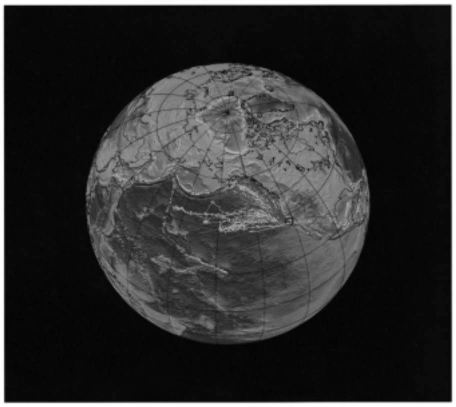

CONTRIBUTIONS

Commentary

How to Manage Data Badly (Part 2)

Preamble. Although managing data badly is difficult, given new developments in hardware and software and many sound recommendations for how to manage ecological data (ESA 1995, NRC 1995), it is still being done. In the spirit of Howard Wainer's (1984) article in *The American Statistician* on "How to Display Data Badly," Part 1 of this essay gave 10 rules for those database managers and administrators who are determined to manage their data badly. Part 2 gives some tips for scientists with the same objective.

Techniques for scientists

Rule 11. Hoard your data

The best technique for scientists to use in mismanaging data is simply to hoard the data. One fringe benefit of hoarding is that it eliminates the need to write those annoying metadata files that are apparently so useful to people trying to steal your results. The argument that citizens have a right to data that were collected using the citizens' money, sounds like a Utopia where the citizens might be smart enough to interpret the data correctly. Untold damage can result when people who cannot conceivably understand your data analyze them on their own and draw their own conclusions. Resolute data hoarders can carry their data to the grave. According to a report by the Ecological Society of America (ESA 1995), there are ". . . numerous incidences of researchers ending their careers without the resources to make provision for the curation and maintenance of their long-term data sets." Data sharing in many collaborative efforts is inhibited by ". . . unclear responsibilities, conflicting goals, misunderstandings, and outright rivalries" (NRC 1995).

Rule 12. It is better to get than to manage

Do not be misled by those who have tried to encourage slovenliness in scientists by proposing that they be rewarded for publishing high-quality data sets in the same way they are rewarded for publishing scientific papers. This could lead to the specter of a publishing person perishing. Do not waste time managing data that have already been used in a journal article. Axiom: It is always better to collect new data than to spend time managing existing data.

Corollary: Get into the field to collect new data before thinking about how those new data will be managed. After-the-fact databases create the same healthy tension as after-the-fact experimental designs.

Rule 13. Avoid tedium

How can you counter the tired cliché that because data are expensive to collect they are therefore worth saving? Simple: make them not worth saving. These next rules show two paths to follow, both of which allow you to avoid tedious work: (1) Do not verify the accuracy of your data; and (2) avoid writing metadata (information about data); instead, let the assumptions, processing steps, and quality of the data remain a mystery to future users.

Amusing things can result when data are not quality assured. For example, who would suspect that a field researcher could make the simple error of swapping latitude with longitude? Let those GIS people (Rule 8) try to lie about why they plotted a station from the east coast of the U. S. on the Greenland ice cap! Nothing is more effective in driving away users than an error-ridden database. Celko (1994) reminds us of opportunities to corrupt good data in the article "When good data goes [sic, but nice illustration of Rule 6] bad." Bad data are so powerful they can make the most elegant database system useless.

The only flawless database is one that has never been used for data analyses. Inspiration for analyses of bad data may be gained from the work of Huff (1954), who gave helpful advice in his book *How to Lie with Statistics*. But remember that although statisticians may be willing to belong to ". . . that group of people whose aim in life is to be wrong 5%

of the time" (Kempthorne and Doerfler 1969), people responsible for data can rarely afford to be this careless.

Rule 14. Avoid boredom

Although metadata are essential for sharing the data needed to understand and deal with the multidisciplinary complexities of natural systems (Michener 1998, Vogel 1998), they are boring to prepare. Just because database management software is fastidious about keeping track of everything it ever did, does not mean that people have to be so inclined. Perhaps you can pick an excuse from Vogel's (1998) account of "Why scientists don't write metadata" when submitting to the Global Change Master Directory. Omitting key metadata can be a foolproof way of rendering a data set useless, as shown by one study of benthic invertebrate communities, where the authors neglected to state what size of sieve was used. Codes in the data set that are not defined in the metadata can be really annoying to users. Another study, conducted by a group of "partners" measuring estuarine health, adopted the popular "Little Red Hen Syndrome," in which everyone wanted to analyze the data and write papers, but no one wanted to document the data. This study had to hire a metadata social worker to get the job done.

Rule 15. Integrity in integration

Apples and oranges can be mixed if they are first converted to juice. Database managers sometimes do this by squeezing all sorts of data into generic database models. As Katha Upanishad says, quoted in Martin (1976), "Who sees the variety and not the unity, wanders on from death to death." (This sounds like some data systems we know.) Scientists can benefit from this technique when they integrate data from different sources of mixed methods and unknown quality. If they avoid writing metadata (Rule 14), who will know the difference? Rotten apples and oranges can be integrated in a compost pile; rotten garbage in, fertile compost out. Further, we can frustrate the practitioners of the statistical art of meta-analysis when they try to rip out of our research articles the vital organs of sample size, mean, and variance. Apparently, we are already doing a good job at keeping these things out of our papers (Gurevitch and Hedges 1993).

You can use a database to integrate data that are too different to be integrated by any other means. This lets you solve an experimental design problem (such as different sampling methods) with a technical solution, as exemplified by one national study of fish pathologies. In another case, a report on watersheds made heavy use of Rule 13 (avoid quality assurance) to integrate data from widely different sources, and tried to make it palatable by use of Rule 8 (show the data on GIS maps). What they might have done is to include another GIS map showing the spatial distribution of data density and sample variance.

Rule 16. There is not much to learn about managing data

Spreadsheet software is so wonderfully versatile that you can use it for everything. A spreadsheet is the perfect tool to use for storing data. You can enter data without being slowed down by meddlesome error-checking routines; within a single column, you can mix units or methods or data formats and enter text into numeric fields; you can calculate new columns from existing columns and easily increase the number of significant digits with the push of a button; or you can add extensive footnotes to individual cells so that no cell can be interpreted in the same way as other cells in the column.

Given the power of spreadsheets, why invest in data managers and their arcane software? Samuel Johnson, in his 1755 preface to the *Dictionary of the English Language*, described the proper attitude toward data managers (lexicographers) as people ". . . whom mankind have considered, not as the pupil, but the slave of science, doomed only to remove rubbish and clear obstructions from the paths, through which Learning and Genius press forward to conquest and glory, without bestowing a smile on the humble drudge that facilitates their progress." An NRC (1995) committee concluded that ". . . there is a critical need to educate scientists about data management principles and to foster improved working relationships between scientists and information management professionals."

Conclusions: reaping what you have sown

Research projects that can bring database managers, scientists, and administrators together and apply several of the above techniques simultaneously will enjoy widespread recognition. The National Research Council (NRC 1995) reviewed the data management practices of six environmental research programs and condensed their findings into 18 recommendations and "Ten keys to success." The latter included such homilies as "Be practical," "Use appropriate information technology," and "Account for human behavior and motivation." Of course, as we all know from Martin's (1976) work almost 20 years earlier, few people would ever actually follow such advice. Alas, these are in effect "Ten keys to failure, or why nobody ever learns anything from lessons-learned reports." Recognition can also be earned from the popular media. One method (failure to share data effectively) was implicated in the very newsworthy nuclear accident at Three Mile Island (Gordon 1997). Although most groups cannot hope to achieve such spectacular results, assiduous application of the 16 rules in this paper will be sure to lead to some form of debased data.

In fairness to Wainer (1984) and his bad graphics, we must admit that data that survive being badly managed can then be badly displayed, thereby achieving the best from both worlds. Or, if displayed cleverly, the wretched data can be offered the comfort of appearing believable. We have seen that many organizations are already making wide use of the 16 rules (and others) to diminish the quality and usefulness of their data. Well done! However, we must never let down our guard against the many research groups who knowingly violate these principles. New developments keep creeping in, making it increasingly difficult to claim that technology limits good data management. And just when you thought you had buried some data for good, along come "data rescue" missions to pull out the corpses and reincarnate them into new databases. We must be strong, ignore those National Research Council recommendations, and keep those horror stories coming!

Epilogue

Although this essay is written in an ironic style, the topic is serious and deserves more attention from administrators, scientists, and database managers, working together. The importance of managing data well increases as the volume and types of data increase and as more information is needed to manage natural resources in an increasingly complex society. Understanding the interactions between humans and the environment requires good-quality long-term data collected at different spatial scales from many scientific disciplines (ESA 1995). Because no single group can collect all the data needed for the necessary analyses and models, we must learn to share data effectively.

The 16 rules given here illustrate only a few of the areas where many of us go wrong, but they point to techniques we could use to do a better job. Three basic things are needed. (1) Place good quality data sets where they can be obtained. (2) Make entries in data directories so data sets can be found. (3) Write metadata files so data sets can be understood. Of course, much more can be done and is often appropriate, but the data system must be sustainable. Administrators and scientists need to make long-term commitments to manage data well, to invest adequate equipment and people, and to better understand data management. Data management people need to understand the science behind the data, to listen to what the administrators and scientists need, to use common standards (and common sense), and to build data systems at the appropriate scale. All three groups would benefit from careful study of the recommendations in the NRC (1995) and ESA (1995) reports.

Too many science data sets collected 20 years ago, back in the time of James Martin's book, are not as useful to us today as they could be because the three basic things listed above were not consistently done well. Even today, too many data sets being created are less than what they could be because those three things are not consistently done well. Will there be a need for a future James Martin or NRC committee to remind us again, 20 years hence?

Acknowledgments

I am grateful to numerous colleagues for supplying the inspirational examples because I swear that I have never used any of the 16 rules in my own work. Nor do the rules reflect the views of my employer, the U.S. Environmental Protection Agency, and no official endorsement should be inferred. I thank Ken Rahn for comments on the text.

Literature cited

Celko, J. 1994. When good data goes bad. American Programmer **7**:17.

ESA (Ecological Society of America). 1995. Report of the Ecological Society of America Committee on the future of long-term ecological data. <http://esa.sdsc.edu/FLED/ FLED.html>.

Gordon, M. D. 1997. It's 10 A.M. Do you know where your documents are? The nature and scope of information retrieval problems in business. Information Processing and Management **33**:107–122.

Gurevitch, L., and L. V. Hedges. 1993. Meta-analysis: combining the results of independent experiments. Pages 378–398 *in* S. M. Scheiner and J. Gurevitch, editors. Design and analysis of ecological experiments. Chapman and Hall, New York, New York, USA.

Huff, D. 1954. How to lie with statistics. W. W. Norton, New York, New York, USA.

Kempthorne, O., and T. E. Doerfler. 1969. The behavior of some significance tests under experimental randomization. Biometrika **56**:231–248.

Martin, J. 1976. Principles of database management. Prentice-Hall, Englewood Cliffs, New Jersey, USA.

Michener, W. K. 1998. Ecological metadata. *In* W. K. Michener, J. H. Porter, and S. G. Stafford, editors. Data and information management in the ecological sciences: a resource guide. LTER Network Office, University of New Mexico, Albuquerque, New Mexico, USA.

NRC (National Research Council). 1995. Finding the forest in the trees: the challenge of combining diverse environmental data. National Academy Press, Washington, D.C., USA.

Vogel, R. L. 1998. Why scientists have not been writing metadata. EOS, Transactions, American Geophysical Union **79**(31): 373, 380.

Wainer, H. 1984. How to display data badly. American Statistician **38**:137–147.

Stephen S. Hale
Atlantic Ecology Division
U.S. Environmental Protection Agency
27 Tarzwell Drive
Narragansett, RI 02882