the Zachry Engineering Center.

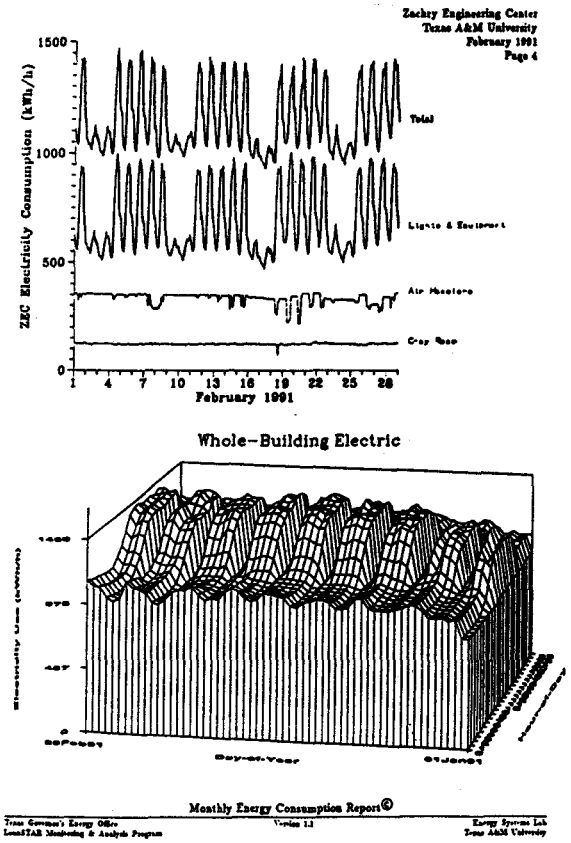


***FIGURE 4:*** *LoanSTAR Polling and Processing Flowchart.* This figure illustrates the primary processing steps involved in retrieving, cleaning and archiving the data from a site. Data are polled once per week and archived in both raw and processed format. Various modules have been developed for removing extraneous characters, checking for missing data, calculating derived weather data and producing various graphs.