Paper 171-28
Transactional Records Access Clearinghouse: SAS®-Based Warehouse and Mining Tools Keep Tabs on U.S. Government

Susan Long, Syracuse University, Syracuse, NY
Linda Roberge, Syracuse University, Syracuse, NY
Jeffrey Lamicela, Syracuse University, Syracuse, NY

ABSTRACT

Transactional Records Access Clearinghouse (TRAC), a research center at Syracuse University's School of Management, has developed a subscription website that allows many different types of users to access transactional data from the U.S. federal government. Subscribers include journalists, lawyers, judges, congressional staffers, public interest groups, and scholars. TRAC's end-users often lack experience with performing statistical or data analyses. Our challenge has been to develop a "point-and-click" interface that allows users to specify what information they would like, without the necessity of knowing how the data are organized or how the statistics are generated.

The power of TRAC's application lies in the combination of the data warehouse and data mining tools. Both of these are SAS-based using a variety of SAS Products including SAS/SHARE®, SAS/IntrNet®, SAS/GRAPH®, SAS/STAT®, and SAS/CONNECT®. The tasks of creating, maintaining, and updating the warehouse plus designing user-friendly, instructional tools have presented an ongoing stream of problems to solve, often with the aid of SAS routines. Accomplishing this on a research center budget (upgradeable to a shoestring) has taught us many valuable lessons. While the challenges faced by TRAC are in some respects unique, other SAS users can adapt our solutions to solve their own unique problems.

INTRODUCTION

Transactional Records Access Clearinghouse, better know by its acronym TRAC, was established in 1989. The purpose of TRAC is to provide the American people -- and institutions of oversight such as Congress, news organizations, public interest groups, businesses, scholars, and lawyers -- with comprehensive information about federal staffing, spending, and the enforcement activities of the federal government. On a day-to-day basis, what are the agencies and prosecutors actually doing? Who are their employees and what are they paid? What do agency actions indicate about the priorities and practices of government? How do the activities of an agency or prosecutor in one community compare with those in a neighboring one or the nation as a whole? How have these activities changed over time? How does the record of one administration compare with the next? When the head of an agency or a district administrator changed, were there observable differences in actual enforcement priorities? When a new law was enacted or amended, what impact did it have on agency activities?

To achieve these ambitious goals, TRAC has created a data warehouse and a suite of data mining tools, both SAS-based, designed specifically with the non-analyst end-user in mind. These tools allow users to generate tables, graphs, and maps that answer a wide variety of questions about how the government functions in different parts of the country and how this has changed over time. The ability to answer questions quickly and easily gives a picture of the workings of the US federal government that is available nowhere else.

The Data Warehouse

The data warehouse contains data from many different sources, including the Executive Office for United States Attorneys in the Justice Department, the Administrative Office of United States Courts, the Office of Personnel Management, the Internal Revenue Service, the Environmental Protection Agency, the Census Bureau, and a range of other specialized federal agencies. Areas covered include criminal enforcement, civil actions, administrative enforcement by the IRS, federal staffing, and federal expenditures, among others.

To build its data warehouse, TRAC begins by searching through government manuals, websites, and other sources in order to identify relevant systems of records and transactional databases maintained by different agencies. Based on these leads, TRAC makes requests for specific data sets and all of the agency documentation describing the details of what is covered and how the information is organized. The requests usually are made under the Freedom of Information Act (FOIA). When release of the data entails a lawsuit and court decision, this beginning step can require a great deal of time and expense.

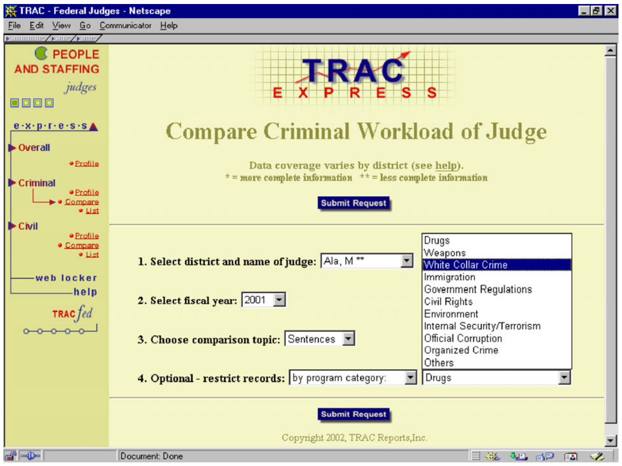
Once the data sets and documentation are in hand, statistical and other kinds of checks are made to test the completeness and reliability of the information that has been provided. Data from different sources about similar events may be compared or merged as a further check on data set reliability. Data sets and fields that fail to measure up are not used for analyses. When this happens, users are warned about the shortcomings. Numerous performance criteria are defined and indicators developed. The linking, grouping, and classification variables that will be used to place the data into geo-political-temporal context are then developed and added to the data. Because the contextual needs of users may be quite different depending on their interests, TRAC attempts to provide as much related information as possible, including geography, population, time trends, constant/real dollars, etc. Finally, the data are added to TRAC's data warehouse and made available via specially designed web sites. Recently, TRAC has begun the ambitious task of incorporating new data into its data warehouse on a monthly basis.

The size of TRAC’s data warehouse is considerable, taking up approximately 300 gigabytes of storage space, with more data being added all the time. For example, because there were more than 140,000 criminal referrals for prosecution in a recent year and the online data go back to 1986, the information in this area alone is extensive. (The Justice Department recently estimated that they have supplied more than 25 million records to TRAC!) In addition, TRAC's enforcement data cover civil matters -- where the government is either the plaintiff or the defendant -- and administrative actions by the IRS. Along with the enforcement data there is information on staffing going back to 1975, and federal spending going back to 1993. As a grant funded organization operating on a limited budget, TRAC constantly strives to automate the production processes for adding new data to the data warehouse.

The Data Mining Tools

As daunting as establishing a data warehouse may be, creating information from the warehouse is yet another formidable task. Given the size of the data warehouse alone, finding information amid all the data is akin to spinning a room full of straw into gold. This is where our specially designed data mining tools come into play. These tools allow our non-analyst end-users to explore the data warehouse looking for trends, relationships, and outcomes. The objective is to find the patterns that will provide a coherent unified view of the organization, and to place this information into a context that will make it understandable and usable for answering a user’s questions.

TRAC has developed three different types of data mining tools that enable users to analyze the data in the data warehouse. The first tool is called "Express" (see Figure 1).



Compare Criminal Workload of Judge

Data coverage varies by district (see help).
* = more complete information ** = less complete information

1. Select district and name of judge: Ala, M **

2. Select fiscal year: 2001

3. Choose comparison topic: Sentences

4. Optional - restrict records: by program category: Drugs

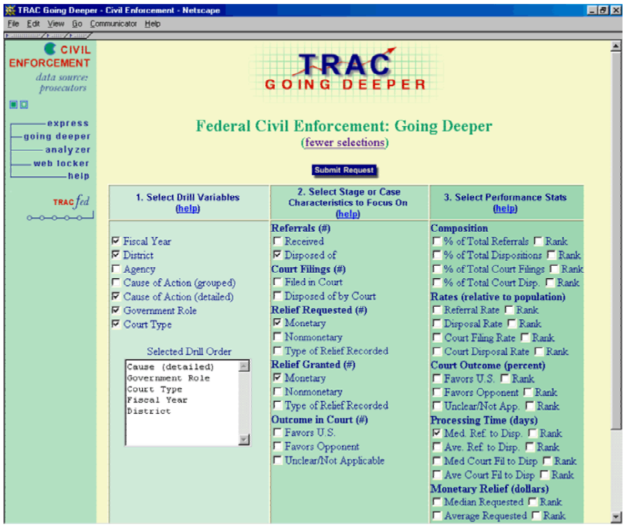
Drugs
Weapons
White Collar Crime
Immigration
Government Regulations
Civil Rights
Environment
Internal Security/Terrorism
Official Corruption
Organized Crime
Others

Copyright 2002, TRAC Reports,Inc.

Figure 1. Express Tool

As the name implies, the Express tool allows users to quickly and easily produce counts, averages, medians, and other specially computed measures that are used to generate rankings, comparisons, and trends. Users can specify if they want the information by district, agency, program area, or lead charge / cause of action. For IRS audits, users can also chose to have the information produced by income class, selection reason, and auditor type. Additionally, users are able to indicate whether they want the information returned to them in the form of tables, graphs, or maps. Because this tool is both powerful and easy to use, it is frequently the tool of choice for novices and power users alike.

Sometimes users need a multidimensional view of the data that allows comparisons across groups, years, organizational entities, etc. For example, I may want to see how my district handles health care fraud. How many of the referrals actually get prosecuted? How does this compare with other districts? Has this changed over the years? To provide this capability, TRAC has developed a second tool called “Going Deeper” (see Figure 2).



Federal Civil Enforcement: Going Deeper

1. Select Drill Variables (help)

☑ Fiscal Year
☑ District
☐ Agency
☐ Cause of Action (grouped)
☑ Cause of Action (detailed)
☑ Government Role
☑ Court Type

Selected Drill Order

```
Cause (detailed)
Government Role
Court Type
Fiscal Year
District
```

2. Select Stage or Case Characteristics to Focus On (help)

Referrals (#)
☐ Received
☑ Disposed of
Court Filings (#)
☐ Filed in Court
☐ Disposed of by Court
Relief Requested (#)
☑ Monetary
☐ Nonmonetary
☐ Type of Relief Recorded
Relief Granted (#)
☑ Monetary
☐ Nonmonetary
☐ Type of Relief Recorded
Outcome in Court (#)
☐ Favors U.S.
☐ Favors Opponent
☐ Unclear/Not Applicable

3. Select Performance Stats (help)

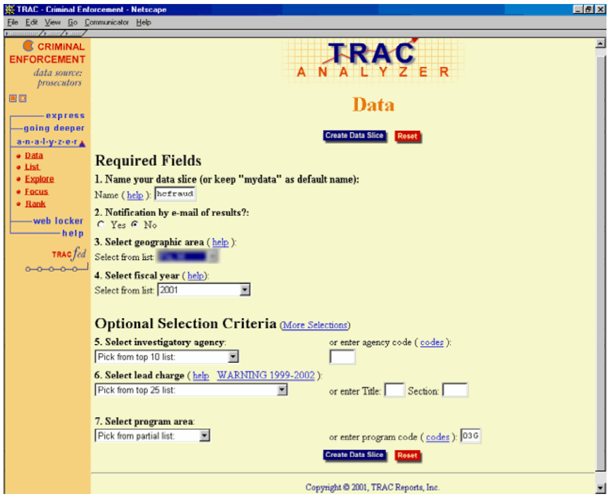
Composition
☐ % of Total Referrals ☐ Rank
☐ % of Total Dispositions ☐ Rank
☐ % of Total Court Filings ☐ Rank
☐ % of Total Court Disp. ☐ Rank
Rates (relative to population)
☐ Referral Rate ☐ Rank
☐ Disposal Rate ☐ Rank
☐ Court Filing Rate ☐ Rank
☐ Court Disposal Rate ☐ Rank
Court Outcome (percent)
☐ Favors U.S. ☐ Rank
☐ Favors Opponent ☐ Rank
☐ Unclear/Not App. ☐ Rank
Processing Time (days)
☑ Med. Ref. to Disp. ☐ Rank
☐ Ave. Ref. to Disp. ☐ Rank
☐ Med Court Fil to Disp ☐ Rank
☐ Ave Court Fil to Disp ☐ Rank
Monetary Relief (dollars)
☐ Median Requested ☐ Rank
☐ Average Requested ☐ Rank

Figure 2. Going Deeper Tool

The Going Deeper tool allows users to focus on a particular stage in the referral process and to generate performance measures such as percentages, rates relative to the population, and outcomes. Going Deeper, as the name suggests, provides a drill-down capability that enables users to produce and view the data as a series of linked tables that focus on increasingly narrower subsets of data down to a listing of the individual matters, federal employee, or judge. As with the Express tool, Going Deeper is easy to use via a point and click interface. And as with Express, users with a wide variety of experience find the Going Deeper tool to be user-friendly and capable of generating complex information.

The most advanced tool is the “Analyzer” (see Figure 3). This tool allows users to specify a particular slice of data that is of interest to them, and to store their own unique subsets of data in personal “web lockers” (see Figure 4). From a web locker, a user can run numerous types of sophisticated analyses, the results of which can also be stored in the web locker. As an adjunct to Express and

Going Deeper, Analyzer provides users with the ability to perform sophisticated analyses on any data in which they are interested.



Data

Required Fields

1. Name your data slice (or keep "mydata" as default name):
Name (help): hcfraud

2. Notification by e-mail of results?:
○ Yes ● No

3. Select geographic area (help):
Select from list: [redacted]

4. Select fiscal year (help):
Select from list: 2001

Optional Selection Criteria (More Selections)

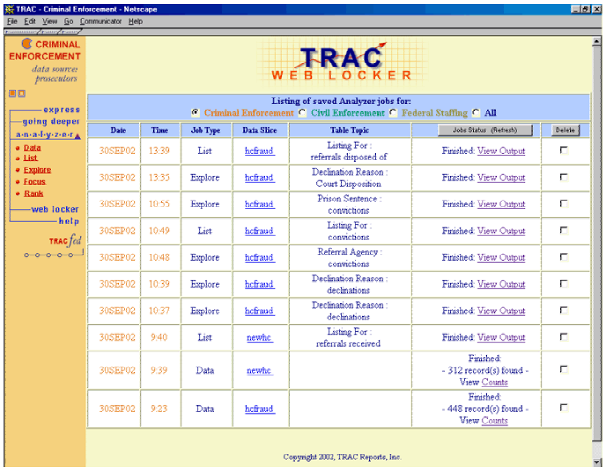
5. Select investigatory agency: Pick from top 10 list:
or enter agency code (codes):

6. Select lead charge (help WARNING 1999-2002): Pick from top 25 list:
or enter Title: Section:

7. Select program area: Pick from partial list:
or enter program code (codes): 03G

Copyright © 2001, TRAC Reports, Inc.

Figure 3. Analyzer Tool



Listing of saved Analyzer jobs for:
● Criminal Enforcement ○ Civil Enforcement ○ Federal Staffing ○ All

Date	Time	Job Type	Data Slice	Table Topic	Jobs Status (Refresh)	Delete
30SEP02	13:39	List	hcfraud	Listing For : referrals disposed of	Finished: View Output	□
30SEP02	13:35	Explore	hcfraud	Declination Reason : Court Disposition	Finished: View Output	□
30SEP02	10:55	Explore	hcfraud	Prison Sentence : convictions	Finished: View Output	□
30SEP02	10:49	List	hcfraud	Listing For : convictions	Finished: View Output	□
30SEP02	10:48	Explore	hcfraud	Referral Agency : convictions	Finished: View Output	□
30SEP02	10:39	Explore	hcfraud	Declination Reason : declinations	Finished: View Output	□
30SEP02	10:37	Explore	hcfraud	Declination Reason : declinations	Finished: View Output	□
30SEP02	9:40	List	newhc	Listing For : referrals received	Finished: View Output	□
30SEP02	9:39	Data	newhc		Finished: - 312 record(s) found - View Counts	□
30SEP02	9:23	Data	hcfraud		Finished: - 448 record(s) found - View Counts	□

Copyright 2002, TRAC Reports, Inc.

Figure 4. Web Locker

Through the use of these tools, users are able to literally "create" information by entering the data warehouse and analyzing the data contained in the individual transactional records concerning each matter. Although the tools are easy enough for beginners to use, they are powerful enough to provide short cuts for experienced analysts.

The Web Sites

TRAC harnesses the power of the Internet by maintaining two broad categories of web sites. First is a series of six free public web sites that mostly focus on the criminal enforcement activities of the Federal Bureau of Investigation, the Immigration and Naturalization Service, the Drug Enforcement Agency, the Bureau of Alcohol Tobacco and Firearms, the Customs Bureau, and the Internal Revenue Service (the sites can be reached from http://trac.syr.edu/).

TRAC's IRS site also includes information about IRS administrative actions -- audits, seizures, levies and liens, etc. Using data from the data warehouse, the free web sites offer pre-selected, but very extensive, views of each agency's enforcement activities, both nationally and within individual districts, along with graphs, maps, and tables that highlight interesting findings and trends over time (see Figure 5). The free sites also offer special studies on such subjects as counter-terrorism enforcement and long-term changes in federal staffing. The free sites do not allow users access to the data mining tools that would allow them to tailor the information to meet their own unique needs.

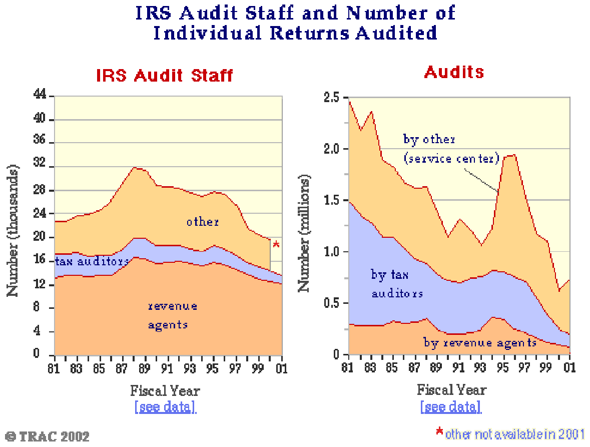


Figure 5. Graph from Public IRS Site

TRAC's second offering consists of a dynamic subscription site that provides vastly more information as well as access to the data mining tools (the subscription site is located at http://tracfed.syr.edu/). In the criminal area alone, for example, enforcement data can be organized by statute, district, Justice Department program category, and by virtually any agency. In addition to data about criminal enforcement, TRAC's subscription site offers a layer with extensive information about the civil matters processed by the U.S. Attorneys, and complete agency-by-agency staffing information -- from statistical overviews by federal judicial district, state, county or city, down to the names and salaries of individual employees. Federal expenditure data -- agency-by-agency and program-by-program -- provide yet another perspective on the government. Areas that allow users to explore the records of individual district court judges and federal prosecutors were recently added.

LESSONS LEARNED

Creating a data warehouse and developing "special audience" data mining tools is complicated under any circumstances. One needs to take complex data and to render them in a form that hides the complexities of structure and processing. The goal is to allow the data "to speak" to end-users without getting mired in the underlying morass of detail. Accomplishing this on a small budget is a real challenge. In our continuing struggles to deliver useful tools and information, however, we have learned some valuable lessons about structuring the data, designing the output, and creating an efficient back-end.

Structuring the Data

We have found that structuring our data is the key factor in maximizing its usefulness for all of the varied end users who are our subscribers. For us, it has required 90% of the effort, but, when done correctly, the rewards are great. Based on our experiences, we have compiled the following general principles that others may find useful.

- **Mirroring Work**
  The structure of the data must reflect the reality of how the organizational entity functions. This means that the data should describe what is important about the organization's work in terms of inputs and outputs, operational processes, etc. In our case, we need to think carefully about what the data may be used for, and what additional fields may be required to permit this use. For example, one portion of TRAC's warehouse consists of cases that have been referred to the U.S. Department of Justice for criminal prosecution. The referral process is complicated and lengthy, with many different stages from beginning to end. We have found it essential to set flags to indicate where in the work process a particular case is so that we can describe the overall functioning of the organization. Pre-processing the data to assure adequate mirroring of work processes requires substantive expertise on the organization, not merely programming expertise.

- **Rates, Percents, Per-Capita And Ranks**
  Producing counts and sums is an essential part of the capability of a data warehouse, capability that is enhanced with the use of appropriate flags. However, counts and sums alone are of limited usefulness. In order to fully understand what is going on within an organization, users need to be able to make meaningful comparisons among organizational sub-units, between a sub-unit and the whole, between the organization and the external environment, and across time periods. The use of rates, percents, per-capitas and ranks can facilitate comparisons as they allow end users to compare like units, i.e. apples to apples.

  Enhancing the capability to make comparisons requires two considerations. First, what comparisons will need to be made? Often a different picture will emerge depending on which type of measure is used. For example, looking at money spent on a regional basis will undoubtedly paint a different picture than looking at the same information presented on a per capita basis. We have found that there is no one measure that is appropriate in all situations, nor is there an a priori way to determine which measure is appropriate in a given situation. Thus we want to facilitate as many comparison measures as possible.

  The second consideration is whether or not the data contain the fields needed to enable measures. An obvious example would be presenting per capita information at a county level. In this case a location field is needed that can be linked to county-level population data. Perhaps not so obvious might be problems encountered with duration measures. For example, although the data may have what appears to be a date field, it needs to be determined if dates entered are defaulted to the first or fifteenth of the month.

  Particularly in the many areas related to measurement, we have found that SAS's capabilities make it shine; standard database calls to a transactional database simply aren't sufficient. In practice, while we have SAS do some of the required calculations online in real time, we commonly build summarized data sets that are then linked with the transactional data before they are loaded into the data warehouse. We have found this step to be essential for efficiency. (More detail on this later.)

- **Categories and Classifications**
  We have found it invaluable to have as many different schemas as possible with which to group results. Characteristics such as place, time, organizational unit, and the particular characteristics of each activity can each serve as the basis of one or more schemas. For example, looking at workload grouped by state is one way of categorizing by place, but it is also possible to look at workload by federal district. Often these categories do not overlap, so creating different schemas may not be simply a matter of summing to a larger category. However, because categorization is a key value-adding activity, we invest considerable resources in building them.

  Often transactional data will have already classified activities recorded, however the existing schemes are rarely sufficient as they are. When the number of categories is too large, the natural variability in the data (i.e. noise) will make comparisons meaningless. At TRAC, we spend a lot of time thinking about how to simplify schemas and reduce the number of categories. But how categories are collapsed can be tricky since poorly designed categories can conceal as much as they reveal. We have found that nested classifications are very useful so you can use broader classifications as well as drill into more detail when needed.

- **Consistency of Measurement**
  With any time series, the possibility always exists that the meanings of a measurement may change over time. Most people are aware, for example, that 1960 dollars are not the same as 2002 dollars. But the same type of phenomenon can occur with other measures also. If there have been changes to category definitions, or changes in department policies and procedures, time series that have been based on affected fields are broken. Consistency of measurement problems can also occur between organizational units using the same database when each unit uses different procedures.

  Whatever the cause, measurement inconsistency greatly reduces the value of the information. At TRAC, during the cleaning process, we look for indications of measurement inconsistency and attempt to find ways to make those measures comparable. Although this is not always possible, it nevertheless must be the goal. If measures cannot be made comparable, it may be feasible to at least have some overlap in series where change has taken place. Simply adding a footnote to a table noting that definitions changed adds nothing to the value of the data.

- **Audit Trail**
  Many of us are a bit lax when it comes to documentation. In this day of graphical interfaces, we have found that an additional benefit of using SAS is that our program code documents what we have done and the decisions we have made along the way. Additionally, a few well-chosen comments document the rationale for those decisions. This audit trail has become essential to us for several reasons. First, using the same programs means that refreshing data is easier in succeeding periods since we can reuse data checks and measure designs. Second, it insures that new data will be pre-processed in the same manner as previous batches, thus helping to protect the integrity of the warehouse. Finally, when questions arise about procedures, we have documentation that we can refer to.

Designing the Output

The information requested by end-users is generated and displayed automatically using a variety of vehicles including maps, time series charts, and tables. Simplicity of the display is important so that the data take center stage. Our goals are to organize the display so that it maximizes the ability of the user to make good judgments as well as to locate important patterns. When a display is prepared manually, it is possible to achieve these goals by choosing the type of display based on the characteristics of the data. In this way, only the essential data are displayed and the point is clearly conveyed. When the display is automatically generated, however, we need display types that work across a wide variety of data characteristics. In general, we do not support a display type unless it has substantial payoff for a wide variety of data characteristics.

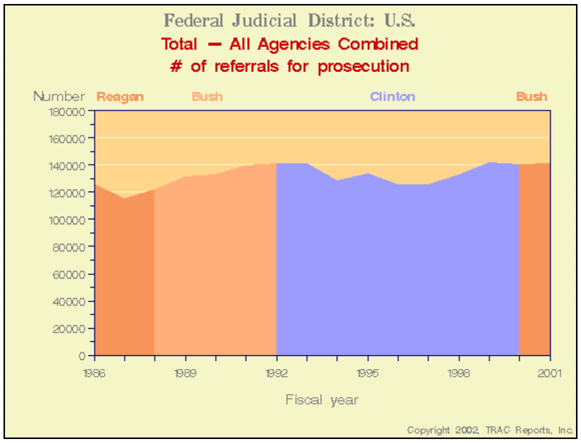
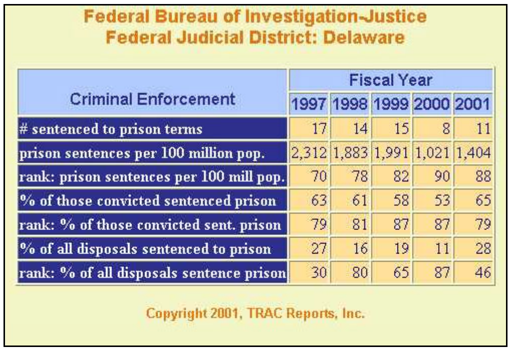


Figure 6. Dynamically Generated Chart

For maps and charts, we attempt to adhere to basic principles relating to visual perception when designing the display. These principles focus on how we perceive combinations of elements such as color, size, shape, and location (see Chambers, 1983; Chap 8). By using a combination of elements we have been able to assist users in visualizing several aspects in one plot; for example, including the time series display of the measure of interest plus presidential administration or political party in power (see Figure 6).

For tabular presentation of information, it is important to present rates, ranks, percents, and other calculated information in ways that give clues as to how they were calculated (see Figure 7). Including the frequency along with a statistic is a good way to do this. Doing so draws attention to those automatically produced results that aren't meaningful. You must also decide how to display missing values and how many digits to display. Design of tables including size and physical arrangement (e.g. which columns are placed next to each other, sort order, etc.) is extremely important. Badly designed tables hide meaningful relationships and can contribute to the familiar "drowning in data" sensation we have all experienced.



Federal Bureau of Investigation-Justice
Federal Judicial District: Delaware

Criminal Enforcement	Fiscal Year				
	1997	1998	1999	2000	2001
# sentenced to prison terms	17	14	15	8	11
prison sentences per 100 million pop.	2,312	1,883	1,991	1,021	1,404
rank: prison sentences per 100 mill pop.	70	78	82	90	88
% of those convicted sentenced prison	63	61	58	53	65
rank: % of those convicted sent. prison	79	81	87	87	79
% of all disposals sentenced to prison	27	16	19	11	28
rank: % of all disposals sentence prison	30	80	65	87	46

Copyright 2001, TRAC Reports, Inc.

Figure 7. Dynamically Generated Table

Tables that have two dimensions can be presented as cross-tabs and, if not too large, are understood easily by most users. However, tables with more than two dimensions can be difficult to comprehend. Suggestions for displaying the latter include nested tables and interactive "drill-down" displays.

Creating An Efficient Back-End

One of our persistent problems concerns the efficiency of our back-end processes. We have found three techniques to be particularly valuable in this regard. First, we use batch processes for our Analyzer tool and store the results in individual "web lockers." Second, we break larger data sets into smaller pieces, then index the smaller pieces. This has been very useful in implementing our "Going Deeper" tool. And third, rather than performing all calculations online in real time, we build summarized data sets for some measures. For the advantage of the extra efficiency gained from pre-processed data, we trade off an increase in the time required to load new data when new summaries must be run.

CONCLUSION

This paper has described the data warehouse and specialized data mining tools designed, created and maintained by Transactional Records Access Clearinghouse (TRAC) at Syracuse University. In the process of working with our data warehouse, we have learned many lessons that we have set forth as guiding design principles. Perhaps because many of our principles seem to be common sense, they are not often discussed. However, we have found that these can be the make-or-break factors in terms of warehouse usability. As a result we put considerable thought and effort into making these design decisions.

SAS software forms the basis for both the warehouse and mining tools. Our data are stored as SAS data sets, and our mining tools are implemented via SAS procedures. Particularly when compared with database software, the power and flexibility of SAS gives us the ability to do more than simply retrieve and display.

REFERENCES

Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey, P. (1983) *Graphical Methods for Data Analysis*. Belmont, California: Wadsworth International Group.

CONTACT INFORMATION

We encourage readers to visit our websites and to contact us with any comments or questions. For more information and a free trial subscription, SUGI participants can go to http://tracfed.syr.edu/sugi.html. All authors can be reached at:

Transactional Records Access Clearinghouse
488 Newhouse II
Syracuse University
Syracuse, NY 13244
Voice: (315) 443-3563
Fax: (315) 443-3196

E-mail addresses:
Susan Long -- suelong@syr.edu
Linda Roberge – lroberge@syr.edu
Jeffrey Lamicela – jlamicel@syr.edu