DATA QUALITY - MORE IMPORTANT THAN EVER IN THE INTERNET AGE

David R. Lide

Editor-in-Chief, CRC Handbook of Chemistry and Physics
E-Mail: drlide@post.harvard.edu

The CODATA Constitution begins with a list of goals in its mission, and first on the list is "The promotion of the evaluation and, in general, the quality control of data ...." Data quality was foremost in the minds of the early leaders of CODATA. The first President, Frederick Rossini, had developed a program for systematic evaluation of thermodynamic data and had established a continuing data center with a large staff of scientists for this purpose. Lewis Branscomb, who was on the CODATA Bureau in the early 1970s, wrote several articles such as "Support for Reviews and Data Evaluation" (*Science* **187**, 603, 1975) and "The Misinformation Explosion: Is the Literature Worth Reviewing?" (*Scientific Research* **3**, 49, 1968). I was invited by the Editor of *Science* to contribute an article "Critical Data for Critical Needs," published in 1981, which gave many examples of unreliable data appearing in the primary literature. Other active participants in CODATA, such as Yeram Touloukian, spread the doctrine of critical evaluation with an almost religious fervor. The programs of the early CODATA Conferences were dominated by papers from data evaluation centers that produced reliable, critically evaluated data sets in various scientific fields.

I believe that the efforts of CODATA during its first quarter-century did help to increase awareness of the fact that much of the data in the primary literature was poorly documented, subject to unknown systematic errors, and often inconsistent with measurements from other laboratories - and consequently that data from critically evaluated sources should be used whenever available. Early CODATA projects such as Fundamental Constants and Key Values for Thermodynamics produced definitive data sets that have been adopted throughout the scientific community. Publications such as the series of *CODATA Directories of Data Sources in Science and Technology* pointed the way to reliable sources of high quality data in a wide range of disciplines.

We now live in an age where finding data is far easier than it was when CODATA started. The sheer volume of scientific data available on the Internet is astounding. Furthermore, modern search engines can search millions of sites and retrieve a list of potentially relevant ones in a few milliseconds. This is particularly valuable to a person seeking data in an unfamiliar field; a search that previously might require many phone calls and visits to remote libraries can now be done in a few minutes. The downside, however, is exactly this sheer volume of data returned by most searches. The ranking algorithms used by commercial search engines hardly take into account the scientific quality of the data, so that the user is on his own in selecting which of the results is best.

Let us start with a rather trivial search, the atomic weight of lead, where Google finds over one million hits. The first three give the current IUPAC recommended value, but only the third on the list includes the uncertainty. Curiously enough, the fourth hit in the ranking is a 1915 paper, and another hit in the first 10 is a 1930 paper. There is a hit from Wikipedia with the correct value, but it is called "atomic mass" and given with units g/mol, which is incorrect. The message is clear: an uninitiated user must take pot luck or ask for help in selecting which result to use.

My next example comes from an actual query posed to me recently about the melting point of calcium oxide. The questioner asked why the value 2899°C that I recommended was so different from the values around 2600°C that he found on the Internet (the implication being that "if it is on the Internet, it must be good"). A Google search did show that most of the values in the top part of the list (1.1 million long) clustered around 2600°C. The frequent appearance of the value 2572°C suggests that many sites simply copied the number from other sites or books. The most reliable sources, which gave values near 2900°C, were presumably far down on the list if they showed up at all.

Almost any Google search for a property of a relatively common chemical compound or material will yield from 100,000 to several million hits. Granted that 99+ % of these will be irrelevant, it would be rare to find agreement among the sources that are potentially relevant. A search for the density of water gives 52 million hits, topped by a small chemical supply house that gives one value without specifying the temperature. Wikipedia simply rounds the value to 1000 kg/m^3, again with no indication of temperature. The definitive data from IUPAC (or other databases

that quote the IUPAC data) are buried somewhere in the crowd; presumably, IUPAC is not able to budget for a high position on the list.

The hits in this type of search cover a bewildering range of sources. Many are web sites created by students; some are maintained by professors to give background information on problem sets to their classes. In either case it is doubtful that much research has gone into the selection of the data. Many come from companies that sell the material; some of these have been carefully done, others not. Some are current, others very old.

There are, of course, many good sources of evaluated data on the web. Some of these, such as the *Landolt-Börnstein Tables*, the CHEMnetBase site (which carries the *CRC Handbook of Chemistry and Physics* and the *Chapman & Hall Combined Chemical Dictionary*) and the Cambridge Crystallographic Data Centre, require a subscription. Others, however, are free; these include the *NIST Chemistry Webbook* and *Physical Reference Data* sites, the *Protein Databank*, and sites maintained by IUPAC, IAU, and other ICSU Unions. The problem is that these rarely appear near the top of a Google search result.

So what can CODATA do to help data users separate the wheat from the chaff? During my years working with CODATA there were occasional discussions of the feasibility of assigning a "CODATA seal of approval" to data sources (then mostly in print form) that passed a test of quality. For technical and political reasons, this idea was never pursued, but I believe it should be revisited now that we live in a new information era. It would not be simple to implement, but with a little imagination might be done. Users could then be encouraged to add that "seal" to their search terms. Another approach is to revive the *CODATA Directory of Data Sources in Science and Technology* (and its electronic offspring, the *CODATA Referral Database*) in the form of a list of selected web sites that provide data of high quality. This listing could be carried on the CODATA home site with appropriate links. Steps of this type are very much needed to restore data quality as a top priority in the CODATA mission.