Beyond words: effective graphics and metadata are keys to concise scientific communication

Geoffrey M. Henebry

Published online: 5 November 2011
© Springer Science+Business Media B.V. 2011

Eric Gustafson (2011) urges us to rethink the format of the research paper in light of (1) a growing literature in general and online in particular, and (2) web-enabled technologies that can facilitate access to voluminous supporting materials, including original and processed data, details about materials and methods, and summaries in terms of tables, graphics, visualizations, simulation outputs, codes, tools, and more.

I concur that changes are needed to improve the clarity, utility, accessibility, and durability of the literature. Here I want to argue that, to make such improvements, far more attention needs to be focused on the quality of our graphical communication and on

metadata broadly construed. My intention here is to point to some resources that can help authors improve the quality of their graphical communication and to consider the value of metadata in making their research discoverable and durable.

Graphs

In my years as a reviewer and an editor at *Landscape Ecology* and other journals, the graphical communication has often been a critical weakness in submitted manuscripts. There is an apparent reluctance to provide richly informative captions for the graphs. Let the caption not only explain what the graph is but also what it means. Don't just rely on the text in the body of the paper to explain the graph; some readers may not venture that far! Structure your story around the graphs and enable the captions to capture the key points of your paper. Another factor contributing to poor graphical communication is, no doubt, the passive acceptance of default settings on software that support graphing. There is also a lack of familiarity with graphical perception theory among ecologists and a strong tendency to mimic the design of published graphs. But how to design an effective graph? It helps to know what works and what doesn't.

Cleveland's pioneering studies on graphical communication in science (Cleveland 1984; Cleveland and McGill 1985) remain germane in the digital era. Indeed, with the rise of powerful yet affordable

G. M. Henebry (✉)
Geographic Information Science Center of Excellence,
South Dakota State University,
Brookings, SD 57007-3510, USA
e-mail: geoffrey.henebry@sdstate.edu

graphics programs and the concomitant decline of graphical design staff to support scientific illustrations, the conclusions may be more relevant now than then:

When a graph is constructed, quantitative and categorical information is encoded, chiefly through position, shape, size, symbols, and color. When a person looks at a graph, the information is visually decoded by the person's visual system. A graphical method is successful only if the decoding is effective. No matter how clever and how technologically impressive the encoding, it fails if the decoding process fails. (Cleveland and McGill 1985, p. 828)

Decoding can be impaired if the graph contains errors. Cleveland (1984) surveyed 377 graphs appearing during 1980 in volume 207 of *Science* to count four types of errors in the graphs: (a) errors of construction—something was amiss in labeling, scales, etc.; (b) errors of degradation—some aspect of the graph was missing or obscured due to poor reproduction; (c) errors of explanation—something appeared on the graph that was not explained in the caption; and (d) errors of discrimination—items on the graph, such as symbols or lines, could not be readily distinguished due to design or size of the graph. Thirty percent of the graphs contained at least one error. Errors of explanation were most common, followed by errors of discrimination. (My recent experience has been that errors of discrimination are now more common than errors of explanation.)

In subsequent work, Cleveland and McGill (1985) set forth a theory of graphical perception, identified elementary sensory tasks in graphical perception, and described the results of experiments that evaluated the performance of these tasks. They provided a rank ordering of elementary perceptual tasks by accuracy, arguing that graphs should be constructed around those higher-ranking tasks to improve the accurate perception of the information encoded in the graph:

1. Position along a common scale,
2. Position on identical but nonaligned scales,
3. Length,
4. Angle or Slope (with 0 not too close to 0, $\pi/2$, or π radians),
5. Area,
6. Volume or Density or Color Saturation, and
7. Color Hue.

Note that discrimination of relative position is most accurate, that discrimination of length is more accurate than area, and that encoding with color is least effective. In addition to these papers, the classic work of Edmund Tufte (1983) is recommended, and a recent work by Stephen Few (2009). In addition, there are several useful blogs that deal with data visualization, including Kaiser Fung's Junk Charts (<http://junkcharts.typepad.com/>) and Nathan Yau's Flowing Data (<http://flowingdata.com>).

As maps are frequently used to display information in landscape ecology, it is appropriate to point out the challenge of using color to encode information. The perception of color varies among the population far more than other basic cognitive abilities. Between 5% and 8% of males suffer some form of weakness in color perception; few females are impaired. Using the wrong color scheme (such as the common default of rainbow) can obscure for a significant portion of the audience the information and contrasts encoded in the map or image. A series of short pieces in EOS a few years ago illustrate both dimensions of the problem and solution pathways (Light and Bartlein 2004, 2005; Stephenson 2005). The online resource ColorBrewer (<http://colorbrewer2.org>) is an excellent place to learn about the strengths and weaknesses of various color schemes. Finally, care also needs to be exercised in the use of color in line, bar, and scatter graphs. Learn to practice robust encoding of differences by using different shapes or line types *in addition to* colors.

Metadata and more

Metadata inform us about data. Indices, catalogs, gazetteers provide the *what* and the *where*, but metadata are distinct in that they provide important contextual information about the *what* and the *where*, including the *why*, the *how*, the *by whom*, and even the *how well*. Metadata enable data to retain coherence and relevance. Think of metadata as preservatives for data, something that is especially needed in this era of ephemeral formats.

There are federal metadata standards for geospatial data (Content Standard for Digital Geospatial Metadata (CSDGM) Version 2—FGDC-STD-001-1998; http://www.fgdc.gov/standards/projects/FGDC-standards-projects/metadata/basemetadadata/v2_0698.pdf) and extensions to handle biological data (Biological Data

Profile—FGDC-STD-001.1-1999; <http://www.fgdc.gov/standards/projects/FGDC-standardsprojects/metadata/biometadata/biodatap.pdf>). In addition to these standards, there are other specialized metadata formats developed by various groups to address the specific needs of their community. For instance, data commonly produced in ecological research are non-geospatial (Michener et al. 1997), that is, they lack referencing to a specific geographic coordinate system, even if they are intrinsically spatial and locational, such as museum voucher specimens of organisms collected on expeditions in earlier centuries. The Ecological Metadata Language (EML 2.1.0; kn.bioinformatics.org/software/eml/) aims to provide support for combinations of data that are common in ecological research using XML (eXtensible Markup Language) schemata.

If *Landscape Ecology* pursues the kind of restructuring of publications advocated by Gustafson (2011), then careful consideration will need to be given to what kinds of metadata to require for online ancillary material. Michener (2006) describes three levels of metadata functionality: (1) supporting data discovery; (2) facilitating data acquisition, comprehension, and use by humans; and (3) enabling automated data discovery, ingestion, processing, and analysis. This last level offers the great promise of scientific workflows that could automatically generate and deliver data products in forms that would alleviate much of the tedium associated with data preparation and thereby increase the time available for data interpretation. Such functionality will require the production of taxonomies, controlled vocabularies, and ontologies that enable the *semantic web* (Ramachandran 2010). An ontology provides a formal, explicit specification of a shared conceptualization (Gruber 1993). The word “formal” here refers to the fundamental requirement of an ontology being processed and understood by machines. Translating our ecological understanding into maps of knowledge that machines can understand is not a simple task. Much technical research is required to get there, many incremental rounds of implementation, and even cultural changes in how we value, as a community of scientists, value both metadata and the effort it takes to produce it (Michener 2006). There are some small steps we can take within the journal toward this goal, such as producing a standard keyword list and requiring its use (in addition to ad hoc keywords) for

submissions to *Landscape Ecology*. Controlled keywords provide digital stubs that are discoverable across the web. The DataONE project (<http://www.dataone.org/>) provides a cornucopia of information on standards, tools, and best practices for data management.

As we discuss whether to reinvent the research paper and how to do it, let us note that not too long ago a small group of researchers at Carleton University—including the distinguished landscape ecologist Lenore Fahrig—sought to reinvent the process of publishing a scientific journal. After a few years of intense work, they debuted in 1996 a pioneering, free, online only, peer-reviewed scientific journal, *Conservation Ecology* (www.consecol.org). It published under that name from 1996 to 2003 and continues now under the broader moniker of *Ecology & Society* (www.ecologyandsociety.org). It is easy to overlook their achievement today when open access journals proliferate and mainstream journals are all online, if usually behind paywalls. Yet big things can happen when people decide to act!

To conclude, if the goal of restructuring the conventional format of the research paper is to improve scientific communication in a concise package, then more attention should be directed to constructing effective graphical displays of data accompanied by richly informative captions and to producing rich metadata that can enable data discovery and more.

Acknowledgements I thank Jingle Wu and Eric Gustafson for the opportunity to respond to Eric’s editorial.

References

- Cleveland WS (1984) Graphs in scientific publications. *Am Stat* 38(4):261–269. <http://www.jstor.org/stable/2683400>
- Cleveland WS, McGill R (1985) Graphical perception and graphical methods for analyzing scientific data. *Science* 229(4716):828–833. doi:10.1126/science.229.4716.828
- Few S (2009) Now you see it: simple visualization techniques for quantitative analysis. Analytics Press, Burlingame
- Gruber TR (1993) A translation approach to portable ontologies. *Knowl Acquis* 5(2):199–220
- Gustafson E (2011) Publishing landscape ecology research in the 21st century. *Landscape Ecol* (this issue)
- Light A, Bartlein PJ (2004) The end of the rainbow? Color schemes for improved data graphics. *EOS* 85(40):385–391
- Light A, Bartlein PJ (2005) Reply. *EOS* 86(20):196
- Michener WK (2006) Meta-information concepts for ecological data management. *Ecol Inf* 1(1):3–7. doi:10.1016/j.ecoinf.2005.08.004

- Michener WK, Brunt JW, Helly JJ, Kirchner TB, Stafford SG (1997) Nongeospatial metadata for the ecological sciences. *Ecol Appl* 7(1):330–342
- Ramachandran B (2010) Enabling technologies to improve the data discovery and science utility of the LP DAAC data collections: an overview of the semantic web technology, and how it would help improve the data discovery and science utility of the LP DAAC data collections. USGS white paper, 12 pp
- Stephenson DB (2005) Comment on “Color schemes for improved data graphics” by A. Light and P.J. Bartlein. *EOS* 86(20):196
- Tufte ER (1983) *The visual display of quantitative information*. Graphics Press, Cheshire