Corporate Environmental Behavior and the Effectiveness of Government Interventions

PROCEEDINGS OF

LUNCH PANEL DISCUSSION

A WORKSHOP SPONSORED BY THE U.S. ENVIRONMENTAL PROTECTION AGENCY'S NATIONAL CENTER FOR ENVIRONMENTAL ECONOMICS (NCEE), NATIONAL CENTER FOR ENVIRONMENTAL RESEARCH (NCER)

April 26-27, 2004
Wyndham Washington Hotel
Washington, DC

Prepared by Alpha-Gamma Technologies, Inc.
4700 Falls of Neuse Road, Suite 350, Raleigh, NC 27609

ACKNOWLEDGEMENTS

This report has been prepared by Alpha-Gamma Technologies, Inc. with funding from the National Center for Environmental Economics (NCEE). Alpha-Gamma wishes to thank NCEE's Cynthia Morgan and Ann Wolverton and the Project Officer, Ronald Wiley, for their guidance and assistance throughout this project.

DISCLAIMER

These proceedings are being distributed in the interest of increasing public understanding and knowledge of the issues discussed at the workshop and have been prepared independently of the workshop. Although the proceedings have been funded in part by the United States Environmental Protection Agency under Contract No. 68-W-01-055 to Alpha-Gamma Technologies, Inc., the contents of this document may not necessarily reflect the views of the Agency and no official endorsement should be inferred.

TABLE OF CONTENTS

Transcript of Tuesday's Lunch Panel/Discussion

Matt Clark (U.S. EPA, Office of Research and Development)
I'll give you a very brief background. As you can see, we have a fairly significant number of researchers who are using facility-specific data from EPA, whether it's TRI or what we call PCS, which is our water pollution system, or some of the enforcement stuff, through ECHO and IDEA, or the air system, or the biennial reporting system for people who are working with hazardous wastes—there's a whole raft of these things that we want to try to align because every time I have a researcher who is trying to pull these things together, it's taking $100,000 off a grant. So, I don't want to keep spending that $100,000 over and over again, and what we're winding up with is—I mean, it would be nice if we could just bring all this stuff in, but you can't, because the researchers are looking at a specific industry, or like Dave Ervin, at a state, and things like that. We're hoping that there's some economy of scale here, some efficiency to be gained, by trying to address sort of holistically what the difficulties are associated with databases and to resolve some of those difficulties.

We had sort of an open meeting on September 3rd, I believe, of last year, and I'm just going to recap it. We had some of the same people who are up here talking, both the EPA database managers, who are describing their plans—a *great deal* of improvement in EPA databases has been made over the last few years—and some of the researchers expressing some of the difficulties that they've had in trying to pull stuff together. In a lot of ways, it works both ways, because the researchers need to know a little bit more than just what the data are—they have to know what the data *mean* and how they were collected.

So, just to recap it, there's a lot of interest in using EPA facilities-specific databases for research; the researchers are cumulatively expending a lot of time and resources matching and cleaning data. There are difficulties in combining information from different databases, most particularly for panel data sets, which is what people are most interested in using. And, there's difficulty on the researchers' end. To people who work in the field at EPA it becomes readily apparent that the researchers don't always understand the data's purpose—how and why it was collected, its limitations, and what it actually means. There has been *substantial* progress in creating single identifiers for facilities and linkages in the Facility Reporting System (FRS) and Envirofacts, which is another sort of a window to all our data world. And, there were problems in linking with Census that Randy Becker talked about, and Randy is *from* Census.

What we're trying to do today is sort of *move on*. We've been talking, and we seem to have had, perhaps not agreement, but a lot of nods in the audience in the September 3rd meeting that it might be a good idea to develop a combined database, a facility research database, and our thought was that we would use the LRD (longitudinal research database) model and develop some sort of sample of the roughly 1.8 million facilities or "things that are regulated out there" in the EPA data sets. When you start talking about panel data, that is just too much to deal with, so we have to figure out a way to limit it and make the kind of matches and corrections that we need. The idea is to match the

data, on a facility basis, from a variety of databases, including enforcement and all the others I mentioned, and some I didn't mention. It would be *nice* if we could link it to Census, particularly the longitudinal research database, in some fashion. There are problems with the IRS, and privacy rights, and things like that, so we have to try to figure out how to address those, but also, somehow, to have ways to identify what kind of company it is and link to Dun and Bradstreet.

Clearly, we want to build a panel data set, which is going to be very hard. We're going to need things like dedicated servers—it's a lot of information and a lot of different facilities. I don't know how *big* it has to be, . . . and one of the things that we discussed in the meeting is that this is something that would have to be *regularly* updated—it has to be a living thing that we have to keep up-to-date over time—on a quarterly, monthly, whatever the right basis is—to keep it going. Ideally, it would be easily aggregated to state, industry sector, etc.

Now, we really have some questions here—it *seems* to be the way to go, but we don't know. We really want our panelists—and we'll give you a few minutes each—to talk about this. Is this a desirable thing to do—certainly from the researchers' perspective, but also from the database managers'? Is it something that would be useful internally? Secondly, is it feasible? And at what cost—half a million dollars--$1 million--$5 million--$10 million? What would it take to make this happen? What are the desirable components? Start talking about sampling protocol—do you want just a *random* sample of all the industries out there? Do we want to over-sample some of the bigger industries, the more highly regulated industries, and have a random sample of dry cleaners, auto part shops, and things like that? How do we stratify the samples, if that's the choice? What should be the elements in the sampling protocol? How would we stratify them—based on the NAIC codes—the size of the firm—the location—the state? Over sampling—what do we want to do, census the *big* facilities and polluting industries? What kind of documentation [do we want]? One of the things that the people at EPA and the researchers both pointed out is that they need some documentation to explain what the data mean and what the purpose was. Finally, what kind of accessibility can be granted? –and any other ideas.

Okay, so I'll just move on down the line of panelists here and get their input.

Pat Garvey (U.S. EPA, Office of Environmental Information)
I'm the National Program Manager for the Facility Registry database that Matt referenced that includes about 1.8 million unique places regulated or monitored by either the Federal government or state government. It has about 2.2 million IDs associated with that—so those large companies, like Alcoa, that might come under the Toxic Release Inventory might have hazardous wastes, might have a discharge pipe, might have an air stack—those are all IDs, programmatic areas that we would try to link together at a given place.

The whole research community has raised five related issues, and I'll review them quickly here. One is the issue of definition of terms. At EPA we have created this large database, and for lack of a better word, we've just used the word "facility." But, the

database also has *monitoring* stations—the air quality subsystem monitoring stations (and it has the ability to have water monitoring stations)—brownfields properties, Superfund sites, all these things that are technically *not* facilities (i.e., there's not a company, an owner, or an operator there). But, the definition of terms, I think, skews the aspect of a researcher coming into a very large data set and saying, "None of these things meet the criteria that I want."

I know researchers have a strong desire for a *historical* perspective, but as Matt just said, the database is a *living* database—it needs to be *refreshed*. Dun and Bradstreet tells us that about 11 percent of all companies in the United States go *into* business or *out* of business *per year*. As soon as you go 3 or 4 years into a historical perspective, you've got potentially 44 percent of your universe either not part of the *initial* sample or not in the *final* sample because of that historical perspective.

We *do* have data gaps in key fields, and we *don't* have *key fields* defined as you [researchers] would want, such as a *primary* SIC code for a North American Industrial Code. We in the Facility Registry System collect *all* SIC codes ever reported by that company, because, as I'm sure all companies do their absolute best job in defining themselves because they have to report to TRI and other regulatory programs and have a high community right to know, they might be *creative* in the designation of their SIC code *or* a change of SIC code process, which would then throw them out of, maybe, historical perspective or trend analysis. The Agency has never dealt with trying to establish a definition of "primary SIC code," so many of our databases collect all the SIC codes, and thus you've got a big melting pot.

You've also got the other key fields that might not be populated because of burden reduction issues, so that the Office of Management and Budget or the Agency or the Administration has decided that, "Gee, for this individual data area, be it small-quantity generators for hazardous wastes or something else, we don't want to collect all the different kinds of data elements that complete a large database like the FRS."

Then, we also have the last issue—many of our *smaller, targeted* programs, such as Green Light or Energy Star and those kinds of things, deal more at the *company* level and not specifically at the facility level. So, if in your research problem you want to deal with small-quantity generators, which obviously constitute a large number, we've got databases that satisfy that need, but if you're looking at some of the *boutique* kinds of programs, such as Energy Star of brownfields or P2—you know them all better than I—their definition of "facility" or "site" versus "company" or "corporation" doesn't match up cleanly.

So, I think these are all very strong challenges that the research community has—*Good luck*!

Matt Clark
Randy will be talking about the length between Census and all the problems dealing with that. He's familiar also with being a researcher, not just being a bureaucrat.

Randy Becker (U.S. Bureau of the Census)
One of the reasons why you'd want to link EPA data to Census data is that we can fill in some of the missing information about plants, such as their size, which isn't generally collected in the EPA databases—that's the critical one. But, you can also fill in the information about industrial detail—we have our own industrial code—as well as production information, so you can look at how productive they are, how many production workers they employ, their input usage, their capital investment—and that's basic information that's collected annually on the Annual Survey of Manufacturers and the Census of Manufacturers. Plus, there are other databases as well that can be matched into those databases that I just mentioned, and most commonly we're talking about the Pollution Abatement Cost and Expenditures Survey, which was collected for a number of years and was collected again in 1999. It is now being redeveloped—we hope to have another one in the field in a couple of years.

So, there is interest in bringing the regulatory data to the census data. A couple of earlier speakers talked about the challenge of the census process, and part of that is that we get our information on what businesses are out there—we call it our business register—from the IRS. You may recall a number of years ago the IRS was raked over the coals by Congress because of its employees browsing the data, so they [Congress] took it out on the IRS and they took it out on agencies that the IRS provides data to, and they wanted to know who's looking at the data provided to the agencies. The Census Bureau is a very large user of IRS data, . . . so the IRS makes the claim that because they provide the underlying frame information, and some of that data makes its way into the final census data—that is, it's co-mingled—they think that they have some say on *who* looks at it and for what purposes, and so forth. So, this is, as of 1999-2000, very subject to a lot of regulatory oversight. Now, most of the Census Bureau is immune from that since there is legislation in place to actually use IRS data to draw samples . . . and so forth. Using the data to do research and handing it over to egghead academics doesn't necessarily sit that well with the IRS. (Of course I don't speak for the Census Bureau when I say all these things.)

But, essentially, that's the conflict that we're faced with, and we are now through the woods, as it were. There aren't much more reporting requirements and more processes in place to get things approved, as Dick Morgenstern was talking about earlier. But, there is a process in place and we're hoping that it's speeding along. So, there is a fixed cost to using Census data, but there is a great return, as well.

So, I think what we envision, and we talked about this at the September 3rd meeting, is that once EPA has their data together and all the facilities in a common identifier, one of the things that could be done is some sort of name-and-address matching with the Census data. Essentially, there would be a "zipper" file—that is, the Census ID and the EPA ID—and because *our* data is confidential (and I think some of the EPA's might be as well), but because *our* data is confidential, that zipper would reside at Census. However, any researchers who would want to link any EPA data could tap into that zipper that's been constructed and bring in whatever EPA data there are.

Some of the issues that Matt was talking about—we have 1.8 million on the EPA site and we have many more on the Census site. You know, if we're talking about sampling and restricting scope, . . . why not just do it all? As soon as you start restricting the scope, there's going to be someone out there who wants to look at something that's outside of that scope. The only real constraint is just computing power—and brain damage, to some extent—but I think it is certainly possible—the Census has experience with linking many, many more records and doing name and address matching. There's sophisticated statistical software out there to do name and address matching and so forth, so I'd say let's think big and if it doesn't work, we can always contract our motivation at that point, but let's not start out that way.

_____

Ron Shadbegian (University of Massachusetts, Dartmouth)
I *do* have a lot of experience using the Census data and linking in EPA data to that, and as a number of people have already said, it's not an easy process to go through. In the old days we didn't have these wonderful computer matching programs, so we did a lot of linking by hand. The other problem is that name and address matching is fine, but not all facilities give their particular address—they give you the address of their headquarters or some post office box—so, there's a lot of hand wringing, and you look at those with the same zip code . . . there are a lot of tricks to sort of looking and trying to identify the matches that you don't get through the computer programs, so it's not an easy thing, necessarily, to do, but as a first crack, maybe that is what we would do.

We've also linked in data from other sources like the Lockwood Directory, which is a paper industry directory that gives information such as the size of the firms in terms of their capacity, whether they use pulping or not—do they by pulp?—do they make their own pulp?—what techniques do they use? So, we've used that sort of data in the past as well, and weaved that in with census data, and we've used Pollution Abatement Cost and Expenditure Survey to collect information on how much these plants spend on pollution abatement capital, on the air side, on the water side—how much they spend on operating costs on the air side, water side, solid waste side.

Wayne Gray and I have used these data sets that we've put together, and continually put together over time, to answer lots of interesting questions. We started off looking at the effect of EPA regulation on productivity, and our EPA regulatory measure was the pollution abatement expenditures—and so that quickly became: well, do pollution abatement expenditures *overstate* actual expenditures on pollution abatement or do they *understate* actual expenditures on pollution abatement? . . . So, we've looked at things like that. We've looked at the effect of regulation on pollution abatement investment, on production investment, and on compliance, as Wayne talked about yesterday.

So, there are lots of interesting things that you can do once you make that link with Census data, so I agree with Randy here in saying that we should go for the whole thing, and if we have to start paring it down, then we can do that as well.

Michael Barrette (U.S. EPA, Office of Enforcement and Compliance Assurance)
Thanks for inviting me today—I'll just give a few remarks about facility data. Let me provide just a little bit of background: First of all, I'm actually involved in two pieces of the puzzle, which is actually *using* the data as somebody that does targeting and analysis with our Regional offices in the enforcement program, and in system design and web help tools, including our integrated data for enforcement analysis (IDEA) system, which is basically a mainframe system that's extremely powerful, but only a handful of people really know how to use it, because it's so complicated. And then two spin-off projects from that—one is what we call OTIS, which is our Online Tracking Information System, an internal web-based tool that basically gets the IDEA data out using a web browser, and then ECHO, which launched about a year or so ago, which provides data . . . and we integrate information on RCRA, the Clean Air Act, the Clean Water Act, TRI, census/demographic data, things of that nature.

So, we're very familiar with the Facility Registry System—we've been working with Pat—we've worked with him basically to put together an error tracking system, so if somebody is using the databases and they see an error, they can submit that and it goes through a fairly elaborate process of getting it to the right data steward. The thing to keep in mind when using the facility data is there are, as Pat mentioned, somewhere in the range of 2 million records. Now, those records come from EPA in many strange ways—we have different reporting cycles; we have different regulations; we have some programs where you have to notify when you shut down and others where you don't have to notify. We have some states that are maintaining their own database and then they decide to just send us uploads, say, once a month, so they don't really view that data set as what they use to manage—they may pay less attention to it. We have other states that are entering [data] indirectly. We also have some systems that are modernized, some that we call legacy, and some that are kind of in between, being modernized now. The bottom line is that facility data that winds up at EPA come in many shapes and forms, and basically Pat's program has to figure out what matches up out of all those things that come in. Over the last 5 or 6 years we've seen a big improvement in the ability to match those things, and we've also added a data steward network on top of the computer programs. But, still, if you really want to work with a very tight data set and ensure there are no mistakes, you're never going to get a hundred percent with that system because things are constantly changing.

At headquarters we don't, obviously, know what's going on in the field, and so—I don't know if I'm stating this the way Pat would, but—what we have with FRS now is pretty much what you're going to have in the future with the exception that maybe more programs will be added. I think his budget's pretty much locked in—he's got a staff that's able to do data quality, to respond to errors, to respond to people who might send in batch files and ask them to clean them up and they always clean them up very quickly—but there's a lot of information out there that people maybe have not scrubbed down or looked at, and there's always going to be some type of mistakes. That doesn't mean that

you should discount the system—you just have to keep that in mind when you're using it. One example is the fact that we have somewhere around a million facilities in the RCRA database. Well, there's no flag in RCRA that really tells you that they're not in existence anymore, so that's something we're working on—we're trying to get that in place by December 2005—to have an "inactive" flag in that database. But, you have to keep in mind that if you're using that database, as many as 30-, 40- 50-percent may be companies that are not in existence.

The other piece is that there has been some progress internally to start looking at corporations and how to match those together, but at this point in time EPA contracts with Dun and Bradstreet and it's not clear exactly what data will somebody outside the Agency use? That's kind of an ongoing thing, but it's something the enforcement office is very interested in—figuring out a good way to profile a company, because every week we get requests, whether from the Administrator or a voluntary program or the White House—somebody needs to know about a company—and there's no way that you could just press a button and say, "Okay, these are the ones that had enforcement actions, this is their TRI release, and this is how many had open violations." You'd have to do a lot of custom work to get that information. Hopefully, if we make some progress with Dun and Bradstreet, we'll be able to automate some of that.

The issue of developing a small subset, or a panel, of data that can be used by researchers—we actually have some experience in that because in 1995 the White House Reinvention Committee basically required EPA to publish data for five industries on the web, which we ended up doing in 1998 under the Sector Facility Indexing Project. Let me relate a couple of experiences with that: First of all, if you tap into the EPA systems and ask how many refineries there are, which is SIC 2911, when we ran that back in 1995 it was something like 850. That doesn't necessarily mean that the linkage data, or FRS, is wrong, but what it means is there's a lot of extra SIC code data in all these systems that it's pulling from, and anything that has any relationship to a petroleum refinery, somebody might have put in a 2911, and then all of a sudden what should be 180 facilities looks like 850. We couldn't put out the list of 850 because it was wrong, and we ended up having to—basically through brute force and a lot of grunt work—figure out which facilities were actually refineries, what were the EPA numbers across all the programs, which ones were operating and not operating, and we used some of the same sources, like Lockwood Post, the Department of Energy for the refineries, the auto industry has trade association data, and we actually got that nailed down and published the data in 1998.

But, it's a pretty intensive effort to make sure that that information stays correct, and we have to have contractor support every year to look at that information. There are all sorts of questions on definition—if a refinery is totally located with a chemical plant and you want to look at emission trends or you want to look at the release amounts versus the production, they may have reported both their chemical and their petroleum together, and how do you piece that apart? So, it's not really that easy to do—it's a very intensive effort that, I guess I would say, unless there's a dedicated staff that is spending money and doing that every year and making sure that there's data surveillance, it's very hard to

accomplish and I think that, as some of the other folks have said, you're going to get a lot of people coming in that want to do other studies.

So, I don't know what the solution is, but one of the things (since I come from the enforcement compliance program) that we've tried to do is to focus in on a subset, which we call the "majors." The "minors" in most of the programs have fewer reporting requirements—the states don't have to tell us all that information. So, unless you're looking at compliance assistance to small facilities or something, we think that the data on the majors, which would be Clean Air Title V permitees, water major dischargers, RCRA large-quantity generators—we spend a lot of time scrubbing that data. The reporting requirements are more frequent, so we know the facility data are better. Once you start entering into the world of Clean Water Act minors or things like that, you never know what you're going to get, and those of you that have used the data probably know that. So, that's *one* way that you *could* develop a subset, but I don't know if that's necessarily the best way.

The only other thing I'll say is that for those of you that *are* interested in Clean Water Act data, we are developing a PCS modernization project. Our hope at this point, and I think we have a pretty solid commitment from the states, is that when that modernization is completed, we will continue to get all the majors' discharge reports, from which we'll calculate a compliance determination. But, we're hoping that we're going to get all the minors in as well, and the reason why we think we can do that is because we're moving toward electronic reporting of that data. The biggest barrier to getting that data in is state staff having to keypunch every little thing every month. So, if we can get electronic reporting of those DMRs right from the facility into a central receiving database, then there's a good chance that we might go from the current situation of about 7,000 facilities with good data up to close to 100,000 in the water program. It's not going to happen overnight, but hopefully we'll get there soon.

Matt Clark
Pete—any suggestions on this—what you would like to see?

Pete Andrews (University of North Carolina at Chapel Hill)
I'm the first of the academic eggheads up here, so I'll try to be as hard-boiled as everybody else.

As a researcher trying to use these data, let me first comment on some of the things that have been really helpful. I want to really thank the gentleman at my left here, Mike Barrette, who has been tremendously helpful with our project, both in getting the data we needed and learning to use it and so forth, and working through the portion of our grant that it took to get these data into usable shape for what we were trying to do with them.

EPA has been working on making all the data more user friendly, and this is a great help. Simple things, like single identifiers, keeping the address and contact information up to date. This comment that was made about which ones have gone out of business—one of

the greatest unexpected learning experiences we've dealt with just in the process of our project is discovering how much of our own sample turns over just in the period of the project. So, that is very helpful.

The other thing that I think would be extremely helpful would be just relatively simple things, like more-user-friendly guides to the data sources—what *were* they collected for?—so that on the web, if you're using this for the first time, they're just easy buttons to click for so you understand how these data were collected, for what purposes, what their updating cycles are, if that's something that can be put up there. That will save *some* costs at the user end; it will certainly save a lot of costs at the EPA end of hand holding all of us individually on those kinds of questions which we've had to ask in the past. Those things are *really* helpful.

I'm more of a skeptic also—I wasn't expecting, after your comment about the nodding heads in 2003 that everybody was going to sound skeptical of the panel data project, but I *am* a skeptic of it. I think it could be a very useful thing to do for some purposes, but I don't think it would solve the problem you're trying to solve with it. There are obviously some areas of research that could come from it and that could be strengthened by it, but one of the basic problems I think it would run into is "bucket size" for doing statistical analysis—how many facilities you've got in each cell. There are so many different kinds of research that cry out to be done in this area, and when we start figuring how many facilities there are in a given sector, in a given state, of a given size, I worry about the *bias* of a panel like that towards: once you've created the panel, you want researchers to use it, *rather* than them looking at some of the new and emerging sectors. You know, we ought to have more people looking at agriculture today, but this is largely a manufacturing facilities and utilities and POTWs database. So, I worry about the bias—I worry about your putting *so much* money into building and maintaining and feeding a database like this that it would be *less* productive of the kind of research that, even with the rough kinds of issues we're having to deal with, *could* be improved by putting the money into *some* things that you're already starting to do and really helping us a lot by doing. And then, let researchers keep having their own head about the kind of questions that look important and how we work with that.

Dietrich Earnhart (University of Kansas)
I will echo most of these points. Going back to the presentation I made remotely September 3rd of last year, what I would want would be a nice, clean, unique identifier across all the databases. If that is already in place, as I understand what Mr. Garvey said, then I have what I want, in general. If 99 percent of the time it's a nice clean identifier, then I'm done—you don't need to integrate anything more for me—I'll gladly run through 1.8 million observations to match up what I need. Now, if that's the case, that's great, and I think it would be helpful to advertise that more, because actually it wasn't until I talked to Sarah Stafford in January of this year that I realized that that had been done.

I, like Ron, have hired research assistants who painstakingly tried to match facility name, facility zip code, etc., so I *don't* want to go back to that. I would echo the point that Pete just brought up in terms of going back to natural underlying databases—it would be helpful to have more accurate information on contact information. In our survey we distributed to every single last chemical manufacturing facility across the entire U.S., and we found that 35 percent of the data that EPA listed was wrong, *way* wrong, not even *close* to being right. We searched *numerous* databases to try to find the people who supposedly were *active* discharging facilities—we could not find them after months and months—we called state regulators—we called the EPA Regional offices—they were no where to be found. In addition to that, the *active* status was way wrong, even for some of the major polluters. So, in some ways, possibly, it would be helpful to clean up the data before you actually integrate it. (Sorry about that.)

Documentation would be really helpful. I will add that the people at EPA have been very helpful—I'll put my plug in for Steven Rubin, a fantastic man who has provided me with the Permit and Compliance System database on an ongoing basis. It would be nice if there actually might be a workshop of this sort to teach us, or teach us in combination with our research assistants—I know Madhu Khanna has an army of RAs working with her. If you could come and say this is how we use databases, especially if we had a database that is really powerful but a bugger to use, then I know I'll never touch it.

All of these things would be helpful as part of this integration/modernization process.

One minor point about the Dun and Bradstreet: I'm wondering how or why that was chosen as the way of connecting things—most people who are looking for financially related data use the Compustat Research Insight Database, which I know definitely does *not* have the Dun and Bradstreet number. So, it would be really nice to find somebody that is *not proprietary*. Once again, maybe there's some great logic behind how it was chosen, and maybe there's a different database that people can match to—I know Michael Lennox made a reference to that—so maybe we all can learn from his previous research.

Michael Lennox (Duke University)
As a doctoral student, I actually *did* that process of matching facilities to Dun and Bradstreet data, then matching to Compustat—the Compustat data, though, *doesn't* give you the structural trees of ownership, and Dun and Bradstreet *does*—and Dun and Bradstreet *has* facility-level data, which Compustat *does not*, so that's why you have to use Dun and Bradstreet.

Michael (Barrette), to make your job even *more* difficult, I just want to throw out another recommendation: It would be *great* to have these over time—affiliations-longitudinally—which I know is *incredibly* difficult but something that we've worked really hard on trying to capture—but it gives you some incredibly powerful statistical powers there by looking at facilities that have changed ownership and how that might affect behavioral performance. [Matt Clark interjection: You're speaking temporally

rather than geographically.] Yes. So, in other words, as Pete was suggesting, corporate affiliation and ownership change more rapidly than you would think across these samples, so you would *almost* have to have a corporate affiliation *per year* and have it reflecting how that changes over time. I know that's a *very tall* order—we've talked to Dun and Bradstreet, and they're not thrilled about dealing with past data—they just want to deal with it here and now. But, that would be *incredibly* valuable for a researcher.

Pat Garvey
We've looked at that a half dozen times and there's no regulatory statute that EPA can sort of "hang its hat on" to do that information collection. We've heard that over and over again—it's just that there's an issue of reporting burden by the regulated community to a regulatory agency, and there's always a *tension* there, as you can imagine.

Michael Lennox
I know that the TRI fields—that's what people were saying—are notoriously ill-reported: The Dun and Bradstreet Number is the wrong one or they get the wrong firm or its subsidiaries—I know it's incredibly difficult to get.

Matt Clark
I have a question: Is there a business service like Lexus-Nexus? I would think that a lot of the sales and changes of ownership would be recorded in the Wall Street Journal or things like that, so that would *not* be an information collection problem. I was just wondering: Is there a service out there, that particularly you business professors have used, that would allow us to go back in time and see what these changes have done?

Pat Garvey
We deal with the number of firms and stuff, and D & B sort of raises its hand highest among those service agents.

Dinah Koehler (University of Pennsylvania)
Wharton has this WRDS database—I don't know what WRDS stands for, but I'll find out—and it's a pretty detailed database related to business-type issues. I think the way it works, and, again, I'm talking sort of "off-line" here, but something along the lines that you can request a data set to be prepared—it's for a fee—and they then prepare that data set, and they will probably match across various different individual data sets. So, we can check up on that and give you guys more information if you're interested as one potential model. Now, it *is* run by Wharton and it *is* for a fee. It's actually a pretty good business for Wharton.

I do have a question: We are in the process of trying to match the RADP Accident Database with TRI. What I'm trying to use is the Dun and Bradstreet facility ID. Now, Mike just said that that's not great with TRI. Can you give us any advice, or is this in the FRS already so I can go use that—or do we have to go this D & B facility ID path, which seems to be the only link that we can find between these two databases?

Pat Garvey
The Department of Homeland Security does not allow me to provide Risk Management (Audit) Program, or RMP, linkages to any FRS data. That was specifically locked down on September 22, 2001, and that's not publicly available. It *is* available to EPA staff.

Dinah Koehler
So, just to qualify, Wharton is in a cooperative agreement with the EPA to work on it, given all the confidentiality issues. So, we *do* have that database and we want to work with it and try to link it with TRI, so I guess we're going to have to do that on our own.

Pat Garvey
The TRI and RMP were our very first two databases that we linked together in 1999. On the Dun and Bradstreet side, we just finished a matching with D & B, and they did match, through an automated process, 65 percent of the 1.8 million facilities to a D & B number. As Mike said, the percentage of the 80,000-84,000 "majors" that we've designated in the database even had a *higher* matching. One could, of course, extrapolate that RMP facilities normally are pretty large, or a certain threshold probably qualified as the "majors" category, but it's not available—we're not allowed to provide that information.

Pete Andrews
Let me just add another comment—maybe this is something else that ought to go into your notes, Matt, as something to work on. Certainly with the Census data, dealing with the IRS, we've had to deal with confidentiality problems and so forth. Particularly with the Department of Homeland Security taking the position it's taking in some places, it may be important to start working on the question of having a validated method for researcher access to this information. It may not be available on the internet to the general public, but certainly EPA researchers and others have got to be able to have access to these data if EPA and the country are going to get the information they need. Maybe they won't get the individual facility coordinates that a terrorist might find useful, but they certainly need to be able to use these data to do valid and useful studies.

Michael Barrette
It's been probably four years since I've even really looked at the RMP database, but from what I remember, the reporting form requires an EPA ID to be provided, and I think it isn't really specified exactly which ID needs to be provided, but I think we recommended in the instructions that you would start with the TRI, so I'm not sure in the RMP program itself what they provide out of their own database. But, if they do provide any of that data and you can get that one field that has the EPA identifier that was self reported, you may be able to construct some of that—I'm guessing.

Pat Garvey
I was in conversation with Homeland just three months ago, and they just made another absolute comment about lockdown on RMP.

Michael Barrette

I believe my comment is related more generally, I think—it might be worthwhile thinking about this in terms of putting this database together, because it's not only researchers who are going to be concerned about this, but there's, I think, some sort of a partnership directly with DHS on this problem of information provision. Letting RMP [data out] is not the only kind of issue—in most of the very important environmental risks that many researchers are concerned about, whether it's nuclear or chemical facilities, oil refineries, water utilities, a lot of these are really potential targets from a Homeland Security point of view. Therefore, it makes sense at the beginning, as you're putting this together, to maybe find some common ground or groups.

I think, also, maybe broadening it to provide incentives rather than only looking at it as an obstacle to research—because I really think there could be some connections on the informational regulation theme, where there have been some potential benefits of broader public information from the environmental protection point of view. The same might be the case from a domestic protection point of view. Some of this information could be put together rather than being at cross purposes. I think otherwise you're really going to run into very serious problems in other fields when you start to put a lot of detailed information together, you start running off facilities, specific locations, and things like that—my guess is that that's when our friends in the anti-terrorism world are, for good reasons, going to sort of become concerned about it.

Matt Clark
I know from working with DHS that our STAR grant researchers actually *do* have to get clearance on some of these things. RFS has also done some research looking at issues about the risk of having too much information available, so that might be able to be continued.

Eric Orts (University of Pennsylvania)
A short story: Just three weeks ago, we had a number of letters come into EPA from local government and county government that said your public website had the words "water tower, water intake, water processing, water association" and they asked that you please strike off facilities that have those combinations. There's a *tremendous* pressure out there—from governors, mayors, county commissioners—to take a lot of information down from public access.

Matt Clark
Maybe research access should be different than public access, and having gone through a screening process, maybe that would work. The Census has a process.

Joel Garner (Joint Centers for Justice Studies, Inc.)
I'm working with Sally Simpson at the University of Maryland on a Justice Department study using the PCS data. I want to re-emphasize that a *big* problem is connecting facilities to firms. Any help we can get on that would be greatly appreciated.

I'd like to suggest a *small* solution as opposed to a "big guns" solution: I think it would be useful if all the projects funded under this program were required to place their data

set at the end of their project in the public archives. Wayne Gray's particular project—those data—publicly funded data—should be sent to the University of Michigan or some other public archive so that others can use that. You can get *more* research per dollar spent—other researchers can use it. Researchers *don't like* giving up "their children" like this. It's just good science—it's good public policy.

Matt Clark
We require it. There are going to be limits to that, of course, with the Census data.

Joel Garner
And there are *other* public archives—Michigan also has mechanisms whereby confidential information is resided there, and it's available only when the original investigator approves researchers using it. The point is there are other places and mechanisms to do that, and it's a . . .

Randy Becker
But there *are* such mechanisms here as well. These data are archived and there *is* a process for getting at it.

Joel Garner
If I wanted to come and use the previous data, I could call you and I could use those data?

Randy Becker
Well, you have to submit a proposal to actually use those data.

Joel Garner
Right. I just think that's a very important thing to do with existing projects. The second thing I would suggest is bringing professional data archivists to the task. There are people who *know* the data and work generating it—it's a *very* different profession, and they like doing this! And they're *good* at it. And the staff *pays* them to do it. They do *good* documentation—we don't have to do that work—they can do it for us.

And then the last suggestion is: Maybe working with the states, generate the data originally, and you talked about doing some of that. If *that* work could be done *originally*, then you don't have to do it in the archives.

Pat Garvey
Matt, I'd be remiss not to tell you that at the facility level there's another large program at EPA called the exchange network. It's got about $20 million a year behind it that we give to States and Tribes. We've made arrangements already with *20 states* to, on a regular basis—as Mike was saying, every two weeks to a month—exchange their master facility record with EPA's master. In that exchange effort, we have a goal of 35 states by the end of this calendar year, so that exchange network . . . is *aggressively* moving forward.

Sarah Stafford (College of William and Mary)

I heard this comment from a couple people—this is straight to the EPA folks: I find mistakes all the time and I don't report them. Do you want me to report them, and if so, how? If I'm trying to do the linkage, and I find a matching that's not there, should I report that? Should I just use the report function in Envirofacts? Is there a better way—to do it as a batch? I'm also concerned, because, you know, it's a 95 percent hunch, because I've done a *lot* of connections, but I could be wrong too. So, what type of information do you want from us to help get these databases better? I think probably everyone in this room has found a number of mistakes—has anyone ever reported them? I find a lot, every day, so any information you have to help me help you would be great.

Pat Garvey
I'll let Mike answer, but I think we had 17,000 reported error notifications on the first year of release of ECHO, and Envirofacts gets about 310-325 per month. We don't see a lot, so *anybody* who finds especially issues of linkage IDs that are poorly done or wrong lat-longs or wrong names of facilities because they've gone out of business, *please, please, please* report. As Mike said, I run Error Tracker *and* I run Facility Registry System, and I'm also the staff director of the exchange network, so I'm in a pretty much unique position.

Michael Barrette
Just quickly on that, in terms of the practicality of reporting errors, first of all, the web site makes it very easy one at a time. So, if you're actually just playing around on the web site—ECHO, Envirofacts, or whatever—and you see an error, hit the button, and if it's a linkage issue it's going to go to Pat's staff, and they're going to figure it out. What I found is when I do a detailed targeting project and I'm looking at, let'say, comparing air releases in one system to TRI air releases or permit data, and I'll look at almost 700 facilities and maybe 200 of them or 100 of them or whatever aren't lining up right. I'll kind of scrub those down, and I'll send a batch file over to Pat and say, "These things look suspect—they're reporting TRI releases to water but they don't have a water permit attached to it, or whatever it is. Normally, he fixes it within 2-3 weeks. If you're dealing with a large data set—I don't know if Pat's willing to do this—but it's easier to send him the file than to key punch them [errors] in one at a time to the web site.

Pat Garvey
I'll take it *any* way—whenever you have a large data set at this level, information and feedback is the most important thing.

Lori Snyder (Harvard University)
This is another solution—I almost think that this already exists—a list-serve, where researchers who are using these databases could subscribe, and then we could post when we discover something—perhaps not every single error we find, but general things that we find that might be of interest to the broader research community could be shared that way, because I know a lot of us are in different fields—we don't always go to the same conferences—we're not always talking to one another—so that might be a way to facilitate communication.

Deanna Matthews (Carnegie Mellon University)
Something that would be helpful—I now know Michael Barrette's name, so I know where to go to for help with these data sets. Perhaps as a Project Officer, something that could be done is assigning somebody who does have control within the database area to each of the new projects, so they have a contact to go to rather than just the link on the bottom of the data page—so that they can say, "I'm looking for this specific data. Who should I go to to find help on that?"

Irene Xiarchos (West Virginia University)
One comment I wanted to make just because it's related to the idea that maybe we should have different access for the research community and for the public. The research community is very broad, and sometimes it includes students that may not be able to be in direct communication with data links, so they are pretty much between the public and the research community, but a lot of research is going to come out from them. I wanted to come out and say that because I am a graduate student.

Another question I have, a personal question if anybody can answer—I don't know *who* could answer, so that's why I'm posing it here—is there any data that you know of on recycling, but for the industry—for facilities—not municipal?

[There was no response from the panelists or the audience, and Matt Clark closed the session.]