Scientific publications and data evaluation in the digital age (a perspective of a thermodynamics researcher)

Vladimir Diky[1][2][3]

Abstract

The state of scientific publications, problems, possible solutions, and underutilized opportunities are discussed on the basis of author's experience as a reader, author, reviewer, and editor. The author feels that significant improvement can be made, which will increase the efficiency of communication and quality of information. The focused area is thermophysical properties related to chemical engineering, but the concerns and conclusions can be applied to a wider scope.

Keywords

Scientific publications; Machine-readable formats; Data; Evaluation; Chemical engineering; Thermodynamics; Thermophysics

Article content



## 1. Introduction

Since the invention of the printing press by Johannes Gutenberg around 1436 (or a possible earlier invention), another most important change in dissemination of scientific information was caused by computerization. However, the nature of scientific communication has remained basically unchanged even since the invention of writing millennia ago: an image of symbols and illustrations for perception by humans. That means tremendous opportunities to increase the efficiency of scientific communication still exist and need to be utilized. The discussion in this article concerns the author's area of expertise, chemical engineering and related subjects, but

[1] Material Measurement Laboratory, NIST, 325 Broadway, Boulder CO 80305. E-mail diky@nist.gov

[2] Any mention of commercial products within this article is for information only; it does not imply recommendation or endorsement by NIST. These opinions, recommendations, findings, and conclusions do not necessarily reflect the views or policies of NIST or the United States Government.

many issues have much wider relevance. I would like to consider six aspects: content of reports, form of presentation, procedures of publication, availability of information, reliability of data, and modeling. Those aspects are closely connected, and cross-references will appear. I will analyze the situation, reveal problems, and discuss possible solutions and perspectives.

## 2. Content of publications

Most publications related to thermodynamic and thermophysical properties of substances and materials result from research funded by public money. Funding agencies usually sponsor research programs associated with practical needs. Measurements, modeling, and prediction of physical properties are then by-products of the claimed goals. Though the importance of thermophysical data in general is broadly discussed [1], some of the intended beneficiaries, such as representatives of industry, cannot directly disclose their needs because of the proprietary nature of that information. Resources revealing and broadcasting the needs and listing outstanding problems may be useful. Presently, the TRC (Thermodynamics Research Center) collection of unresolved inconsistencies contains 365 cases. Its content was published in 2019 [2], but it is growing faster than the issues are being resolved, and a Web-based interface would provide a better service.

It is not strictly defined what should be communicated after publicly funded research is completed. Funding agencies are generally satisfied by the facts of publication with an appropriate acknowledgment. An attempt to explicitly say what should be published was made in the IUPAC Good Reporting Practice project, where representatives of readers/users, as well as experts in material science expressed their needs and visions [3]. Essentially, their statement is that all knowledge and data acquired in research should be published, well defined, presented in a numerical form, traceable to the origin, and convenient to consume. The final report [3] provides details for specific areas of material science, while its general principles can be applied to any scientific publication. Unfortunately, many publications nowadays are still not compliant with those principles. They are declarations of the research done rather than reports of the results. In my estimation, about 30 % of the scientific results in the area may be lost due to not reporting or poor reporting. Another 30 % of the research funds may be wasted due to production of erroneous or misleading information as discussed in the Reliability section. Possible reasons of poor reporting may be lack of motivation of the participants of the process (authors, reviewers, publishers) and not involving the interested parties. For example, additional efforts to provide better reports would not increase researcher's chances to obtain another grant or publisher's profit. Those issues will be discussed in the Procedures section. I would say here that the existing metrics of researchers' success may not sufficiently correlate with their actual impact. The existing metrics are based on the numbers of publications and citations. Unfortunately, these may not reflect the quality and reliability of the information, even the amount of scientific production. A frequent practice is repeating, basically, the same information in multiple publications, and splitting measured data into several fragments and publishing each in a separate article. Citations frequently reflect popular topics of research rather than the quality and actual impact of the cited publications. A more adequate metric based on the validation of the results would require expertise, labor, and additional time to accumulate the basis for evaluation. However, practitioners commonly associate the reliability of publications with the laboratories of their origin.

An improvement in the content of scientific reports can be achieved through acceptance of Good Reporting Practice [3]. The most effect can be achieved if funding agencies accept it and modify their procedures to support it, but someone should reach and motivate them.

3. Form of presentation

First, let's consider perceiving of scientific publications by a human reader. From my perspective, attributes of a good publication are the following. A good summary allows the reader to quickly evaluate the relevance of that publication for their needs. A clear structure allows one to easily and completely identify all components of the report such as experimental results, data collected for review, mathematical models, model parameters, and derived data. References and cross-references would facilitate navigation through the article, e.g., references to derivation of equations, references from figures to tables with corresponding numerical data, etc. Examples of good article structure can be found in J. Phys. Chem. Ref. Data (e.g., [4]). It has become a good tradition to have an "Experimental" section where the studied substances and their actual samples are described, especially convenient in tabular form, as well as the experimental methods, but the results are usually mixed with modeling, comparisons, and discussion in a messy section "Results and discussion." That is similar to "spaghetti code," a practice discouraged in computer programming. It is frequently needed to scan the whole article to discover or understand a single piece of data, table, or figure. Such a way of presentation causes enormous losses of time for extracting the information from a scientific publication. Again, there is, unfortunately, little motivation for the authors to produce high-quality publications other than their own esthetical feelings. Possible solutions are related to the procedures of publications and will be discussed in the corresponding section.

Another highly underutilized opportunity provided by computerization is machine-readable scientific publications. First of all, that concerns numeric data and mathematical equations. The main use of numeric data nowadays is input for processing by computers. That means they should be captured as structured symbolic information. Modern publishing technologies normally produce scientific articles in the form of computer files with rendered symbols that should obviate typing (or scanning, doing OCR – optical character recognition, and fixing OCR errors), but transferring data from publications intended to be read by humans is still a labor-intensive, stressful, and error-prone effort, especially when transferring data from compact "good-looking" tables, highly fragmented and containing common values of state variables in separate places (Figure 1). Publishing in formats directly readable by computer software would both save time for a more productive activity and achieve another goal: assuring that the data are well defined, which means defined completely and unambiguously. It is very difficult for a human to not miss anything, and only parsing by computer software can assure the correctness of reporting as defined in [3]. It is amazing that authors spend, basically, the same amount of time constructing such tables from plain data files they initially possess, when such files should allow more easily comprehensible information than perhaps more attractive compacted tables. In addition, typographical errors are frequently made by the authors composing “good-looking” tables.

Mathematical equations usually need to be reimplemented in computer programming languages. That is also a major effort, more than copying data from the source document. Again, errors are easily made producing traditional visual images of equations, and re-derivation is needed to prove the correctness. Cases are known when typographical errors were revealed after publication of equations, and implementation of those equations as published would lead to incorrect results. At the same time, the authors usually possess the data and equations in computer-readable form, such as equations implemented in programming languages, and publishing them in the hard-to-consume traditional way typically requires additional efforts. A good example of reporting equations in machine-readable format is [5] where both C++ code and Python script are given in attachment. What prevents the authors publishing data and equations in machine-readable form, especially as journals provide effectively unlimited space for supplements? I can guess, lack of motivation as discussed in the previous section. However, publishing data in a machine-readable form may be more difficult than publishing equations, because of the absence of commonly accepted formats and convenient data capture tools.

There are examples of successful electronic data communication of scientific information. The Cambridge Structural Database (CSD) [6] accumulates crystal structure information in a machine-readable format. It is a requirement that crystal structures communicated to scientific journals are submitted in that database format. The Protein Data Bank [7] is another example. Similar attempts have been made in material science, but they have not been successful. One of the reasons is the complexity and diversity of the data, in contrast to uniform data in the mentioned cases. Instead of basically one property, material science deals with hundreds of properties reported in different variations as functions of multiple variables. Consider one of the simplest properties, density. In addition to various conditions and phase states, it may be reported as specific or molar density, specific or molar volume, compressibility factor, as a direct value, difference, ratio, or relative difference from a reference state, and in different units. That makes initial data capture and development of convenient tools and procedures for that an extraordinarily difficult task not yet fully addressed. Quite powerful data capture software (Guided Data Capture or GDC) [8] is used at TRC (NIST), but it requires installation on a user computer, certain knowledge and efforts to do data capture, and still undergoes constant improvement to be more flexible and convenient. A pioneering attempt of electronic publication of thermophysical data was ELDATA journal [9]. It consisted of a paper version and electronic files with numerical data. Moderate amounts of data were completely published in both forms, while large data sets were complete in the electronic form only, and only part of the data were in the paper version. I do not exactly know why the journal was discontinued after five years of existence. Of course, the dawn of the digital age 30 years ago when ELDATA was founded did not provide the capabilities we have nowadays, such as Internet, high-performance computers, and powerful software, and all that reduced the chances of ELDATA to survive. Probably, the absence of sponsors’ support doomed that project, which might not have commercial success. Probably, the authors were motivated to contribute by the personality of Henry Kehiaian (founding Editor of ELDATA) [10] rather than other reasons. Unfortunately, the surviving electronic files are impossible to interpret, even having the ELDATA software. I was able to recover part of them by contacting the authors of the corresponding ELDATA publications.

Another attempt started at TRC/NIST in 2002 [11]. The idea was to establish the infrastructure and procedures similar to CSD [7]. An additional motivation was revealing of a significant number of errors in articles during data capture. That triggered a number of corrigenda published in 2002 published by the authors after TRC comments. To make necessary corrections prior to publication, an agreement was established between TRC/NIST and 5 journals from 3 publishers (ACS, Elsevier, and Springer) to access the data and provide feedback during the review stage of manuscript publication. The authors were requested to install GDC software, capture the measured data, and submit data files together with manuscripts. The involvement of the authors was terminated soon by TRC. TRC's justification of that termination was too many data capture errors in the data files made by the authors. I believe, the main problem was the burden of installing GDC software and handling data files. In my opinion, the only convenient way of data capture and submission by the authors would be a Web-based interface, but development of such an interface is a challenge even nowadays, and the short life of Web development platforms causes high maintenance costs. So, all data capture from the cooperating journals has been done at TRC since then, but it does not seem to be a sustainable solution. Development of a self-service Web interface for data capture is critically important. I believe, scientific publications will become machine-readable earlier or later saving processing time, reducing the number of mistakes, increasing productivity of researchers, and accelerating overall progress.

4. Procedures of publication

As said above, a significant part, if not majority, of research is done with public money distributed by funding agencies, and communication of research results, which is a requirement for publicly funded research, is done in the form approved by the funding agencies – articles in scientific journals. Publication in traditional journals is usually free for the authors. That scheme removes publication expenses from research budgets. Those expenses are paid by readers and subscribers. Publishers are supposed to assure quality and sustainability of publications by arranging review and providing storage and dissemination. Scientific editors are appointed by publishers on a compensation or volunteer basis, and reviewing is a volunteering activity. There are certain benefits and deficiencies in this scheme. Required research budgets can be seemingly smaller because of not including publication expenses, but a significant part of those expenses is still paid by researchers indirectly, through overhead costs, which include subscription prices for researchers' organizations. Publication charges can be paid by the authors, and articles can then become free to the readers. Paying for open-access publications directly from research budgets reduces the funds available for other needs, which may be critical in some cases. The weakest part of the existing procedure, in my opinion, is reviewing. Reviewers have the most freedom among the participants of the process and the least responsibility. They have little motivation: their job as reviewers is not paid, and reviewing may be a hard and time-consuming job, especially reviewing of ill-formed manuscripts. As a result, reviewing is frequently perfunctory and inadequate. Reviewers may lack the necessary expertise, and their comments may be inappropriate. Ultimately, the reliability of the information from scientific publications may become equal to that of messages in social networks, and any piece of information may just as likely be correct or wrong. Editors are in charge of the final decisions, but they are usually overwhelmed with their workload and unable to thoroughly assess reviewers' performance.

The author of this article is familiar with nearly all roles in the process of publication: reader, reviewer, editor, and author. The first two roles are similar; I would say a reader is the most motivated reviewer, but without any power to influence the publication, which has already happened. Both as a reviewer and as a reader, I face basically the same challenges. Poor structure of documents, incomplete reporting, and lack of clarity make reading, discovery, and interpretation of the interesting parts of the document difficult and time-consuming. If that happens to a manuscript, it may indicate an insufficient qualification of the authors to do that research. If it happens to a published article, it means a failed review. Finding reviewers, especially good reviewers, is a challenge for editors. Once, handling a submission far from my area of expertise, I contacted 10 professionals until I received three reviews, of which only two provided useful information. In another case, a colleague, an editor of another journal, complained he could not find a reviewer for a manuscript from a well-known and highly ranked laboratory. The storyteller offered his service and enjoyed commenting that manuscript. Many other cases fall in one of two categories: positive reviews not addressing serious or even critical deficiencies on one hand; or reviewers' unreasonable demands for unnecessary revisions, on the other hand. While poor and misleading articles are frequently published, an obstacle for publication of good articles may be formatting and style demands imposed by certain publishers and reviewers. Satisfying those demands takes much time and does not increase the quality of the information or the ease of its consumption.

According to my observations, there are different common patterns of editors' actions. Editors, probably overwhelmed by the volume of submissions, rely entirely on reviewers and have no chance to read manuscripts. I am in a better situation handling a handful of submissions. I trust only those reviewers, whose qualification has been proven. Quite frequently, they are top experts in their areas, and I can manage to contact them and not abuse their attention only because I invite them infrequently. In such cases, comments from one reviewer suffices. Respecting the reviewers' time, I try to reveal and, if possible, fix any problems before appointing them. A paradox is that the worst manuscripts consume most of my time as a reviewer or an editor, unless it is immediately obvious that the manuscript should be rejected. If an editor receives 2-3 poor manuscripts a month, handling them with the needed scrutiny can consume all editor's time. In reality, many editors have a much higher workload. As mentioned, finding good reviewers is not easy, and potential good reviewers are usually overwhelmed with their duties. Shortened reviewing times required by many publishers may repel reviewers. Accepting services of whoever accepts an invitation may result in poor and misleading publications.

Does the existing state need improvements, and how can it be improved? If the answer is no, we may need to treat scientific publication in a way we treat news from the Internet or social networks: thoroughly validate any piece of information before accepting it and reveal reliable and wrong reporters. A target audience such as industrial chemical engineers could probably be good reviewers, but I did not see their enthusiasm when I raised that question at one of their conferences. Young researchers can make good reviewers: they may be less experienced, but they are often diligent and enthusiastic. An underutilized resource may be advisory boards of journals; their active role in the review process can be made an explicit part of their duties. Enthusiastic retired experts may be a resource. Another option could be establishing an institute of paid reviewers and keeping a high level of requirements to them. It is doubtful if publishers can afford that, but funding agencies can.

5. Availability of information

The existing schema of dissemination of scientific information seems to be rather traditional than optimal. The idea is, scientific results are publicized to promote progress worldwide, but the actual availability for each reader is determined by the subscription capabilities. In combination with the inability of that schema to assure the quality and reliability of publications, it may indicate the need to change it and directly invest part of national research funds in evaluation and dissemination of the results. Practices of making scientific publications freely available exist, for example, in government organizations of USA and certain Japanese and Canadian journals. A "curated research-sharing platform" [12] can be mentioned.

Prior to accessing scientific information of interest, it is necessary to discover it. Traditionally, that discovery is provided by paid abstracting services such as Chemical Abstracts (which acquired historic Zentralblatt) or Beilstein (now Reaxys). They provide tremendous services, but physical properties of substances seem to be outside of their main scope. Property data search in one popular service is apparently based on analysis of natural language similar to Web searches rather than structured information, which used to appear in the paper version in the past. Search is mostly based on natural language phrases, and the results are highly sensitive to the phrases used. None of them are able to assure all needed responses, and the vast majority of the results returned are irrelevant, which requires an enormous human effort to filter them. The situation is even worse when one needs to find data for mixtures, as compared to pure compounds. Structured property data searches are provided by other services, but their scope and coverage are limited. A lot of property data are included in reference books and reviews. A problem for readers is that many of them highly duplicate each other, but none of them covers the content of all others, so data search that way is too labor intensive. In addition, some of them have poor content indexing or are not traceable to the original sources. Similar observations are relevant to databases and services, as well. Those observations are supported by the state of literature reviews from scientific publications: different reviews of the same topics frequently have significantly different coverage. The problem is prominent, and solutions are needed. One may hope that artificial intelligence can provide better content analysis, but its present state is far from solving such tasks. A possible solution could be self-indexing by the authors, and the motivation could be expected increase of citation due to self-indexing, but such a service would need some review of the entered information and, of course, investment. Publishing and consolidating indexing information may be beneficial. The TRC index database (not publicly available) contains over 130 thousand records indicating data sets, most of which are not yet captured to the SOURCE database [13].

As part of the agreement between TRC/NIST and the cooperating journals and publishers, the captured data could be posted in a machine-readable format for free access at the TRC Web site. The format selected was ThermoML [14], and that was probably the only use of that XML-based format outside TRC. However, the expenses to produce those free data proved to be enormous, because of the reasons discussed above: difficulties to discover, interpret, and extract data from publications. The amount of work to review and capture the experimental data from those journals providing, probably, 40 % of the published property data for molecular compounds and mixtures, consumed the in-house capabilities of TRC. To capture basically all published data and collect the data published in the past, it would be necessary to have a facility with at least a dozen high-level experts with technical support, and the cost of data capture may be in the range 1 to 10 U.S. dollars per data point. The amount of property data for molecular compounds and their mixtures published yearly may be over 500 thousand values, in my rough estimation. Probably, the consolidated resources involved in data capture worldwide would be able to keep up with that data production, but coordination of their activities seems impossible. A possible solution could be creation of a public data domain supported by a Web collaboration interface and data storage. In that case, the output from projects like the Solubility Data Series SDS [15] could be accumulated, consolidated, refined when needed, and made easily available. In my vision, two features should distinguish such resources from the existing projects such as Webbook [16],TRC ThermoML data collection [17], ILThermo [18], and Gas Hydrates database [19], which are static collections presenting the work done by others: a dynamic nature reflecting the current state of information with corrections when needed and an active role of contributors other than the project keepers. Many community members are interested in free resources. However, free resources always involve cost to somebody. Implementing them as collaboration interfaces could eliminate or spread a significant part of those costs.

6. Reliability of data

While the fact of publication may address the needs of authors, readers need reliable data, at least, those with an adequate estimation of the possible error. Providing uncertainties of reported property data is standard nowadays, but the claimed uncertainties are misleading in most cases. There are two main reasons of that. First, only one contribution to the uncertainty is usually reported rather than the uncertainty budget combining all contributions, and the reported contribution is usually not identified (Manufacturer's specification? Repeatability? Authors' belief?) and is frequently a minor one. A brilliant comment has been made by my colleagues: "As water is the only liquid whose viscosity is known to an uncertainty as low as 0.17 % … all measurements in which the authors quote uncertainties of less than 0.01 % (*e.g.,* 0.003 %!), characteristic of investigators that do not understand how to assess their measurement uncertainty, have been placed in the secondary data set" [20]. Second, mistakes and deviations from experimental protocols frequently happen. As a result, the readers need to make their own uncertainty estimations. The extent of possible errors is revealed as additional information becomes available [2]. A simple case is illustrated in Fig. 2. Known inconsistencies indicate the lower bound of a possible error. It is more difficult to determine the upper bound. Sometimes, the magnitude of error makes the data useless or even harmful. Over 61 thousand of 8.2 million data points are presently identified as flawed in the SOURCE database [13].

An attempt has been made to make independent uncertainty estimations during data capture [8]. Those estimations are based on the nature of the substance, reported purity, property, measurement method, property value, and conditions. For example, inclined piston is more accurate than Knudsen effusion for vapor pressure measurements (in the appropriate pressure range), and adiabatic calorimetry is more accurate than DSC. However, errors as large as 40 % have been discovered in reported adiabatic calorimetry data, but they are associated with particular laboratories and researchers rather than the method itself. A paradoxical situation appears: as we acquire more data and improve our knowledge, the estimated uncertainties increase because we reveal more inconsistencies with other data. Another paradox is, the complete validation of any data is possible only if we don't need it, i.e., if we can derive the value from other available information for the validation purpose. The methods used at TRC for data validation were described in [2]. I should note that some colleagues [21] do not accept the term "validation" considering it to be a claim of the absolute truth and prefer to say "corroboration." The authors of [2] defined validation as a process rather than a result. There are two processes complementing each other: validation and invalidation. Validation shows to what extent we can support (corroborate) the value by other information (which should be done cautiously because any validation is mutual validation). For example, prediction methods, being generally less accurate than measurements, are more robust if correctly used and can give a band, within which true values can be reasonably expected. Invalidation reveals experimental errors and inconsistencies with other data, so that the uncertainty is believed to be not less than that extent. The expected uncertainty caused by equipment, protocols, and practices can be associated with laboratories or researchers and revealed by their previous publications compared to other data. To do that, unambiguous identification of the laboratory and researcher is needed, which is difficult nowadays, but things like mandatory ORCID (Open Researcher and Contributor IDentifier) may help.

Some typical patterns and possible reasons of erroneous reports were discussed in [22]. I do not support optimistic claims made in that article about having solved the problems. The errors may originate from the identity and purity of the studied substances, inappropriate use of the methods and equipment (e.g., measuring viscosity outside of the apparatus operation range), not following the protocols of experiments and data processing (for example, proper DSC calibration, especially for mixtures), unjustified interpretation of the observed phenomena (again, DSC peaks may be an example).

I would distinguish three kinds of errors in scientific publications. One is a data processing or typographical errors, which can easily be revealed and fixed in many cases. The other is experimental or computational errors not revealed by the authors, which may though be persistent in the results from certain laboratories. The last one is dishonest reporting, which may be difficult to prove. A typical case is not reporting known impurities in the samples affecting the results. Other quite frequently encountered situations are when a wrong assumed chemical form of a substance or flaws in data processing are revealed, and the numeric values are not properly corrected after that. If rejection happens, the manuscript is usually submitted to another journal either keeping obviously erroneous data or removing the information indicating the problems. My colleagues and I made such conclusions when we accidentally received such resubmitted rejected manuscripts for review again. Sometimes, the changes of the data in sequential revisions of the same submission do not look credible (e.g., changing property values while keeping exactly same mixture compositions and claiming new measurements) or unexplained data changes happen. The problem of honest reporting has been widely discussed [23], but no solution has been offered. My reviewer's experience has shown that there are few chances to improve the results submitted for publication. I see the goals of a review to assure complete and honest reporting and fixing obvious typographical errors, which would warrant a corrigendum.

TRC/NIST journal cooperation, which had the initial goal of making new experimental data available for the community in a computer-readable form, transformed to data reviewing because of revealing data errors [24]. That cooperation allowed catching and fixing certain errors prior to publication. Because of the increased workload, an agreement was achieved to do data review after acceptance by conventional review, which eliminated approximately 1/3 of manuscripts rejected at that stage. Though an additional review at TRC may cause additional labor and delays to the authors, that happens only if deficiencies are detected in the manuscript overlooked by the previous reviewers. Obviously, the conventional review has failed in such cases. If the authors are motivated to produce a high-quality publication, TRC data review helps them to accomplish that. However, TRC cannot serve all journals, and routine reviewing cannot justify activities at NIST. It is necessary to develop and create resources and services supporting the traditional review. They may include uniform and effective guidelines, checklists, document structure templates, and data capture, visualization, and uncertainty assessment tools. As a pioneering project addressing the needs of one experimental method, combustion calorimetry, the TRC Combustion Calorimetry Tool [25] can be mentioned.

If reporting problems are discovered after publication, the chances of clarification are much smaller. When I discovered deficiencies or ambiguities as a reader and contacted the authors for clarification, my requests have been ignored in most cases, even when supported by the publishing editors. The last resort is submitting comments and initiation of retraction procedures, such as [26]. To assure honest reporting, privacy practices may need to be revised. A possible option is following. Once submitted, a manuscript becomes public, and all modifications, including resubmissions to other journals, if it happens, are traceable.

## 7. Evaluation and modeling

In a narrow sense, I consider data evaluation as producing the best possible estimates of the true property values in the form of single values or equations based on the available data. Data evaluation is generally based on a balance of experimental data and predictions with an increasing role of the latter as the theoretical science develops. The need for and the idea of dynamic data evaluation (DDE) able to update the results as new information arrives was claimed long ago [27] and was implemented in ThermoData Engine (TDE) software [28], which also enforces thermodynamic consistency. Ideas of consistent evaluation of properties represented by separate equations [29] appeared before TDE software. TDE is based on similar ideas, but involves more properties, which can ultimately make evaluations comparable to equations of state. Our experience required us to revise the initial DDE implementation: discarding the previous results and redoing the automated evaluation from scratch every time new data arrive. One reason is that engineering applications need stable models, and only well justified changes should happen. Another reason to abandon that approach was caused by the fact that additional data may sometimes not be better, and they may increase the ambiguity rather than clarify the picture. The algorithm may not perform better than the expert in complex situations, and an expert judgment may be needed, which should not be overridden by software. In addition, significant errors caused by the absence of sufficient am accurate data may happen, and some of those errors are revealed by property correlation for a series of compounds. That requires storing evaluation results for comparisons and correlations.

The things the automations can perform the best are revealing inconsistencies and possibilities to improve the recommended values. The present model of data evaluation at TRC is based on keeping static evaluations and automatically reviewing them as new data arrive or the algorithms are improved. That procedure requires more complex algorithms than automated evaluation alone and solid criteria for making decisions to update the results. Maintaining thermodynamic consistency between different properties makes re-evaluation of a single property impossible in most situations. Because of the complexity of the task, improvement of the automated updating procedures may be a never-ending process.

There are several challenges in data evaluation. As mentioned, it is difficult to judge the reliability of most data sets, even for setting relative weights to different data before model fitting. Another challenge is assessment of model uncertainties. The probabilistic methods consider random errors with known distribution. Bayesian methods require knowing data uncertainties for assessing model uncertainties. In any case, a judgment about the proportion of random and systematic errors should be made. Model bias, extrapolation error, and overtraining are other contributions, which should also be assessed. No general solution exists, to my best knowledge. My colleagues and I employed different methods in our evaluations [30] and assessed the adequacy of our uncertainty estimates by comparing the uncertainties to the scatter of the data allowing some decrease due to decreasing the random error when multiple consistent data sets are averaged. A useful additional option seems to be building and comparing alternative models, e.g. NRTL and UNIQUAC for VLE in mixtures.

Considering an evaluation, in addition to the quantitative questions about uncertainties, qualitative question also appear: Is that evaluation the best possible one and is it acceptable even if not best possible? The answer to the second question is quite obvious if one knows the application and trusts the uncertainty estimates. The first question is more challenging. As discussed above (Availability section), it is difficult to collect and even discover all relevant data. The extent of the relevant data is also difficult to define because any additional data, even for different compounds or mixtures, may contribute to refinement or corroboration through correlations.

New models are being developed for smoothing, extrapolating property data, and consistent fitting of multiple properties. However, the process of implementation of each model is usually labor intensive and may even require their re-derivation. Authors of some models generously made their computer code available, e.g., [5]. Assembling them under a TDE [27] interface as independently developed dynamically linked libraries (dll) called on demand with access to the public-domain data is considered at TRC. A free researcher version of ThermoData Engine with those capabilities may be released.

## 8. Conclusions

As shown above, certain problems exist in production and communication of scientific information. Cheap science is too expensive for society because of a high redundancy needed to get trustworthy results. Development of effective solutions requires identification of the major players and understanding their interests. Those players are funding agencies, researchers themselves, reviewers (mostly from the researchers' pool), publishers, and readers/users. Journal editors act mostly like reviewers, and they may also be readers' advocates. Unfortunately, the interests of different players do not always coincide and may even contradict each other. Possible solutions can be based on motivation of the existing parties, appointing of motivated parties not presently involved on certain stages of the process, or specific actions needing investments. The most radical solutions seem to be revision of the approaches and criteria utilized by funding agencies, accepting machine-readable communication formats, and creating interfaces for self-service and collaboration.

Opportunities to significantly increase the efficiency of scientific publications and research itself have opened due to technical progress. Certain decisions and relatively small investments in the infrastructure can make those opportunities reality.

| $T$/K | $m_A$/(mol·kg$^{-1}$) | $\rho$/(g·cm$^{-3}$) |
|---|---|---|
| water + 0.10 mol·kg$^{-1}$ B | | |
| 303.15 | | ($\rho_o$ = 0.9972) |
| | 0.0570 | 1.0045 |
| | 0.0814 | 1.0076 |
| | 0.0961 | 1.0094 |
| 308.15 | | ($\rho_o$ = 0.9956) |
| | 0.0570 | 1.0029 |
| | 0.0814 | 1.0060 |
| | 0.0961 | 1.0078 |
| water + 0.30 mol·kg$^{-1}$ B | | |
| 303.15 | | ($\rho_o$ = 1.0003) |
| | 0.0436 | 1.0060 |
| | 0.0525 | 1.0072 |
| | 0.0914 | 1.0122 |
| 308.15 | | ($\rho_o$ = 0.9986) |
| | 0.0436 | 1.0043 |
| | 0.0525 | 1.0055 |
| | 0.0914 | 1.0104 |

Figure 1. An example of a difficult-to-consume table.

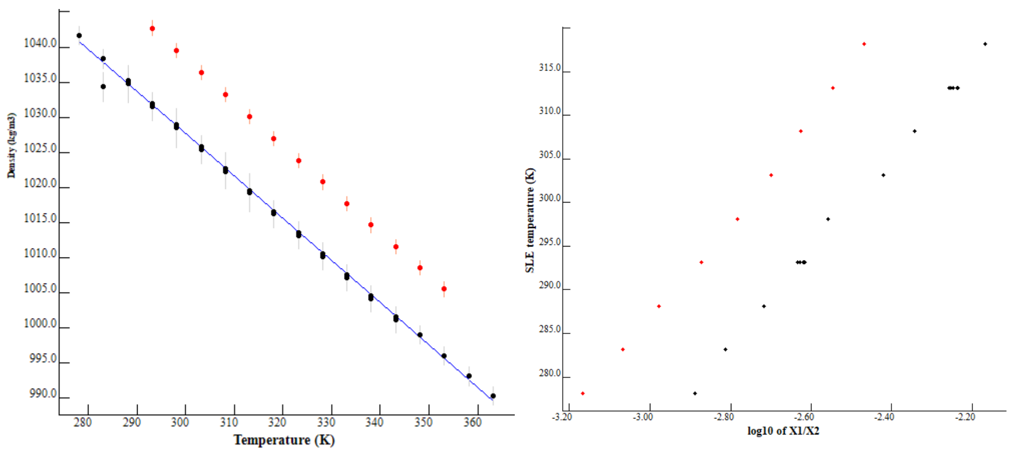


Figure 2. Examples of data inconsistency revealed from multiple publications: density of 1-hexyl-3-methylimidazolium dicyanamide and solubility of Benorilate in acetonitrile.

---

[1] Georgios M. Kontogeorgis, Ralf Dohrn, Ioannis G. Economou, Jean-Charles de Hemptinne, Antoon ten Kate, Susanna Kuitunen, Miranda Mooijer, Ljudmila Fele Žilnik, Velisa Vesovic, Industrial Requirements for Thermodynamic and Transport Properties: 2020, Ind. Eng. Chem. Res. 60 (2021) 4987-5013, doi: 10.1021/acs.iecr.0c05356
[2] Vladimir Diky, Ala Bazyleva, Eugene Paulechka, Joseph W Magee, Vikina Martinez, Demian Riccardi, Kenneth Kroenlein, Validation of thermophysical data for scientific and engineering applications, J. Chem. Thermodyn. 133 (2019) 208-222, doi: 10.1016/j.jct.2019.01.029
[3] Ala Bazyleva, Jens Abildskov, Andrzej Anderko, Olivier Baudouin, Yury Chernyak, Jean-Charles de Hemptinne, Vladimir Diky, Ralf Dohrn, J. Richard Elliott, Johan Jacquemin, Jean-Noel Jaubert, Kevin G. Joback, Ursula R. Kattner, Georgios M. Kontogeorgis, Herbert Loria, Paul M. Mathias, John P. O'Connell, Wolffram Schröer, G. Jeffrey Smith, Ana Soto, Shu Wang, Ronald D. Weir. Good reporting practice for thermophysical and thermochemical property measurements (IUPAC Technical Report), Pure Appl. Chem. 93 (2021) 253-272, doi: 10.1515/pac-2020-0403
[4] M. L. Huber, R. A. Perkins, M. J. Assael, S. A. Monogenidou, R. Hellmann, J. V. Sengers, New International Formulation for the Thermal Conductivity of Heavy Water. J. Phys. Chem. Ref. Data 51 (2022) 013102, doi: 10.1063/5.0084222
[5] I. H. Bell, A. Jäger, J. Res. NIST 121 (2016) 238, doi :10.6028/jres.121.011
[6] The Cambridge Crystallographic Data Centre (CCDC), https://www.ccdc.cam.ac.uk/ Accessed 3/31/2022
[7] Protein Data Bank, https://www.rcsb.org/ Accessed 3/31/2022
[8] Vladimir V. Diky, Robert D. Chirico, Randolph C. Wilhoit, Qian Dong, Michael Frenkel, Windows-Based Guided Data Capture Software for Mass-Scale Thermophysical and Thermochemical Property Data Collection, J. Chem. Inf. Comput. Sci. 43 (2003) 15-24, doi: 10.1021/ci025534t

[9] Eldata, the international electronic journal of physico-chemical data, https://www.worldcat.org/title/eldata-the-international-electronic-journal-of-physico-chemical-data/ Accessed 3/31/2022
[10] Biography, http://www.codata.info/codata02/bios/bio-kehiaian.htm Accessed 6/30/2022
[11] Electronic data submission to NIST Thermodynamics Research Center. Fluid Phase Equilib. 226 (2004) v.
Electronic data submission to the NIST Thermodynamics Research Center, J. Chem. Thermodyn. 36 (2004) iv.
Electronic data submission to NIST Thermodynamics Research Center. Thermochim. Acta 421 (2004) 241.
K. N. Marsh, New process for data submission and dissemination, J. Chem. Eng. Data 48 (2003) 1.
W. M. Haynes , Electronic data submission for the International Journal of Thermophysics, Int. J. Thermophys., 26 (2005) 307.
[12] arXiv, https://arxiv.org/ Accessed 6/30/2022.
[13] M. Frenkel, Q. Dong, R.C. Wilhoit, K.R. Hall, TRC SOURCE database: a unique tool for automatic production of data compilations, Int. J. Thermophys. 22 (2001) 215–226, doi: 10.1023/A:1006720022161
[14] ThermoML – An XML-based IUPAC Standard for Thermodynamic Property Data, https://iupac.org/what-we-do/digital-standards/thermoml/ Accessed 6/30/2022
[15] Solubility Data Series, https://iupac.org/what-we-do/databases/solubility-data-series/ Accessed 3/31/2022
[16] NIST Chemistry WebBook, https://webbook.nist.gov/ Accessed 3/31/2022
[17] NIST/TRC ThermoML Archive, https://trc.nist.gov/ThermoML/ Accessed 3/31/2022
[18] ILThermo database, https://www.nist.gov/mml/acmd/trc/ionic-liquids-database Accessed 3/31/2022.
[19] NIST Gas Hydrates Database, https://gashydrates.nist.gov/ Accessed 3/31/2022
[20] Marko Mebelli, Danai Velliadou, Marc J. Assael, Marcia L. Huber, Reference Correlation for the Viscosity of Ethane-1,2-diol (Ethylene Glycol) from the Triple Point to 465 K and up to 100 MPa, Int. J. Thermophys. 42 (2021) 116, doi: 10.1007/s10765-021-02867-0
[21] D. Kirk Nordstrom, Models, validation, and applied geochemistry: Issues in science, communication, and philosophy, Appl. Geochem. 27 (2012) 1899-1919. doi: 10.1016/j.apgeochem.2012.07.007
[22] Robert D. Chirico, Michael Frenkel, Joseph W. Magee, Vladimir Diky, Chris D. Muzny, Andrei F. Kazakov, Kenneth Kroenlein, Ilmutdin Abdulagatov, Gary R. Hardin, William E. Acree, Jr., Joan F. Brenneke, Paul L. Brown, Peter T. Cummings, Theo W. de Loos, Daniel G. Friend, Anthony R. H. Goodwin, Lee D. Hansen, William M. Haynes, Nobuyoshi Koga, Andreas Mandelis, Kenneth N. Marsh, Paul M. Mathias, Clare McCabe, John P. O'Connell, Agilio Pádua, Vicente Rives, Christoph Schick, J. P. Martin Trusler, Sergey Vyazovkin, Ron D. Weir, Jiangtao Wu, Improvement of Quality in Publication of Experimental Thermophysical Property Data: Challenges, Assessment Tools, Global Implementation, and Online Support, J. Chem. Eng. Data 58 (2013) 2699–2716, doi: 10.1021/je400569s
[23] J.P.A. Ioannidis, Why most published research findings are false, PLoS Med 2 (2005) e124, do: 10.1371/journal.pmed.0020124

R. Horton, Offline: What is medicine's 5 sigma?, Lancet 385 (2015) 1380, doi: 10.1016/S0140-6736(15)60696-1
J.P.A. Ioannidis, An epidemic of false claims, Sci. Am. 304 (2011) 16, doi: 10.1038/scientificamerican0611-16
M. Felaefel, M. Salem, R. Jaafar, G. Jassim, H. Edwards, F. Rashid-Doubell, R. Yousri, N.M. Ali, H. Silverman, A cross-sectional survey study to assess prevalence and attitudes regarding research misconduct among investigators in the middle east, J Acad. Ethics 16 (2018) 71-87, doi: 10.1007/s10805-017-9295-9
J. Belluz, B. Plumer, B. Resnick, The 7 biggest problems facing science, according to 270 scientists, VOX, September 7, 2016. https://www.vox.com/2016/7/14/12016710/science-challeges-research-funding-peer-reviewprocess (accessed 3/31/2022)
[24] P. T. Cummings, Theo W. de Loos; J. P. O'Connell, W. M. Haynes, D. G. Friend, A. Mandelis, K. N. Marsh, P. L. Brown, R. D. Chirico, A. R. H. Goodwin, J. Wu, R. D. Weir, J. P. M. Trusler, A. Pádua, V. Rives, C. Schick, S. Vyazovkin, L. D. Hansen, Joint statement of editors of journals publishing thermophysical property data: Process for article submission for The Journal of Chemical Thermodynamics, Fluid Phase Equilibria, International Journal of Thermophysics, Thermochimica Acta, and Journal of Chemical Engineering Data. Fluid Phase Equilib. 276 (2009) 165-166; Int. J. Thermophys. 30 (2009) 371-373; J. Chem. Eng. Data 54 (2009) 2-3; J. Chem. Thermodyn. 41 (2009) 575-576; Thermochim. Acta 484 (2009) vii-viii.
[25] Combustion Calorimetry Tool, NIST Standard Reference Database 206, https://trc.nist.gov/cctool/ Accessed 3/31/2022
[26] M. Ijaz Khan, A. Alsaedi, Sumaira Qayyum; T. Hayat, Retraction Note to: Consequences of Binary Chemically Reactive Flow Configuration of Williamson Fluid with Entropy Optimization and Activation Energy. Int. J. Thermophys. 42 (2021) 166. doi: 10.1007/s10765-021-02918-6
[27] Michael Frenkel, Robert D. Chirico, Vladimir Diky, Xinjian Yan, Qian Dong, Chris Muzny. ThermoData Engine (TDE): Software Implementation of the Dynamic Data Evaluation Concept. J. Chem. Inf. Model. 45 (2005) 816-838 doi: 10.1021/ci050067b
[28] ThermoData Engine, NIST Standard Reference Database 103b, https://www.nist.gov/mml/acmd/trc/thermodata-engine/srd-nist-tde-103b Accessed 3/31/2022
[29] Kvetoslav Ruzicka, Vladimir Majer, Simple Controlled Extrapolation of Vapor Pressures Towards the Triple Point. AIChE J. 42 (1996) 1723-1740
[30] Ala Bazyleva, William E. Acree Jr., Vladimir Diky, Glenn T. Hefter, Johan Jacquemin, M. Clara F. Magalhães, Joseph W. Magee, D. Kirk Nordstrom, John P. O'Connell, James D. Olson, Ilya Polishuk, Kurt A. G. Schmidt John M. Shaw, J. P. Martin Trusler, Ronald D. Weir, Reference Materials for Phase Equilibrium Studies. 2. Solid-Liquid Equilibria (IUPAC Technical Report), Pure Appl. Chem. (2022) in press.
Ala Bazyleva, William E. Acree Jr., Ian H. Bell, Robert D. Chirico, Theodoor W. De Loos, Vladimir Diky, Glenn T. Hefter, Johan Jacquemin, Jeong Won Kang, M. Clara F. Magalhães, Joseph W. Magee, D. Kirk Nordstrom, John P. O'Connell, James D. Olson, Ilya Polishuk, Kurt A. G. Schmidt, John M. Shaw, J. P. Martin Trusler, Ronald Weir, Recommended Reference Materials for Phase Equilibrium Studies, IUPAC Project No. 2011-037-2-100. https://iupac.org/projects/project-details/?project_nr=2011-037-2-100 Accessed 6/30/2022