

## Paper 132

**EPA's SAS Analytic Library on the Web**

by Jay Jacob Wind, American Environmental Institute, Inc.

Chapman Gleason, Center for Environmental Information and Statistics, US EPA

Michael Yu, Digital Intelligence Systems

Kamau Njuguna and Myles Powers, Northrop Grumman Corporation

**Abstract**

The U.S. Environmental Protection Agency (EPA) Center for Environmental Information and Statistics (CEIS) has developed and administers **EPA's SAS Analytic Library** using SAS/Share Servers, SAS/Data Warehouse Administrator, SAS/EIS, SAS/MDDB Server, and SAS/IntrNet.

This paper describes the strategy, architecture, and sample products of Library. We show how SAS analysts worldwide with SAS/Connect and TCP/IP can access the Library, obtain metadata about each data system and table in the Library, and produce environmental statistics, graphics, and maps. From CEIS's website and SAS/Share Server, SAS users at all skill levels on all operating systems can analyze environmental data for the whole nation or individual states, counties, ZIP codes, and communities. As far as we know, no other SAS/Share Server in the world provides as much environmental data in SAS format for analysis purposes.

**CEIS's Strategy**

**Figure 1** illustrates the audiences for data products from EPA's SAS Analytic Library. In essence, the Library publishes EPA's data via the Internet, to be used by SAS programmers in the environmental community worldwide.

CEIS chose to denormalize data to simplify analysis. **Normalized data structures** store each data point exactly once at its appropriate level, except for variables to link various data levels. **Denormalized** data structures merge all relevant variables onto one record. Normalized structures store data more efficiently, but programmers need detailed knowledge of the data structure. Denormalized structures take more space, but programming is simple.

**Normalized Data Example**

Facilities or Monitoring Stations Dataset
Sampling Circumstances Dataset
Environmental Measurements Dataset

**Denormalized Data Example**

Facility or Station Variables	Sampling Variables	Measurement Variables
-------------------------------	--------------------	-----------------------

**EPA's SAS Environment**

Since its creation in 1970, EPA has used SAS products extensively to manage, consolidate, analyze, and display data. SAS is EPA's primary tool for statistical analysis, presentation, and data management. Every EPA office and program uses SAS to analyze and display data.

Currently, EPA has a contractual agreement with SAS Institute, similar to the Enterprise Computing Agreement, for 21 desktop SAS products on seven desktop operating systems. The agreement also includes the networking and data server products. Besides the 21 desktop products, EPA has licensed at least one copy of SAS's server and Internet publishing products, SAS/IntrNet, SAS MDDB Server and SAS Data Warehouse Administrator (DWA).

**Figure 2** illustrates CEIS's Intranet website. From this website, CEIS administers SAS at EPA. Via the website, CEIS distributes SAS products using FTP, subject to license agreements, e.g. SAS Version 7 WinZipped; and consolidates and publishes technical knowledge for SAS users.

In addition, CEIS administers EPA's SAS Analytic Library using SAS/Share Server on EPA's IBM Mainframe. This server stores large environmental datasets in denormalized form at the proper unit of analysis for analysis and display. CEIS also administers aggregated data on an MDDB Server on one of EPA's DEC Alpha computers. CEIS's website supports drill down into this large SAS server as shown in **Figure 3**.

**CEIS Information Architecture and Products**

CEIS uses SAS software in all aspects of its web strategy and intends to deploy and release information products as tables, graphs, maps and comma-separated-variable (CSV) download files for enterprise level environmental data bases at the national, state, county, community (city, town or place) and ZIP code.

The CEIS computing architecture relies on base SAS, SAS/Share Server, SAS/IntrNet, and SAS/MDDDB Server to produce tables, graphs, maps, and download files via the Internet.

The view of CEIS's information architecture in **Figure 1** shows the relationship between audience, SAS product, level of summary, and CEIS product. This product architecture is not operational yet on EPA's public access website (<http://www.epa.gov>), but some components will be released later in 1999. We expect each CEIS database to produce environmental statistics nationally and by state, county, community, ZIP code, and record-level. Comparisons by geography will be provided in a future release.

Any SAS programmer worldwide can easily any of CEIS's record-level data on the IBM Mainframe SAS/Share server or aggregated data on the DEC Alpha MDDDB server. The SAS programmer must have SAS/Connect and TCP/IP connectivity installed on their client and two simple configuration changes to two files:

- /WINDOWS/HOSTS
- /WINDOWS/SERVICES (in DOS) or
- /etc/hosts
- /etc/services (in UNIX)

In HOSTS, add:

```
134.67.180.1 epaibm.rtpnc.epa.gov epaibm
```

In SERVICES, add:

```
ceissas 3005/tcp
```

Once these changes are in place, SAS use data in EPA's SAS Analytic Library. For instance, to print the TRIS data for 1996 for Dade County, Florida, a user could submit:

```
LIBNAME TRI96 SERVER=EPAIBM.CEISSAS ;
PROC PRINT DATA=TRI96.TRI9696 ;
WHERE COUNTYNM='DADE' AND SPC='FL' ;
RUN ;
```

**Figure 4** shows how the SAS Analytic Library presents metadata for each denormalized data system. CEIS calls it an analytic library and not a warehouse, since its primary purpose is to provide yearly statistics, rather than transactions.

CEIS stores the data at the proper unit of analysis for statistical purposes, unlike most relational databases that require complex JOIN commands before they can be analyzed.

It is said that only six programmers in the entire U.S. can write a moderate-to-complex JOIN for one particular EPA data system. Using

EPA's SAS Analytic Library, however, SAS programmers can use their skills directly to produce statistics, charts, and maps of EPA data.

### **Appendix: EPA's Data Collections**

EPA collects data under many laws:

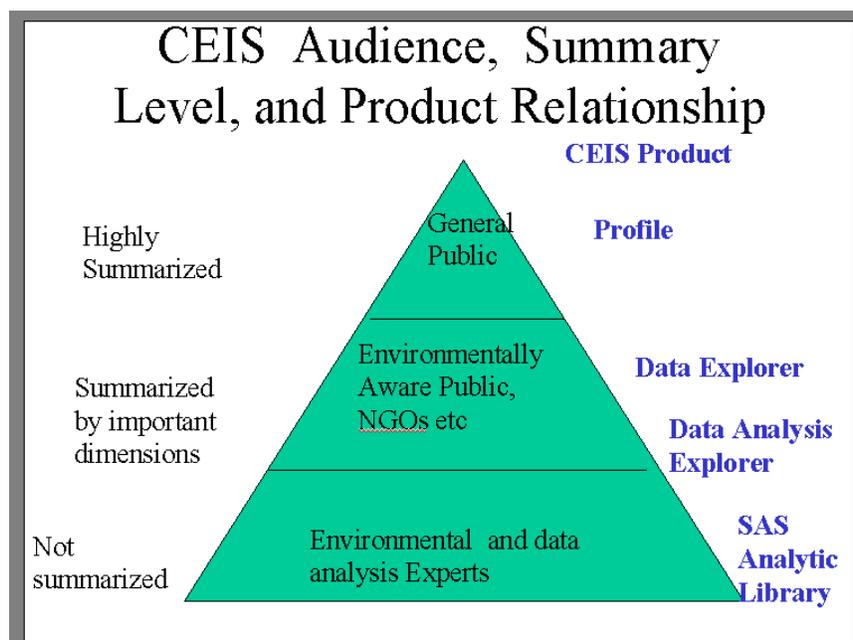
- Under the Clean Air Act (CAA), EPA collects data on air pollutant emissions and monitored air pollutant concentrations nationwide. EPA stores air emissions data in Aerometric Information Retrieval System (AIRS) Facility Subsystem (AFS) and air monitoring data in Air Quality Subsystem (AQS). Both AFS and AQS are written in ADABAS.
- Under the Clean Water Act (CWA), EPA collects data on water discharges and chemical concentrations. EPA stores water discharge data in Permit Compliance System (PCS) and monitoring data in STorage and RETrieval System (STORET). PCS is written in ADABAS; STORET is written in an EPA-proprietary database language.
- Under the Resource Conservation and Recovery Act (RCRA), EPA collects data on hazardous waste permits and wastes generated, managed, received, and shipped. EPA stores permit data in RCRA Information System (RCRIS) and waste amount data in Biennial Reporting System (BRS). Both RCRIS and BRS are written in FOCUS.

- Under the Title 3 of Superfund Amendments and Reauthorization Act (SARA) Emergency Preparedness and Community Right-to-Know Act (EPCRA), EPA collects data on air emissions, water discharges, underground injections, land disposals, off-site transfers, and pollution prevention for 600+ toxic chemicals into Toxics Release Inventory System (TRIS), written in ADABAS.
- Under the Safe Drinking Water Act (SDWA), EPA collects data on violations by drinking water providers in Safe Drinking Water Information System (SDWIS). SDWIS is written in DB2.

From these data systems, we built EPA's SAS Analytic Library. We followed standard data warehouse denormalization processes to extract, scrub, validate, integrate, and summarize each data system, although at first without the use of SAS/Data Warehouse Administer. Eventually, we intend to place all the code to extract, clean, and summarize the data into SAS/DWA.

SAS and all related products are trademarks of SAS Institute, Inc.

**Figure 1**  
The Audience for EPA's SAS Analytic Library



**Figure 2**  
The EPA Intranet SAS Support Website

EPA SAS Support Web - Netscape

File Edit View Go Window Help

Bookmarks Netsite: http://intranet.epa.gov/epasas/ What's Related

EPA SAS

[Home](#)

[Index to base, Macro and Procedures Documentation](#)

[SAS Online Documentation](#)

[SAS Online Documentation non-Java version](#)

[Business Geographics](#)

[Business Geographics Download](#)

[Interfacing SAS to EPA's Networks and Systems](#)

[Rapid Applications Development](#)

[Rapid Applications Development Download](#)

[SAS/IntrNet Software](#)

**SAS Online Documentation**

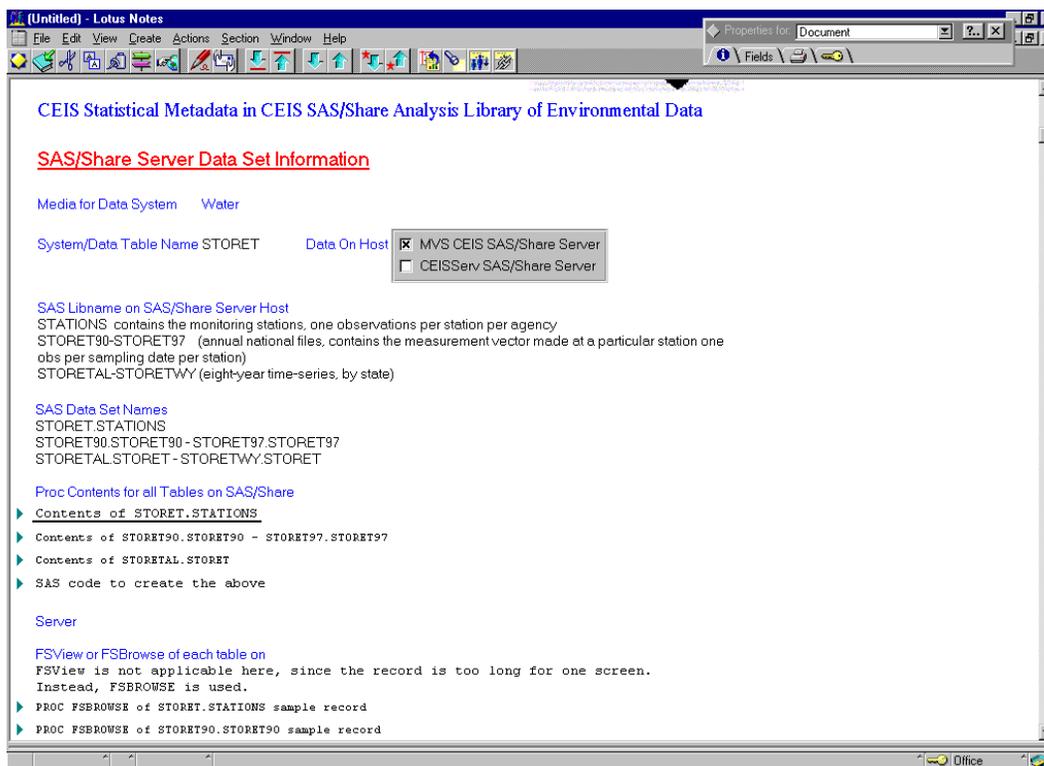
SAS OnlineDoc, Version 7, provides you with reference information for SAS software (see the Contents for a complete list of products that are included). It is intended for users who are already familiar with these products but who need to look up information such as

- the syntax of SAS language elements such as functions, statements, and procedures
- elements of the SAS language whose parameters or behavior differ somewhat in different operating environments, or elements that are available in some operating environments but not in others
- the fields in a particular window for windowing products such as SAS/SPECTRAVIEW
- how to use DBMS data in SAS programs
- how to present your data in the form of charts and graphs
- communications access methods that you can use with SAS software products in a networking environment.

SAS OnlineDoc  
VERSION 7.0

Document Done

**Figure 3**  
**CEIS SAS/Share Server of Environmental Data Displayed via a Web Browser**



**Figure 4**  
**Example of Metadata in EPA's SAS Analytic Library**

