PREPRINT
Published in: *Issues in Science and Technology Librarianship*, Viewpoints, Fall 2012
http://www.istl.org/12-fall/viewpoint.html

The Unbearable Lightness of Data: Touloukian's Quest and Why It Still Matters

David Flaxbart
Chemistry Librarian
University of Texas at Austin
flaxbart@austin.utexas.edu

There has been much discussion recently about what is broadly called "e-science," which focuses on the preservation of and access to the mass of underlying data that is generated by scientists in the course of their work, but which is typically inaccessible after the research summaries have been published in the literature. The recent move by the NSF to require grant applicants to specify a "data management plan" has lent some urgency to a situation that has been on the back burner for a long time. The question of what if any role libraries should play in this process seems to be the main concern for librarians – but that is a question for another column (such as Haas and Murphy, 2009).

I recently had occasion to update some information on physical property resources for my web pages, and I had another look at an artifact that only librarians of a certain age might recall: an index with the mellifluous title *Thermophysical Properties Research Literature Retrieval Guide*, and its compiler, Y. S. Touloukian.

Now, the if-it-ain't-on-the-internet-it-don't-exist readers out there are probably rolling their eyes and furiously clicking their back buttons. Everyone else, bear with me.

Y.S. Touloukian (1920-1981) was a physicist at Purdue University who in 1957 founded the Thermophysical Properties Research Center (TPRC, now known as CINDAS). He spent his career tirelessly advocating for improving the quality of scientific data collection and evaluation, and, equally importantly, improving access to that data after the fact. His *magnum opus* was a fifteen-volume blue-covered data compilation titled *Thermophysical Properties of Matter* (IFI/Plenum, 1970-79) that still sits, most likely festooned with dust bunnies, on many a library shelf. More recently, CINDAS, now a private company, has offered a subscription database derived from the book set.

Others had observed that traditional bibliographic indexes such as *Chemical Abstracts* did a relatively poor job of helping the researcher locate actual physical property data buried within the primary literature. (O'Connor, 1977) And they were right: anyone who tries to use an *index* to find a *specific piece* of numeric data is likely to miss much. Abstracts tend to be of little help in determining how much and what kind of data may or may not be present in an article, and the human-supplied metadata is also lacking in granular detail. And while these observations date from the era before there was ready access to online searching, the situation is not that much different now that indexes are all-digital, because the traditional literature indexes themselves, and the methods used to compile them, remain largely unchanged. (1) The problem is organic and fairly obvious: It is often not fruitful to use text to search for numbers.

The *Retrieval Guide* was Touloukian's attempt to remedy this problem by creating an evaluated bibliography of publications reporting physical property data for various materials. The source documents were examined for their data content and coded for parameters such as physical state and temperature range, and meticulously indexed by substance. (Ironically, the sources were all identified in the first place by – you guessed it – searching in bibliographic indexes, because there was truly no alternative.)

A while back I pulled our copy of the *Retrieval Guide* off the shelf and took a fresh look at it. It has an immediate ick-factor for modern users: it's basically a typeset printout of a computer database, it's fat, and it has a highly arcane organization that takes a bit of mental effort to comprehend. A user guide on the end sheets aims to help if one takes the time to read it. Looking anything up requires several steps, consulting in turn three separate, minutely printed indexes, to arrive eventually at a literature reference that, as likely as not, is to something like "FIZ TVERD TELA". After which you're on your own finding a copy of it.

Using it always reminds me of the famous scene in the Marx Brothers' *A Day at the Races*, when Chico sells Groucho an extensive set of racing reference books, none useful without the other, all of them leading to a tip that comes too late for the race. The *Research Guide* is not a tool for the faint of heart or the short of time.

We all know that users take shortcuts whenever possible, and that for every researcher who knows about and uses an appropriate resource there are probably five others who stick to Google, whether out of ignorance or personal preference. That isn't going to change. The "good enough" ethos of information-seeking has won out, but it guarantees that pertinent information will remain unfound. I'm not saying that we should all dust off our copies of the *Retrieval Guide* and start using a magnifying glass instead of a computer when we need to look for the specific heat of Rhenium hexafluoride at low temperatures. But a promising lead to that very data point is right there in its bibliography section, and who knows if it would ever turn up in a tool we're more comfortable with. (2) And heaven help the Googler!

While I was studying the *Retrieval Guide* I came across an article Touloukian wrote shortly before his death. While it's mainly a self-congratulatory overview of his organization, he made some interesting statements about data discovery and access that should resonate in today's e-science debate.

> While the nation's scientific and technical community starves from a lack of critically *evaluated* information, it is being smothered by an overwhelming document birth rate with the associated emission of polluted information referred to as "original data." The nation spends large sums of money on its research and development only to waste it by ignoring its results. *Information discovery* is indeed futile in the absence of an adequate means of *information recovery.* Should there be continued research, then the evaluation and proper dissemination of the results of this research to the end-user can hardly be questioned. (Touloukian, 1981)

In other words, all the data collection in the world is fairly pointless if that data cannot later be found and used by other researchers. And Touloukian was talking primarily about *published* data; extending the scope to unpublished data increases the problem by many orders of magnitude, which is the "pollution" he refers to. What is the true value of petabytes of unpublished, unevaluated raw data? Will storing it on centrally managed servers really help move the scientific process forward? Or will the sheer inchoate mass of it merely overwhelm those few gems that might exist therein? What tools will

be used to mine it, and how effective will they be? Might it just be easier to recreate data with new measurements than to find, convert, and analyze old data that may well turn out to be useless anyway? What would Touloukian have thought about the idea of open e-science?

As we ponder these questions, his bibliography sits there on our shelves, daring us to look beyond our computer screens for that elusive bit of *published* data that may not be discovered any other way.

Notes

1. It should be understood, however, that not all indexes are created equal. A search for chemical data in CAS' SciFinder has a much greater likelihood of success than the same search in, say, Web of Science or Engineering Index. This is the result of the variable levels and quality of controlled vocabulary and compound identification from tool to tool. Success or failure also depends very much on the skills of the individual searcher. Just because you have a Steinway doesn't make you Rachmaninoff.
2. Happily, a 1963 proceedings reference from Touloukian's Guide *does* show up in a SciFinder topical search on "heat capacity of rhenium fluoride." But you wouldn't have found it if you used the term "specific heat" – CAS used "heat capacity" instead. Nor is the value found in Reaxys, because Gmelin did not cover Re compounds extensively in that time period.

References

Haas, Jennifer; Murphy, Sharon. 2009. "E-Science and libraries: finding the right path." ISTL, Spring. http://www.istl.org/09-spring/viewpoint1.html

O'Connor, John. 1977. "Data retrieval by text searching." *Journal of Chemical Information and Computer Sciences*, 17 (3), 181-186.

Touloukian, Y.S. 1981. "Twenty-five years of pioneering accomplishments by CINDAS--a retrospective review." *Int. J. Thermophysics* 2, 205-222.

Touloukian, Y.S. 1982. *Thermophysical Properties Research Literature Retrieval Guide.* 3rd ed. (New York: IFI/Plenum).