"Truthiness" and the National Soil Information System: How Are They Related?

Richard L. Livingston

While listening to the news recently, my attention was caught by one of the short closing stories. The reporter stated, “The word 'truthiness' has been chosen as the 2005 word of the year by a group of linguists." After some research on the internet, it seems this story originated with the American Dialect Society (www. americandialect.org).

The origin of **"Truthiness"** is attributed to Comedy Central's satirical news show, the Colbert Report “Truthiness" is defined by the American Dialect Society as **"The quality of stating concepts or facts one wishes or believes to be true, rather than concepts or facts known to be true."**

I usually try to put these "words of the year" into some kind of personal perspective. This new word seems to lend itself to my professional life. Directives to populate soil data in the National Soil Information System (NASIS) per USDA-NRCS National Bulletin 430-5-7 (USDA-NRCS, 2005), dated August 25, 2005, have prompted much discussion among soil scientists about data—quality, quantity, and integrity, as well as scientific methods, professional ethics, and data population methods.

In the search of alternate definitions for "truthiness," I went to Merriam-Webster online (www.m-w.com) only to find they too have a list of "2005 Words for the Year." Based on user searches for 2005, the word "integrity" was named the Merriam-Webster word of the year. I find this strikingly related to the newly coined "truthiness" word of the year named by the American Dialect Society. The first definition for integrity listed by Merriam-Webster is *"firm* adherence to a code of especially moral or artistic values." Here was another "word of the year" by a different entity that relates to the above-mentioned National Bulletin on data population.

I consulted some related literature for guidance while attempting to justify the requirements of my data population assignment, including “The Soil Survey as a Paradigm-based Science" (Hudson, 1992), "Soil Interpretation in the Soil Survey" (Kellogg, 1961), and the book Data Reliability and Risk Assessment in Soil Interpretations (Nettleton, 1996).

A good account of my dilemma can be given from my work populating data in NASIS for the published Soil Survey of Marion County, TN, issued in April of 1958 (Elder et al.,1958). The following data elements were required to be populated for hydric soil inclusions by December 31, 2005 per National Bulletin 430-5-7: component name, composition, component kind, slope, T factor, wind erodibility index, wind erodibility group, drainage class, hydrologic group, (current) taxonomic classification, flooding and ponding frequency, ponding and soil moisture status depths by month, restrictions such has lithic or paralithic contact, duripan, fragipan, and densic horizons, hydric rating, geomorphic description feature name and modifiers, horizon designation, horizon depths, particle size separates, organic matter, Ksat, linear extensibility, Kw, Kf, pH, and a representative texture modifier and class for each horizon.

In the Marion County, Tennessee Soil Survey, specifically in the “Hamblen loam" map unit description, the following sentence mentions some of the inclusions in this map unit:

> *Although most of it [the map unit] is imperfectly drained, some very small included areas are poorly drained, whereas others are well drained.*

The Hamblen series at that time belonged to the alluvial soil group of Azonal soils as defined in the 1938 Yearbook of Agriculture (Baldwin et al., 1938). Today, the Hamblen series classifies as fine-loamy, siliceous, semiactive, thermic, Fluvaquentic Eutrudepts (Soil Survey Staff, 2003). At some point, the classification was updated from the 1938 system to "Soil Classification—A Comprehensive System: Seventh Approximation" (Soil Survey Staff, 1960) and/or Soil Taxonomy (Soil Survey Staff, 1975).

I presume this “re-correlation" was done so existing information could be incorporated into the Soil Interpretation Record (Form 5) in the early 1970s or into the State Soil Survey Database (SSSD) system in 1987. At the time of this soil survey, most soil profile descriptions only extended to about 20 to 40 inches. This appears to be an earlier case where the needs of a database encouraged the population of data that exceeded the availability of supporting information.

To meet the requirements of the National Bulletin, enough tabular information must be given about the “very small included areas [which] are poorly drained" so these areas will show up as hydric inclusions within the

R.L. Livingston, Soil Scientist, USDA-NRCS,
201 Main Street, Suite 103, Dayton, TN 37321
(email: rick.livingston@tn.usda.gov).
Published in Soil Surv. Horiz. 47:43-46 (2006).

“Hamblen loam" map unit. It took a leap of faith, a bunch of prognostication, a fair amount of interpretation, the application of paradigm principles and tacit knowledge, and finally, some expert estimates for me to populate the required information. And then, what do we do about the part "whereas others are well drained "?

We have been estimating soil properties since the inception of the soil survey program, and making reliable estimates will be a necessary part of soil science in the future. Reliable and reasonably accurate estimates of many physical and chemical properties are very much part of soil science and are well understood in many cases. Sometimes, our initial estimated data are used in algorithms to populate other data elements, which equates to sort of a second generation estimate. Then the second-generation estimate is combined with other first and second-generation estimates for interpretations.

What really is the "truthiness" of the interpretation? **Have we crossed the science-pseudoscience boundary?** Is this scientific misconduct? Can we explain this interpretation as some kind of paradigm or tacit knowledge transfer?

Documentation seems to be the answer to many of these conundrums. In earlier times, it seems we were quite candid, at least within our agency, about the origin of data and interpretation on the outdated Soil Form 5 “Soil Interpretation Record." Phrases like:

> *A. RATINGS BASED ON TEST DATA ON SAMPLES FROM SIX PEDONS. LINE 1 ALSO INCLUDES SC, SM-SC*
> *B. RATINGS BASED ON NATIONAL SOILS HANDBOOK.*
> *C. WOODLAND RATINGS BASED ON NATIONAL FORESTRY MANUAL*
> *D. WILDLIFE RATINGS BASED ON SOILS MEMO-74, JAN. 1972*

(from footnotes on the Waynesboro Series Soil Interpretation Record. 0811990) were routinely used to help document the data and interpretations.

When we discuss populating data by estimates, algorithms, tacit knowledge, or other methods used in lieu of measured data, the discussion usually turns to a line of reasoning that goes something like “if someone has to make an estimate, we (Soil Scientists in the NRCS) are better qualified than anyone else" or “our data is the best out there." Sometimes the phrase "because nobody else has any data" is added. I cannot argue with any of those points. The questions that usually come to my mind are "Why are we doing this?" and '"Who needs the information?" Since there are better sources of this information or we do not have measured data, are we really required to put something in that field?

Interestingly, while hunting for literature, I found a reference pertaining to soil data and interpretations that was published in 1996. Brasher and Benham (1996) outlined a system of assessing and documenting soil survey information. This chapter originated from the proceedings of a symposium at the Soil Science Society of America meetings in November 1993. During this time, the NRCS was about to venture into many unknown areas of the information age and databases.

So, what can we do differently? Here are a few things that possibly could help with some of my concerns. Many are in the form of questions because I am not sure if implementation is even possible.

1. Could we have status boxes to indicate data quality? We have a status block next to some data elements to indicate if data is C (calculated), or M (manual override). Why could we not have a status box next to others that indicates status such as D (data from laboratory), E (expert estimate), A (estimated data by algorithm), S (estimated from data from similar benchmark soils), W (wild guess in lieu of laboratory data)? This emulates the "Quality Class Code" of Brasher and Benham (1996). Other important things could be placed in text notes, but some indication of the data quality would be included where the user could see it easily. This kind of ranking could also help us identify where data are needed and where to concentrate sampling priorities. This could perhaps be used in generating part of the metadata for tabular information. Some data users are not concerned with the quality of our data, but others would like to know.

2. Could we define the RV (representative value), high, and low values where possible? This “RV" could be the mean, median, or mode of a group of data or just an expert estimate. Which of these is more "representative" when sufficient data are available? That brings up the discussion of the range. Is the range defined by ±1 standard deviation of the mean, ±15% of the representative value (RV), or possibly an absolute value of ±10% of the representative value (RV)? I realize that measured data are not available for many situations and probably will not be available for some time. If the concepts of representative value (RV) and its range are not defined, and data are available, different data summary methods and statistical techniques could lead to different values for the representative value and its range for the same dataset.

3. Should the RV or a range be used in interpretations? It certainly appears the range would be a better estimate of performance of a map unit in our interpretations. Maybe only ranges should be populated in lieu of measured data and RV used when at least

some data are available to substantiate the claim. On a similar note, should ranges and RVs be calculated to one-tenth of one percent, when the numbers are derived from algorithms based on expert estimates?

4. In my opinion, documentation will cure many ills. Documentation is the difference between science and pseudoscience in many cases. Even if a number or range is not statistically significant, many times it is still very useful and meaningful. A professor once told me that we get much more information from trends and relationships in a data set than we ever get from a single number, no matter how accurate the measurement. It is our duty to supply information to our users, but it is also our responsibility to give the users some idea how the data were obtained.

5. Much effort and resources have been spent populating data voids by various methods. Why do we not apply an equal amount of effort and resources for sampling and acquiring measured data? With careful planning and forethought, we can only increase our knowledge base and ultimately our ability to provide accurate estimates and interpretations.

Maybe this article is my way of being a conscientious objector to some parts of my job? Anyway, I think **it helps all of us to ask "What are we doing"? and "Why we are doing it?" every now and then.**

Upon completing a senior level soils class, a professor gave each member of the class a "diploma" that deemed each class member a "Prognosticating Pedologist" with "all of the responsibilities and accolades that the title holds." (A muddy thumbprint of the professor was the "official seal"!) Presently, I may have reached the point in my career where I can fully utilize this title.

References

Baldwin, M., Kellogg, C.E., and James Thorpe. 1938. Soil classification. p. 979-1001. *In* Soils and men. USDA Yearbook of Agriculture for 1938. U.S. Gov. Print. Office, Washington. DC.

Brasher. B.R., and E. Benham. 1996. Quality assessment of soil survey information. p. 21-31. *In* W.D Nettleton et al. (ed.) Data reliability and risk assessment in soil interpretations. SSSA Spec. Publ. 47. SSSA, Madison. WI.

Elder, J.A., E.K. Yates, and I.B. Epley. 1958. Soil Survey of Marion County, Tennessee. U.S. Dep. Agric., Soil Conserv. Serv., Ser. 1950 no. 2. U.S. Gov. Print. Office, Washington, DC.

Hudson, B.D. 1992. The soil survey as paradigm-based science. Soil Sci. Soc. Am. J. 56:836-841

Kellogg, C.E. 1961. Soil interpretation in soil survey. USDA-Soil Conserv. Serv., Washington, DC.

Nettleton, W.D., A.G. Hornsby, R.B. Brown, and T.L. Coleman (ed.) Data reliability and risk assessment in soil interpretations. SSSA Spec. Publ. 47. SSSA, Madison, WI.

Soil Survey Staff. 1960. Soil classification—A comprehensive system: 7th Approximation. USDA-SCS. U.S. Gov. Print. Office, Washington, DC.

Soil Survey Staff. 1975. Soil taxonomy: A basic system of soil classification for making and interpreting soil surveys. USDA Agric. Handb. 436. U.S. Gov. Print. Office, Washington, DC.

Soil Survey Staff. 2003. Keys to soil taxonomy. 9th ed. USDA-NRCS, Washington, DC.

USDA-NRCS. 2005.National Bulletin 430-5-7. Available at http://policy.nrcs.usda.gov/scripts/lpsiis.dll/NB/NB_430_5_7.htm (verified 1 Aug. 2006).