Material EASE

AMPTIAC 9

Jorge Luis Romeu
IIT Research Institute
Rome, New York

DATA QUALITY AND PEDIGREE

Introduction

Data are scientific or technical measurements represented by numbers or other means [1]. Their importance in scientific and engineering work can never be stressed enough. The heart and soul of statistical analysis are the data. It is well known that bad data induces the so-called *GIGO* model: Garbage In, Garbage Out. In plain English, no statistical procedure will yield good results if the data used with them are bad.

In a similar way, pedigreed data is essential during material selections, such as in the design of a critical component for an airplane. The use of bad data may result in poor or improper material selections having very serious consequences in manufacturing and product performance. Discerning good data from bad data is neither readily nor easily accomplished. At first glance, much of the data may look alike, even when they may be completely different in terms of their source and their quality. The objective of this article is to provide some direction to the materials engineer as how to recognize and distinguish good data, thus setting it apart from bad data.

Broadly speaking, *good data* refers to accurate, complete and trustworthy data, which may be easily accessed by the material engineer from a reliable source, such as a handbook or database. Good quality data has first been carefully collected by an accredited testing organization that follows a set of strict quality control guidelines. As a part of this process, the collecting organization has carefully reviewed the experimental and test procedures of the data originators, checked it for consistency and registered this ancillary information, known as *metadata*, jointly with the analysis data itself. The reporting organization also checks the analysis results via a series of internal and external procedures, known as data validation and certification. Finally, the entire data information package is placed in an accessible medium, such as an electronic database, where data may easily be retrieved and used by interested practitioners.

In short, good data may be recognized because they bear a *good pedigree* – and bad data do not. The concept of data pedigree is difficult to define, but easy to understand at an intuitive level. In the same way that one would look at a dog's pedigree, one may look at a data pedigree. A dog may look very good as it is presented at a dog show, but a prospective owner may only be sure that it is worth paying a high price for it when its pedigree is verified. A pedigree is established by determining who the dog's parents and grandparents were, how many prizes they won, its health history, siblings, breeding organization, etc. Finally, all this information is certified by a respected dog breeding society to affirm that the information is bona fide and has not been sloppily or hastily evaluated.

Understanding the practical importance of using good data and its associated metadata and pedigree requires answering several important questions: How is good data generated; how is it requested and used; how does a trustworthy data bank accept and process them; what is the data accreditation process like; and how does one justify the cost of generating quality data sets?

This article discusses the importance of determining whether data are "good" or not. It also deals with the closely associated concept of *data pedigree*. Moreover, it addresses some of the issues that arise when answering questions about data quality and pedigree. Lastly, the associated *metadata* characteristics required for the appropriate application of the statistical procedures, as described in MIL-HDBK-5G [2] and MIL-HDBK-17.1D [3], are discussed.

Data and Metadata

Several well known specialists have discussed material data problems at length [1,5,6]. Since materials data originate from tests developed under specific conditions [1], the corresponding data about the data, or *metadata*, need to be recorded. Without such ancillary information, the experimental results will lose their contextual meaning. For example, the use of fatigue information is closely associated with the conditions in which the fatigue occurred, and with the related material specifications.

Examples of metadata information include characteristics of test materials, specimens, experimental test conditions, measurement and calibration procedures, recorded readings, specific ambient conditions, etc. In addition, metadata is used to perform statistical analysis, compare different test samples, and establish smoothing curves. It is also used in the process of data validation.

Metadata is often missing or incomplete, creating voids in data collection. An easy solution would be to collect and store everything about the data, but this would create even larger problems. It is essential to consider the ease of information retrieval by potential users via electronic databases or other storage media. If this information is ever to be used to facilitate its retrieval, the storage of materials information must be well planned, then implemented in such a way that it is easily and uniformly accessed. To this end, extensive standard formats have already been established [4].

Metadata can also be used in assessing which data sets to pool together. For example, apparently similar data sets may have some specific difference (say an ambient condition) that sets them apart. Also, experimental techniques improve or change with time. New parameters are identified that affect test results. If metadata are available, we may correct the original data for these new developments. Finally, the ancillary information obtained from the metadata provides the variables for regression and analysis of variance or covariance, among other statistical procedures. The functions obtained can then be used to correct or reclassify the data, as well as to fill in data gaps.

Material E A S E

One of the technical publications by the ASTM Committee E-49, the Computerization of Materials Property Data [5], deals with the problems of facilitating data storage and retrieval. The committee presents a list of materials descriptors and guidelines for reporting test data. The list emphasizes the importance of a unique format for the identification of metals and metal alloys and of polymers. A standard data format for the computerization of test data and mechanical properties is necessary to make comparisons between data sets. Such comparisons are valid when all relevant fields are obtained. This shows the importance of recovering all the information requested in standardized formats, in addition to just reporting test data.

Standardizing the information content again raises the problem of data evaluation for quality and reliability. This is another crucial issue for those who generate materials data as well as for those who use the data in their engineering design work. Similarly, collecting all available information about the data is not a solution, neither is storing all available data sets. When confronted with this issue, engineers must perform a selection.

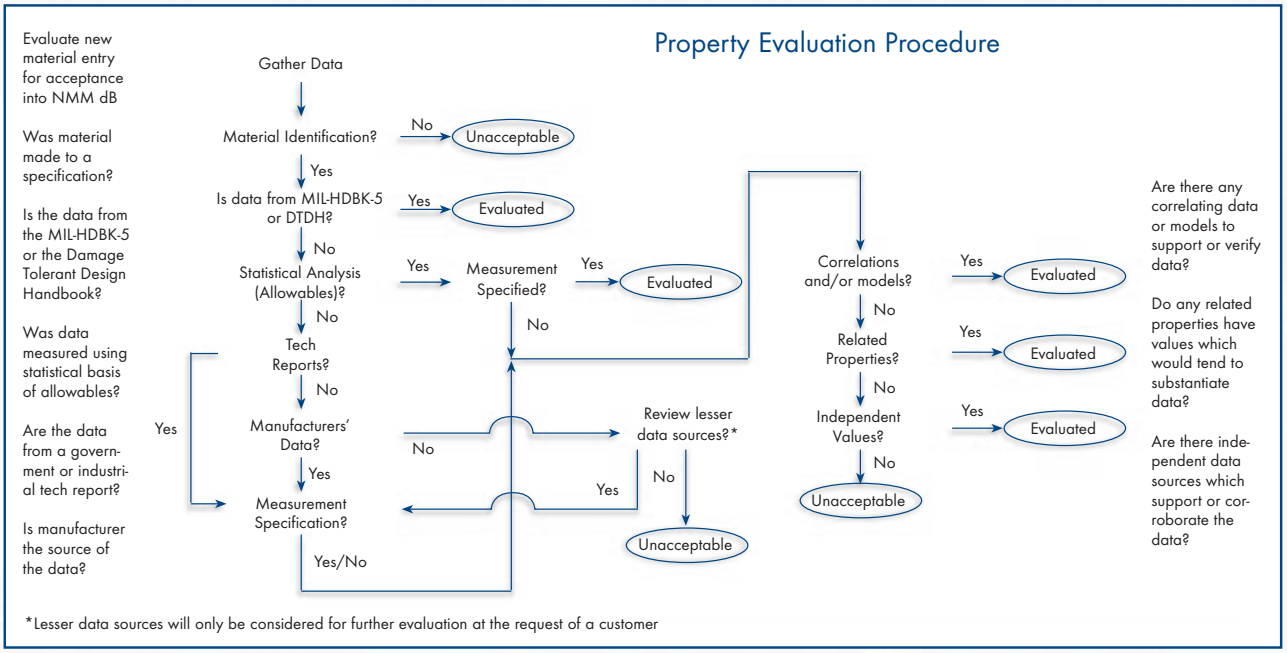
This problem has also been thoroughly studied. ASTM STP 1140 provides a thorough treatment of data quality and reliability issues [6]. It provides several lists of guidelines for subjective assessment, validation, analysis and certification of material data based on the ASTM Committee E-49.05 report on data quality. It identifies and discusses several data levels. These levels, from lowest to highest are unanalyzed (raw) data, analyzed individual results, mathematically reduced data, evaluated, validated, and finally certified data. The precise definitions for these material data classification levels are included in the reference [6].

Also discussed are standard guidelines for database management, regarding quality and reliability, emphasizing identification of data sources, proof checking of data, correcting errors, and assessing user satisfaction. ASTM STP 1017 also provides extensive guidelines for data evaluation [5]. They are classified into subjective assessment, validation, analysis and certification, giving lists of activities for each category.

These guidelines as well as the problem of quality assessment of data sets [1], are further discussed in ASTM's database manual. Mixing good and bad data does not improve a data bank – on the contrary, it lowers the quality of the mix. In particular, mixing good and bad data increases the data variability, which in turn lowers the accuracy of the derived allowables. Data may be evaluated through a complete process that starts with assessing the organization that generates it and ends with a comparison of the originated test results with well-accepted and certified results.

An organization that creates data can be evaluated through its experience, accountability, bias, calibration practices, and management attitudes such as the separation between data generators and evaluators. To avoid conflicts of interest, an independent group should carry out the data validation, if such validation is done within the same organization.

The ASTM database manual provides a well-defined set of activities for the validation team [1]. It provides lists of guidelines for the validation process and for establishing data quality indicators. The most important guidelines are to work with plural teams that include members of uniformly high experience and ability, that base their decisions on true consensus and whose members work within the limits of their knowledge and experience. It also suggests avoiding inclusion of members of questionable reliability, experience or known bias. He also provides a glossary of terms concerning data, quality and their validation process.



Lastly, the manual states that *certification*, as opposed to validation, is the recognition by a *warranting authority*, of the quality of the data. These authorities have to be uniformly recognized and well established and should certify only for their area of expertise. Examples include committees of professional societies and official organizations such as Underwriter's Laboratories and the Society of Automotive Engineers.

Types of Data and Databases

There are as many different types of material data and databases as there are different types of uses for them. Material data are thus collected, processed and organized accordingly. Material databases may be classified according to different schemes that include data, user, and application and access types [7]. Material information should flow from data generators (e.g. testers) to data users (e.g. handbooks) as flows a slow moving river. Such an information flow consists of four stages: data generation, analysis, aggregation and reanalysis.

As the computer has become universal, more work is automated and performed through or with computers. Much of the materials testing found these days is done this way. The resulting data collection from materials test equipment is thus entered directly onto computers. These computerized collections of original test results data are referred to as *laboratory notebook databases* [7]. They can be computer searched, analyzed, updated and manipulated, among other functions. And they also contain very useful ancillary data.

Report databases are those that provide analysis results of test data [7]. They may include sophisticated correlations, graphical comparisons, coefficients, parameters, etc; and may appear in the literature (journal articles and technical reports) or in handbooks. They serve several functions, including derivation of properties, extension of data domain and improved understanding. One cannot under-emphasize the importance of the data analysis stage and of the need to preserve the results of these (intermediate) analysis procedures.

Handbook databases, conversely, compile data and other results into collections (MIL-HDBK-5 and 17, for example) and constitute the data source of first choice [7]. Not too many of them exist and the need for them is great. Organizations such as AMPTIAC foster the creation and development of such materials databases.

Data targeted to specific applications may be classified as *applications databases*. These are derived for convenience, or for the quality of their data, and are built for solving specific problems They may be custom-built for some specific project, but they are usually not maintained nor are they updated beyond the life of such specialized work.

He also discusses the classification of databases by user groups and presents tables of such uses. Database uses include the calculation and evaluation of material properties; the design, development, selection and performance evaluation of materials; and failure analysis and product information. Databases may also be classified into personal, group, institutional, collegial or public types, according to their users and their source.

Data Accreditation

Collecting good data and rejecting bad data is paramount to any engineering or design activity. To emphasize this, the following paragraphs highlight the process that leads to providing the users with assurances about the quality of the data.

The data evaluation procedure for each application is unique. ASTM E-49 provides some general evaluation procedures, but to serve any specific purpose, these expansive flow charts must be reduced and customized. Figure 1 provides an example of the evaluation process for defense-grade structural materials. A complete methodology for data evaluation of high temperature semiconductors is provided in the reference [9]. Data are divided into seven acceptability level classes: unevaluated, research (preliminary and work in progress), typical (from surveys), commercial (manufacturer's), evaluated (basic acceptance), validated (confirmed via correlations and models) and certified (standard references).

If materials are not well specified, data is classified as *unacceptable*. If dealing with manufacturer's data and the measurement methods are not described, it is classified as *commercial*. If it is survey data, it is classified as *typical*. Subsidiary data is classified as *unevaluated*. Data is also classified as *unevaluated* if none of the above apply. If the data provides (or is checked against) standard reference values, it is classified as *certified*. Otherwise, if correlation or models have been applied, it is classified as *validated*. If data is checked by independent values, it is classified as evaluated. If the data is not checked, but real properties are provided, the data is also classified as *evaluated*. If peer-reviewed and part of an interim report, the data are classified as *research in progress*. Otherwise, if results are incompatible, data may have to be reassessed and reclassified as either *evaluated or unacceptable*. Precise definitions of these classifications are provided as is a discussion of the activities involved in working with them. In addition, the authors provide specific examples of applications of analytical, statistical and graphical methods to the validation of the data.

There is an additional treatment of quality and reliability issues of material databases, as well as the ASTM Committee E-49 criteria [6]. Standardization of the information is basic, because it allows uniform and universal access to it. Standardization is obtained through uniform fields in a database. The recommended field content descriptions include database name or acronym, full title, name of producer, address, telephone, types of data, materials classes, property classes, independent variables, testing variables, updating frequency, evaluator name and organization, availability, and delivery media.

Other database quality indicators include data presentation issues (e.g. accuracy), unit conversions and other data manipulations [1]. Such issues are often taken for granted, but they are relevant to the values recorded. Barrett provides a list of quantifiable quality indicators for assessing data records or databases. These indicators are grouped into data quality (e.g. source, statistical basis, evaluation status), database quality (e.g. completeness, support) and database operation (e.g. availability, access). This list of indicators may be regarded as a vector in a multidimensional space. Under this multivariate approach, database comparisons may be established by looking into each component.

Uses and Cost of Good Data

So far, we have discussed materials data, their quality and their pedigree. Obtaining good data however, does entail a cost. But a benefit is also derived from its uses in engineering design. Some of the advantages of creating materials information systems include the provision of a central source of best *available* data, of preferred materials and processes, of experience gained in manufacturing. The fact that the data used is traceable and the metadata can be compared to that of other data sets is most valuable [10].

The five stages of engineering design each require materials information

[10]. These stages are R&D, product scheme, detail drawing, production qualification and in-service product report. In all of them, the materials information process has a valid and useful input.

Many factors affect materials selection: specific properties (e.g. fracture mechanics, fatigue, strength and ductility), compatibility (e.g. corrosion, wear, thermal mismatch) and manufacturing (e.g. availability, cost, machinability, inspection, formability). The best materials data information systems contain information on all of these factors. For example, if one only considered property data with no regard to availability or cost, diamond would frequently top the list as the best material for many applications. When the real factors of availability and cost are considered, diamond falls far down the list of desirable materials for a given application.

Good and reliable data does cost money to collect, validate, install, deliver and maintain [11]. In these economic times, one is required to perform a cost-benefit analysis of the engineering information system and to show it's value-added to the design process. The problem is that the information activity hides its benefits quite well. Frequently it is easier to show the losses incurred by not using good information in the design process than it is to show the gains obtained by using good information systems. Regarding this situation, the cost-benefits relationship should be uncoupled until benefits are better characterized and understood. In addition, different viewpoints on information benefits need to be recognized. These viewpoints should include not only those from system developers, but also the viewpoints of users (of existing systems) as well as of potential users (of new systems under development) and of those non-users who can influence the process (e.g. managers).

A new approach to the quantitative evaluation of benefits would be to presuppose that economic benefits do exist and hence should reflect somewhere in the system. Therefore, database functions and features should be linked with tangible user benefits – some of which may not be readily identified or appreciated. Some economic and social advantages perceived or sought by the user of a materials information system include reduced design cycle time, lower labor, material and capital costs, improved product quality and reliability and enhanced education and work interest for the information system user.

Conclusions

In this article, we have summarized the main discussion issues regarding materials data, their quality and their pedigree as well as their relationships to the construction, maintenance and use of materials databases for engineering design. With it, we expect to generate additional questions for the AMPTIAC Forum and to receive further comments and suggestions for developing useful materials in future newsletter issues.

References

1. Barrett, A. J; **Data Evaluation, Validation and Quality**. *ASTM Manual on The Building of Material Databases*. Crystal H. Newton, Editor; ASTM Manual Series: MNL 19; 1993, pp 53-67.
2. **MIL-HDBK-5G**; *Metallic Materials and Elements for Aerospace Vehicle Structures*. November 1994.
3. **MIL-HDBK-17.1D**; *Composite Materials Handbook*.
4. **ASTM Standards on the Building of Materials Databases**; ASTM Series, 1993.
5. Kaufman, J.G.; **Standards for Computerized Materials Property Data –ASTM Committee E-49**. *Computerization and Networking of Materials Databases*; Glazman and Rumble, Editors; ASTM STP 1017; 1989, pp 7-22.
6. Kaufman, J.G.; **Computerization and Networking of Materials Databases: Third Volume**; Barry and Reynard, Editors; ASTM STP 1140; 1992, pp 64-83.
7. Rumble, J. R.; **Types of Materials Databases**. *ASTM Manual on The Building of Materials Databases*; Crystal H. Newton, Editor; ASTM Manual Series: MNL 19; 1993, pp 27-33.
8. **Navy Metallic Material Property Database/Materials Information System Final Report: Phase I Summary**; AMPTIAC; SPO700-97-D-4001.
9. Munro, R. G., Chen, H.; **Data Evaluation Methodology for High-Temperature Superconductors.** *Computerization and Networking of Materials Databases*; Nishijima and Suichi, Editors; ASTM STP 1311; 1997, pp 198-210.
10. Newley, R.A.; **The Integration of Materials Information into Engineering Design; Computerization and Networking of Material Databases**; Barry and Reynard, Editors; ASTM STP 1140; 1992, pp 192-205.
11. Barrett, A. J.; **The Benefits and Economic Consequences of Materials Property Databases**. *Computerization and Networking of Materials Databases: Second Volume*; Kaufman and Glazman, Editors; ASTM STP 1106; 1991, pp 17-25.

EMAIL: amptiac@iitri.org
http://amptiac.iitri.org

PHONE: 315.339.7117
FAX: 315.339.7107

AMPTIAC is a DoD Information Analysis Center Administered by the Defense Information Systems Agency, Defense Technical Information Center and Operated by IIT Research Institute