FROM DATA WAREHOUSE TO INFORMATION CRAFT SHOP: THE CHANGING SHAPE OF INFORMATION SUPPORT FOR ENVIRONMENTAL PROTECTION

Malcolm K. Sparrow
Kennedy School of Government, Harvard University, Cambridge, MA 02138

INTRODUCTION

When William Ruckelshaus became the first Administrator of the U.S. Environmental Protection Agency in 1970, he selected the phrase "Pollution Abatement" as his mission statement (Wiehl 1974). In 1990 the language of senior EPA officials includes, instead, "pollution prevention", "risk management", "cross-media integration," and "managing for environmental results" (Grizzle 1989; Reilly 1989; Reilly 1990a,b,c). To many these phrases signal important new ideas about the business of environmental protection and about its ends and about its means; they signal a shift in the strategy of environmental protection (USEPA 1989b).

New strategies require new institutional structures and processes to support them. In particular, they require new uses of information and of information technology.

The task for this paper is to examine the new ways in which information technology is being used within Federal and State environmental protection agencies and to determine the extent to which they fit new or emerging strategies. The focus of the paper is the nationwide effort to coordinate and integrate data collection—The State/EPA Data Management Program.

PIONEERS IN GEORGIA

In 1985 Georgia’s Environmental Protection Division (EPD) of the Georgia Department of Natural Resources established a "Comprehensive Data Management Program", introducing on-line State access to EPA databases. The federal databases were physically resident on EPA’s mainframe computers at Research Triangle Park (RTP) in North Carolina. There were several motivations for the changes.
In terms of mandated provision of data, the U.S. EPA had viewed the Georgia agency as one of the worst in the country (GDNR 1984). Georgia's EPD had been seriously embarrassed at data discrepancies brought to light during Congressional hearings. State and federal officials, called to testify, both talked about the same facilities in Georgia, but they had two completely different sets of data, with two different sets of opinions and conclusions (Sparrow 1990a).

Also Georgia's EPD, faced with increasing demands and diminishing resources, was learning the importance of accurate data as the proper basis for intelligent priority setting.

Third, the agency was having some highly information-intensive tasks added to its work load. A new state law required EPD to regulate agricultural users of ground water (GDNR 1988d), and another required regulation of underground storage tanks (GDNR 1988c). Georgia had over 10,000 farmers using ground water (whereas they were previously monitoring only a hundred or so industrial users) and roughly 50,000 tanks. Hence, the agency faced a massive explosion in the regulated community, quite apart from the continuing rapid expansion of the population of Georgia with its expanding industrial base.

Fourth, the Georgia agency faced serious difficulties with the computer support that EPD received from the State's central data processing facility (GDNR 1984; Sparrow 1990a)

EPA offered assistance to Georgia's EPD in the form of a Prime 2550 minicomputer, which could act as a front-end interface to the national databases. The Prime would be directly connected through a high-speed data link, and would be able to handle all of the state's local requirements not met by the available federal databases.

The promise of the scheme, for EPA, was the prospect of obtaining reasonable quality data from Georgia—hitherto one of the most seriously delinquent states in that regard—since EPD would rely upon the data in EPA's databases for its own operational purposes. The advantages for EPD included cost savings and resolution of their immediate problems.

**Phase I: Data sharing**

There were two distinct aspects to the program. The first established on-line connections from EPD to seven EPA databases at Research Triangle Park, North Carolina:

<table>
<thead>
<tr>
<th>System</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) NEDS</td>
<td>National Emissions Data System: provides data on air emissions from industrial facilities;</td>
</tr>
<tr>
<td>(2) SAROAD System</td>
<td>System for the Storage and Retrieval of Aerometric Data: provides ambient air quality data from air quality monitoring sites;</td>
</tr>
<tr>
<td>(3) CDS Compliance Data System</td>
<td>Compliance Data System: provides air compliance data. (Superseded, along with NEDS and SAROAD, in 1989 by AIRS, the Aerometric Information and Referral System);</td>
</tr>
<tr>
<td>(4) GICS Compliance Data System</td>
<td>Grants Information and Control System: keeps track of municipal sewage treatment plant construction projects when supported by federal grants.</td>
</tr>
<tr>
<td>(5) STORET Water Quality Storage and Retrieval System</td>
<td>Water Quality Storage and Retrieval System: provides water quality data from monitoring sites;</td>
</tr>
<tr>
<td>(6) PCS Permit Compliance System</td>
<td>Permit Compliance System: provides compliance data and effluent data for facilities having surface water discharge permits; and</td>
</tr>
<tr>
<td>(7) HWDMS Hazardous Waste Data Management System</td>
<td>Hazardous Waste Data Management System: provides information about hazardous waste sites.</td>
</tr>
</tbody>
</table>

A Memorandum of Understanding stipulated that EPA would not alter the State's data. Previously EPA would quite happily have altered data they mistrusted or did not like. A formal management procedure was established for resolving any disputes about accuracy, and EPD became solely responsible for the quality of its data.

The Phase I objective was met, in Georgia, by the end of 1986, although some of the benefits of the data-sharing scheme were apparent long before that. So much so, in fact, that the EPA headquarters in Washington, D.C. was beginning to pay close attention to what was happening in Georgia. Lee Thomas, EPA's administrator, took a close personal interest, and staff at EPA Headquarters' Office of Information Resources and Management began to see the potential impact of the data-sharing approach if extended to other states (ASI/WPCA 1986; Sparrow 1990b).

The phenomenon was simply that Georgia's EPD had turned around from being one of the worst providers of data to one of the best, in less than two years.

**Phase II: Decision making for environmental results**

Phase II of Georgia's program, partially supported by a grant from EPA, related more to the uses to which the data would be put, than to the mechanisms through which it would be collected. The explicit goal of Phase II was "the use of automated systems to better support decision making for environmental
results" (GDNR 1988a). Under that rubric, three specific objectives were highlighted:

(1) Data integration to support decision making;

(2) Use of Geographic Information Systems (GIS) to relate and analyze data from the different media (air/water/land etc.); and

(3) Support of the toxic substances and risk management programs. This was a state initiative to help coordinate emergency response to toxic spills, and to make rapidly available the necessary chemical information for dealing with such an emergency.

It was the use of Geographic Information Systems that most dramatically demonstrated the value of cross-media integration of data (GDNR 1988a). The EPD in Georgia had five early successes. They are worth examining in some detail as they shed some light on the philosophy and strategy of this state agency.

Example 1: Siting sanitary landfills. Traditionally municipal applicants for permits to construct sanitary landfills had to face rejection of their proposals by the Georgia EPD in at least 50% of cases. Applicants tended to choose rocky, swampy, or steeply sloping sites because these were unsuitable for other development and consequently low priced. They also had high environmental impacts, being groundwater recharge areas, or wetlands, or upstream of drinking water sources.

To make relationships with applicants more productive, the EPD used GIS to construct a map, for public release, which showed vast areas of Georgia that were unsuitable for siting sanitary landfills. The map itself resulted from the overlay of nine separate digitized databases. The map increased the productivity of both the EPD in processing applications and the local governments in their selection of possible sites. It also produced a more cooperative relationship between them.

Example 2: Prioritizing hazardous waste sites. There are 3230 facilities in Georgia which generate, treat, store, transport, or dispose of hazardous waste. Georgia's EPD could not pay attention to them all, so they plotted their locations in relation to significant groundwater recharge areas. A crude scoring system, incorporating proximity to recharge areas, together with the toxicity and the subjectively assessed risk of a release, enabled EPD to rank the sites in order of environmental impact. The product of the GIS application in this case was a list of priorities, not a map.

Example 3: Impact of agricultural chemicals. Georgia recognized the risk posed by overuse of agricultural chemicals. By superimposing the ground water recharge areas on the agricultural land-use areas, EPD highlighted particular agricultural areas as targets for educating farmers on management practices that would minimize the risk.

Example 4: Siting regional water supply reservoirs. Georgia's drought problem required intelligent siting of new reservoirs for drinking water. EPD used GIS to help identify appropriate sites by integrating information about measured stream flows (during recent drought periods), local population density, local land uses, acceptability of dam sites, and stream classification (i.e., they had to avoid building dams on trout streams etc.)

Example 5: Communicating the risks of radon exposure. The Georgia General Assembly requested an assessment of the risks posed by radon in the state. EPD used GIS to superimpose zip-code areas on a geologic map of rock formations with high natural concentrations of radioactive minerals.

Other methods of data integration

The EPD's local Prime 2550 computer became host to new integrated databases, designed and implemented by EPD in high level database programming languages. These local databases served specific local management needs, but drew their information from EPA's databases, combining data elements from two or more of those databases and presenting it in an integrated format.

The first was the Facility Identification Module, which stored basic information about the states' regulated facilities—their name, location (coordinates), address, and contact names, together with a unique EPA identification number—irrespective of which program was obliged to regulate them.

The second was the Permit Application Tracking Module, which was designed to provide a management overview on EPD's accomplishments, backlogs, and productivity with respect to processing permit applications.

A third was the Compliance Management and Enforcement module, designed to facilitate inter-program coordination on routine facility inspections and on enforcement actions. It enabled, for example, the water branch to examine the status of the air permits and vice versa. It also provided the basis for scheduling the multi-media inspections that EPD ran from
its four Regional Offices. The data it drew from program-specific EPA databases included the results of inspections, details of notices of violation, consent orders, and dates upon which facilities returned to compliance. It provided a quick and comprehensive view of any one facility's compliance record and status, thus more clearly revealing which were the good guys and which were the bad guys.

**EPA HEADQUARTER'S PERSPECTIVE**

EPA also made data management a priority, and established its Office of Information Resources and Management in 1985. That Office, with strong support from its administrator Lee Thomas, formally established the State/EPA Data Management Program as a nationwide umbrella for the further development of information management. It was modelled on the experience of Georgia's EPD and was similarly divided into two phases (USEPA 1989a).

Phase I (Data Sharing) concentrated on the establishment of the physical network infrastructure and cooperative relationships and understandings necessary to ensure provision of "complete, accurate and timely data". Phase II (Data Integration) was more concerned with making intelligent use of the data once they were available. The explicit goals of the two phases were as follows:

**Phase I**

- Obtain and enter complete, timely, and accurate information in all EPA enforcement, monitoring, and compliance data bases.
- Provide states with high-speed, on-line access to all EPA data bases.
- Ensure the integrity of state data in EPA data bases and due process for states in resolving data disputes with EPA.

**Phase II**

- Provide EPA and the states the data, methods, and technology required to conduct integrated environmental analyses and to plan and manage cross-media programs.
- Build effective, long-lasting arrangements for sharing data and technology between environmental agencies at all levels and with our colleagues in federal, state, and local agencies that are responsible for commerce, agriculture, science, and natural resources conservation.

In 1987, one state in each EPA region (and two from region IV) were selected as Regional pilot states for Phase I. In 1988, a further 19 states were added to the list. As of April 1991, high speed data links to all states except Hawaii (also to Puerto Rico) have been connected.

However, installation of the communications channels is no guarantee that the various programs within state agencies will adopt on-line access to EPA's databases. Nor is technical adoption of Phase I's data-sharing provisions any guarantee that states will grasp the fundamental purposes and significance of Phase II.

Many states have experienced substantial difficulties in linking the federal databases to their own existing computer systems (Sparrow 1990c). Even when the technical connections are completed, some states report that their staffs find EPA's databases unfamiliar and more difficult to use than their own (Sparrow 1990c; 1990e). Most states find that EPA databases do not meet all of their operational requirements (USEPA 1989b; Sparrow 1990a; 1990d; 1990e). And often the long term benefits of the Data Management Program are not sufficiently visible or not sufficiently understood to make the implementation obstacles worth overcoming.

**STRATEGY AND CULTURE AT GEORGIA'S EPD**

Therefore it is prudent to ask what was so special about the culture and philosophy of Georgia's EPD. Why did these new information management ideas fit there, when they do not fit so well in some other states and regions? (Sparrow 1990c; 1990e). Six features of Georgia EPD's philosophy are quite striking (Sparrow 1990a).

**Cross program integration**

For Georgia's EPD, cross-program integration was the routine way to use data. Data management was only one of ten functions within the Program Coordination Branch whose budget was roughly $5 million annually, representing between one third and one half of the EPD's program budget (GDNR 1988b).

One major program within that Branch was the administration of the multi-disciplinary inspection teams, based around the state in four regional offices. Fifty interdisciplinary field officers handled all of the inspections for 6000 of the state's 10 000 regulated facilities. The others, requiring specialist engineering or complex technical skills, were regulated by program specialists from the State central EPD office. The EPD moved to this multimedia approach over a decade ago, in 1976.
Cooperative attitude towards industry

EPD is almost service oriented, as opposed to viewing industry as the involuntary object of enforcement and regulation. And it views enforcement as only one means for compliance.

Another sign of the agency's commitment to working with, rather than against industry, is its New Industry Program, designed to smooth the whole process of site selection and permit issuance for industry contemplating moving into Georgia (Sparrow 1990a). This EPD scheme helps to attract industry to Georgia by eliminating unnecessary obstacles.

Selection of performance measures

EPD's performance measures are described by its staff as results oriented, not numbers oriented. They have rejected the traditional measures of output or productivity: the numbers of enforcement actions taken, the numbers of permits issued. Instead, they set goals such as to insure that dissolved oxygen in a given reach of a river achieves a particular level within six months. In other words, they measure their performance by environmental results.

Attitude towards public release of data

They view release of data and information to the public as lightening the agency's load, not increasing it. They see it as a method of empowering the public to participate in procuring a cleaner, safer environment.

Pioneering spirit

Georgia EPD also prides itself on being quick to adopt new ideas, new practices, and to experiment with new programs. Staff members are encouraged to publish papers, attend conferences, and to help out agencies in other states who want to learn from Georgia's experience.

Attitude towards other state agencies within Georgia

EPD consciously assumes a cooperative rather than an adversarial position. In particular, the interests of economic prosperity and environmental protection—so often in conflict—seem well integrated.

NEW STRATEGIES FOR ENVIRONMENTAL PROTECTION

The Georgia story provides some useful clues about the directions in which strategies of environmental protection might be moving. The signals about reorientation in philosophy emanating from federal EPA have been quite unmistakable, especially since the arrival of William Reilly as Administrator in February 1989 (Reilly 1989; Reilly 1990a,b,c). The initiatives being pursued by EPA's Office of Information Resources and Management properly reflect some of those strategic shifts.

Phase II of the Data Management Program is, in fact, predicated on the assumption that environmental agencies will have to change the way they do business in at least three important ways: first, there is the perceived need to break out of the strict programmatic straightjacket (water/air/hazardous waste etc.) in order to do cross-media risk analysis and problem solving; second, there is the push to focus on environmental impact rather than on the more traditional "bean-counting" output performance measures; and, third, there is a new emphasis on drawing the public into the decision-making process through provision of public access to data.

Cross program integration

The need for cross program integration is obvious. It is mandated by the advent of new environmental problems such as global warming.

When EPA was formed in 1970 from a number of disparate programs, the congressional oversight arrangements were never integrated. EPA Headquarters staff is primarily disaggregated along program lines. These divisions within EPA Headquarters are mirrored by similar divisions within the EPA Regional Offices and within almost all state agencies. The strength of the programmatic divisions present a constant challenge for the proponents of the Data Management Program, in the details of Phase I as well as in the more philosophical concerns of Phase II.

Managing for environmental results

Equally obscure is the extent to which managing for environmental results can become a practical reality for EPA within the foreseeable future. The Program Staff at EPA recognizes William Reilly's commitment to it, but still feels the compulsion to maintain or increase the numbers of judicial referrals as a method of demonstrating their productivity. Some of the staff are still more conscious of the requirement to collect data, than the need to use it intelligently. For many the most obvious purpose for having national data collections is for congressional reporting, not for operational decision support. Their focus on getting the data from states, makes them natural fans of the Data Management Program, at least of Phase I. Anything that helps make it easier to get the states' data is a bonus. The need for Phase II seems much less immediate.
Looking further ahead, to measuring states' performance by the quality of their water rather than by the number of enforcement actions and the extent of permittee non-compliance, EPA officials point to the extreme difficulty of designing generally applicable water quality measures.

There seem to be two major obstacles facing any major shift towards managing for environmental results: first, the scientific problem of establishing meaningful measures of environmental quality; and, second, the necessity of re-educating not only agency staff, but the legislatures, the media, and the public as to appropriate new forms of accountability. The scientific debate about the first and the political debate about the second are active and vigorous.

Public access to data

While EPA continues the struggle to design more meaningful indicators of environmental quality and better measures of environmental impact, it simultaneously has to deal with ever increasing public interest in environmental affairs.

The Data Management Program has not yet included public release of data within its formal Phase I or Phase II goals. But senior EPA staff know that it has to come (Habicht 1989). They describe EPA's past attitude to data as being secretive and possessive. And they see the public availability of toxic discharge data as a small indication of what will come.

Since 1989, a National Library of Medicine computer provides on-line access to one of EPA's databases. Using a PC and modem, anyone can access the files on 17,000 different manufacturing facilities across the country. Included are emissions to each media on an annual basis. The Toxic Release Inventory, as it is called, contains data on over 300 different toxic chemicals, and has a total of 75,000 records, submitted by industry itself (USEPA 1986).

Officials of EPA's Office of Toxic Substances, responsible for directing this implementation, refused to protect the chemical industry from public misinterpretations of the data. The burden of educating the public had to be shared by the industry itself. The result, according to a senior EPA official, was that the industry began a meaningful dialogue with the public about the data, rather than stonewalling. "The result has been more two-way communication about toxics in the last year than we've ever seen before" (Sparrow 1990b).

FROM DATA WAREHOUSE TO INFORMATION CRAFT SHOP

These trends in environmental protection strategies require new forms of information support. Data collection cannot remain focused on levels of activity; regional and national aggregation cannot remain the dominant form of analysis. Four broad trends for information support will emerge.

First, the process of analysis of incidents comprising any one risk will require the facility to aggregate and disaggregate data along a number of different dimensions—geographic, temporal, type of pollution source, affected medium, causal mechanism for the pollution, offending chemical, or class of chemicals. That requires flexible database structures and versatile analytic capabilities.

Second, the information and analytic support for risk identification, risk selection, problem-solving, and resource allocation will be required at many different levels—ranging from support for isolated, street-level problems dealt with by a single Field Engineer, to major and protracted national priorities.

Third, agencies' information and analytic support has to be available for environmental problems whether they are covered by traditional enforcement programs or not. Public partnership may result in the nomination of new priorities.

Fourth, increasing attention should be paid to educating and thereby empowering the public regarding environmental matters. Release of appropriate information should be seen as a powerful tool, not a regrettable drain on resources.

The emerging strategies will therefore require the production of more appropriate information products, in addition to enhanced data collection. Good quality data only has to be accurate, up to date, and complete. Good quality information has to be relevant, useful, comprehensible, well designed and presented, and in the right hands. Environmental protection can no longer be adequately supported by a "data warehouse". It has to be supported by an "information craft shop".

It could be argued that Phase I of the State/EPA Data Management Program seeks to perfect the traditional national data warehousing system and that Phase II, with its emphasis on GIS units, introduces the information craft shop, elevating the status of data analysis.

CONCLUSION

It is not yet clear to what extent the Data Management Program itself will ultimately be successful in supporting strategic changes within the environmen-
Information support for environmental protection

tal protection community; nor is it clear to what extent the progress of the Data Management Program will be constrained and hampered by the traditional enforcement-oriented and programmatically divided status-quo.

But it seems an inescapable reality that environmental information is one of EPA's major assets; and that the agency's future credibility will depend to a considerable degree on how it manages, uses, and communicates that information. Collecting the data is a necessary prerequisite. But, the aggregate picture has to be presented in such a way and to such an audience that it serves some useful environmental purpose.

The implications for the State/EPA Data Management Program are clear. First, Phase I must never be regarded as an end in itself. The EPA national databases, however complete, accurate, and up-to-date, are of no use unless significant attention, resources, and creativity are applied to generating useful and usable information products from them.

Second, Phase II should never be made to wait for Phase I. The transformation of the strategy of environmental protection does not depend on perfecting the national information infrastructure. Many information products supporting intelligent environmental decision-making are now being generated without using data from the major national databases. Supporting a wider variety of risk-assessment and resource-allocation decisions utilizes a wider variety of data sources and demands many non-traditional forms of analysis.

REFERENCES

ASIWPCA (Association of State and Interstate Water Pollution Control Administrators). Comprehensive data management conference in cooperation with the USEPA; conference report. Georgia Department of Natural Resources, Atlanta, GA; 1986.
GDNR (Georgia Department of Natural Resources) Georgia data management recommendations; consultant's report. Environmental Protection Division of the Georgia Department of Natural Resources, Atlanta, GA; 1984.
GDNR (Georgia Department of Natural Resources). Use of integrated information system for environmental decision making. Report on Phase II of the Georgia/U.S. EPA Comprehensive Data Management Project. Environmental Protection Division of the Georgia Department of Natural Resources, Atlanta, GA; 1988.
GDNR (Georgia Department of Natural Resources). Annual Report. Georgia Department of Natural Resources, Atlanta, GA; 1988.