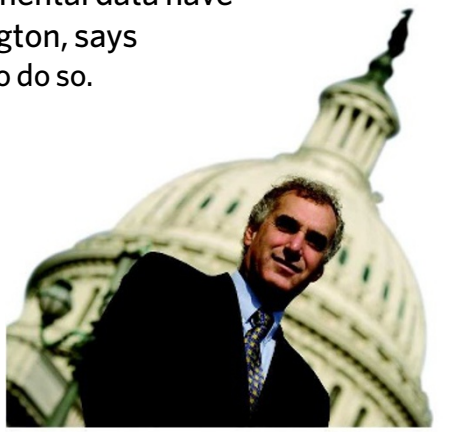
Data wrangling

Collecting and releasing environmental data have stirred up controversy in Washington, says **David Goldston**, and will continue to do so.



PARTY OF ONE

Data sound like a grey, non-partisan and unemotional topic for political discussion. But decisions on what data to collect and release for use in research or policy-making are hardly neutral in their impact. This may be clearest in the arena of environmental policy, where hard-fought disputes over the collection and dissemination of data frequently break out. Indeed, perhaps the only thing politicians agree on about environmental data is that more data are always better — in theory.

Maybe only in theory. One of the continuing debates has been over how much data collection to fund. Although advocates on both right and left frequently call for more data — if sometimes just to delay decision-making — sensors that measure such matters as air and water quality are often the first items to be cut when budgets get tight. And efforts to develop a set of environmental indicators that would be regularly updated — something akin to economic statistics — have not got very far. This can be seen in the second *State of the Nation's Ecosystems* report (www.heinzcenter.org/ecosystems), released in June by the Heinz Center, a private think tank in Washington DC. Although the report includes many environmental measurements, it is chock-full of lists of subject and geographical areas for which few if any data exist. Indeed, a companion volume urges a significant expansion of federal funding for environmental indicators.

Such an expansion, though warranted, seems unlikely any time soon. Politicians like to talk about the need for more data, but it is rarely anyone's top priority. In fact, when the Bush administration responded to the Heinz report by announcing that it would put more money into indicators — a move only symbolic this late in its term — one environmental organization took the White House to task for spending money on measuring pollution rather than cleaning it up. Broad data collection not connected to any single controversy isn't very sexy and must compete with many related activities presumed to have more immediate impact (although it may be hard to tell without the data). Even when instrumentation is regularly funded, as some kinds of satellites are, money is often lacking to maintain the data or to make them sufficiently accessible or digestible.

And if data collection and processing were to be institutionalized, another ongoing debate would have to be resolved — how insulated the operations should be from politics. For years, there has been talk of establishing a Bureau of Environmental Statistics (BES), which would not only gather data but also analyse them. Data are never as straightforward a matter as they seem; just deciding what information to collect involves judgements about what's important. You don't measure, say, pesticide levels in food unless you think they're a problem. And deciding that a problem exists is different from deciding what to do about it. The frequently heard claim that 'the data speak for themselves' has to be one of the most misleading sentences in the English language.

Still, a statistical agency needs to be free from political manipulation to have any credibility. Around 2002, I was involved in lengthy, closed negotiations on Capitol Hill between moderate and conservative Republican congressmen interested in setting up a BES. But the effort was eventually scuttled when the Bush administration rejected all proposals to keep the agency at arms length from politics, arguing that the heads of all major agencies should be responsible to the president.

Even without a BES, the US government releases a lot of environmental data. Much of this is information to determine compliance with regulations, but increasingly just making data available is seen as a way to encourage companies to clean up their operations. The model for such efforts is the Toxic Release Inventory (TRI), established by Congress in 1987, which requires companies to publicly report their annual emissions of certain chemicals. The TRI has resulted in substantial cutbacks in emissions as companies try to 'green' their reputations. Bush administration proposals to save industry money by reducing the frequency of reporting or the number of companies required to report have met with widespread opposition. But expanding reporting has also been controversial. Chemical firms have resisted reporting what chemicals they use (as opposed to release), arguing that doing so would reveal too much about their operations to competitors and would provide too much information for potential terrorists.

Industry has also been concerned about alleged inaccuracies in data on government websites. The Data Quality Act, enacted in 2000, requires federal agencies to enable private parties to challenge the accuracy of information being disseminated by the government. The law has been anathema to environmental groups, which have seen it as a way to stymie regulation. And it has been primarily invoked by corporations questioning studies that raise alarms about their products. The statute was written after academic researchers declined to release the raw data behind epidemiological studies that were being used to toughen clean-air regulations in 1997, citing privacy concerns.

Data sharing by individual, non-governmental scientists has increasingly become a topic for public debate. Charging that a scientist has been unwilling to share data is a good way for politicians to raise suspicions about someone's work, especially when the work itself is too technical to be easily evaluated by laymen. But different fields have different mores about data sharing, and the issue is not clear-cut. For example, Michael Mann, the author of a controversial study on the history of Earth's temperature — the 'hockey stick' graph — was attacked by conservatives for not sharing his data. But what he had actually held close was not his data, but his computer code, which he claimed, with government backing, was his intellectual property. He did eventually release the code.

In the political sphere, talking about the need for public data is always a good way to sound objective and above-the-fray. But data are a complicated matter, and by themselves rarely resolve an underlying controversy or problem. Nonetheless, the siren song of data has a long history. When the idea of publicly collecting and releasing information was in its infancy, an eighteenth-century Enlightenment thinker proclaimed that statistics and tyranny were incompatible. That turned out to be untrue, too. ■

David Goldston is the former chief of staff of the House Committee on Science. Reach him at partyofone@gmail.com.

See Editorial, page 1.