Journal of Thermal Analysis and Calorimetry, Vol. 60 (2000) 1101–1110

DATABASES IN MATERIALS SCIENCE
Contemporary state and future

J. Fiala[1*] *and J. Šesták*[2]

[1]ŠKODA Research Ltd, Tylova 57, CZ-31600 Plzeň & University of Mining and Metallurgy Tř. 17 listopadu 15, CZ-70833 Ostrava–Poruba
[2]Institute of Advanced Studies of the Charles University, Legerova 63, CZ-12000 Prague & Institute of Physics, The Czech Academy of Sciences, Cukrovarnická 10, CZ-16253 Prague, Czech Republic

Abstract

The state-of-the art of databases in materials science, understood as ordered and stored information, is discussed and the outlook for their development is deliberated. The text is divided in sections dealing with the glory, misery and future vision of databases, thermal analysis having its own comment.

Keywords: bibliography, databases, information science

Introduction

Science and information

Motto: 'For in much wisdom is much grief: and he that increaseth knowledge increaseth sorrow' [Preacher 1:18]

It is clear that the main product of science is information. From this point of view, there is, in science, seldom anything more respectable than databases which store the

* Author for correspondence: Šimerova str. 16, 32006 Plzeň, Czech Republic, Fax: (+ 420 19) 7734335

1418–2874/2000/ $ 5.00

Akadémiai Kiadó, Budapest
Kluwer Academic Publishers, Dordrecht

information gathered by scientists and put it in order. On the other hand, there are still controversial issues and open problems to be solved in order that the information (and the derived databases) will better serve the ultimate endeavour of science – the pursuit of discovery and truth.

Intermezzo: preface related to thermal analysis

The ideas discussed in this paper are the individual approach of ordinary scientists (specialised in X-ray diffraction (Fiala) and thermal analysis (Šesták)) from a 'laboratory battlefield' where they are both producing and consuming information and have to use information services. Well-trained specialists in information services, systematisation and assessments might have different views on how to rank and discuss databases grown inside big information corporations, usually pharmaceutical corporations, linking computational chemistry with molecular modelling, etc. Debates have started recently and will probably continue for a long time on various national (e.g., local dispute of [1] *vs.* [2]) and international levels.

Let us make some initial remarks related to a specific field of our interest, i.e., thermal analysis (TA), as well as to the general scope of data publication. The Journal of Thermal Analysis and Calorimetry (JTAC) is one of the periodicals covering the field of thermal analysis and related thermochemistry and material science and belongs to a family of about 60 000 scientific journals that publishes annually about 10^6 papers on 10^7 pages. The questions then arise as to the appropriate role of such a specific journal, and the place of JTAC among so many presently existing scientific periodicals. The answers to these questions may be useful not only for the JTAC Editors, but also for prospective authors trying to locate their articles properly, as well as for researchers needing to identify suitable journals when the interaction between specialities or disciplines pushes them beyond the borders of familiar territory.

It is generally recognized that almost three-quarters of all published articles are never cited and that a mere 1% of all published articles receives over half of the citations from the total number. These citations are also unequally distributed over individual journals. Articles written by a Nobel-prize winner (or other high-profile scientist) are cited about 50 times more frequently than an average article of unknown affiliation cited at all in the given year. About 90% of all the actual information ever referred to represents a mere two thousand scientific volumes, each volume containing roughly 25 papers. The average library also removes about 200 outdated volumes each year, because of shortages of space, and replaces them with newer issues.

What is the driving force for the production of scientific papers? Besides the natural need to share the latest knowledge and common interests, there is the often repeated factor of 'publish or perish' which is worthy of serious re-thinking, particularly now in the age of resourceful computers. We have the means of safeguarding the originality of melodies, patents and even ideas, by rapid searching through a wide range of databases, but we are not yet able (or willing?) to reduce repetitions, variations and modifications of scientific ideas. Printed reports of scientific work are necessary to assure continued financial support and hence the survival of scientists and, in fact, the routine continuation of science at all. It would be hypothetically possible

to accelerate the production of articles by applying a computer-based 'Monte Carlo' method to rearrange various paragraphs of already-existing papers so as to create new papers, fitting into (and causing no harm in) the category of 'never-read' articles. Prevention or restriction of such an undesirable practice is mostly in the hands of scientific referees (of those journals that do review their articles) and their ability to be walking catalogues and databases in their specialisation.

The extent of the task facing a thermal analyst is potentially enormous. For the 10^7 compounds presently registered, the possibility of 10^{14} binary reactions exists. Because all reactions are associated with thermal changes, the elucidation of a large number of these 10^{14} reactions could become a part of the future business for thermochemistry and TA and, in due course, the subject of possible publications in JTAC and other journals. The territory of thermal treatment and analysis could become the most generally enhanced aspect of reactivity studies.

The thermal properties of samples are monitored using various instrumental means. Temperature control is one of the basic parameters of TA experiments, but there are only a few alternatives for its regulation, i.e., isothermal, constant heating/cooling, oscillating and modulated, or sample determined (during quenching or explosions). Heat exchange is always part of any experiment, so reliable temperature measurements and control require improved sophistication of TA instruments. These instruments can be considered as 'information transducers', invented and developed through the skill of generations of scientists in both the laboratory and manufacturers' workshops. The process of development is analogous to the process for obtaining useful work, where one needs to apply, not only energy, but also information, so that the applied energy must either contain information itself, or act on some organised device, such as a thermodynamic engine (similarly understood as an 'energy transducer' [3]). In TA, the applied heat may be regarded as a 'reagent' which, however, is lacking in information content in comparison with other 'instrumental reagents', richer in information capacity, such as various types of radiation, fields, etc. We, however, cannot change the contributed information content of individually applied reagents and can only improve the information level of our gradually invented transducers.

For comparison, the method of X-ray diffraction (XRD) involves measuring the directional distribution of X-rays diffracted by matter. One obtains not only information on the molecular structure (i.e., the organisation of atoms in a molecule of matter involved), but also information on the supramolecular structure as expressed by the arrangement of molecules in a crystal and the assembly of crystals in polycrystalline bulk, including such minute structural aspects as the orientation, shape, size and internal defects of individual crystalline blocks. Such measurements can also be made from low to high temperatures, using sophisticated measuring chambers in addition to the sophisticated state of the XRD instrumentation itself. This increased information content is reflected by the richness of the structural databases, containing about 10^5 XRD reference identification spectra, while the number of known TA spectra is at least one order lower. Although the extent of TA studies is broader than that of XRD, XRD and its associated disciplines of crystallography attract more public attention, as

reflected in the larger membership of crystallographic societies compared with the memberships of TA societies (such as ICTAC). It follows that a better-directed method and co-operation towards more restricted but well-specified goals is more appropriate in scientific society. This may be related to the built-in information content of each distinct 'reactant' (special X-rays *vs.* universal heat) which is important for the development of the field in question. It certainly does not put a limit on the impact of combined multiple techniques in which the methods of TA can play either a crucial or a secondary role. Both fields then claim superior competence (e.g., thermodiffractometry). These simultaneous methods, can extend from ordinary combinations of, e.g., DSC with XRD or microscopy, up to real-time WAXS-SAXS-DSC, using synchrotron facilities. Novel combinations, such as atomic force microscopy fitted with an ultra-miniature temperature probe, are opening new perspectives for studies on materials, and providing unique information rewards.

Glory of databases

Motto: 'And ye shall know the truth, and the truth shall make you free' [John 8:32]

In January 1999, the Chemical Abstracts Service (CAS) registered the 19 000 000th chemical substance. Since 1995, more than a million new substances have been registered annually. Table 1 shows how the number of newly registered substances has been increasing since 1965 when the CAS Register System started. The world's largest and most comprehensive index of chemical literature, the CAS Abstracts File, now contains more than 18 million abstracts. During 1998, this file increased by 681 008 new abstracts. Of these, 559 009 were abstracts of papers (that appeared in some 14 000 journals from 150 nations), 117 815 patents (from 29 nations) and 4 184

Table 1 Growth of the Chemical Abstracts Service Registry File in the three year periods

Year	Number of substances registered each year	Substances on the cumulative file at the given year end
1965	211 934	211 934
1968	270 782	796 479
1971	351 514	1 952 447
1974	319 808	2 988 020
1977	369 676	4 077 703
1980	353 881	5 141 872
1983	418 905	6 346 713
1986	528 966	8 803 687
1989	615 987	9 912 619
1992	690 313	11 950 526
1995	1 186 334	14 594 302
1998	1 679 913	18 920 403

books related to chemistry. The last, 13th Collective Index to Chemical Abstracts, covering the years 1992–1996, cited 3 130 955 documents (i.e., journal papers, patents and books) of interest for chemistry. About a half of the one million papers that are published annually in scholarly journals deal with chemistry (which we consider a natural part of materials science [4, 5]). Derwent, the database producer and world's largest patent authority, registers some 600 000 patents and patent equivalents annually; 45% out of them bear upon chemistry. Beilstein Substance Database now covers more than 8 million organic compounds and Beilstein Cross Fire plus Reactions (BCF&R), the largest and most comprehensive database of chemical reactions available, delivers information on 11 million reactions. The largest factual database of inorganic substances, Gmelin, covers more than one million inorganic substances. One of the most extensive printed sources of physical properties and related data, Landolt–Boernstein Numerical Data and Functional Relationships in Science and Technology, has more than 200 volumes (occupying some 10 metres of shelf space).

In recent years, we have witnessed growing cooperation between information institutions and their integration which makes the future development of databases and their usage more efficient. Since the beginning of 1998, Elsevier Science, which offers databases and electronic library products and publishes approximately 1200 scientific journals in all major scientific technical and medical disciplines, has acquired Beilstein Informationssysteme GmbH and entered into an exclusive licence with the Beilstein Institute to market and support the Beilstein Database which will be updated and enhanced by this Institute in the future. Over the past several years, The Information Centre for Diffraction Data (in Pennsylvania), which coordinates the production and dissemination of critically-evaluated reference spectra for identification of substances by X-ray powder diffraction, has concluded important agreements with three other database organisations: Fachinformationszentrum Energie, Physik und Mathematik GmbH (Karlsruhe), Cambridge Crystallographic Data Centre, and National Institute of Standards and Technology (Gaithersburg in Maryland) which control the world's largest structural databases 'Inorganic Crystal Structure Data', 'Cambridge Structural Database' (organic compounds) and 'Crystal Data' . Cooperative activities frequently take place within the framework of such organisations as the International Council of Scientific Unions and its Committee on Data for Science and Technology (CODATA), established in 1966 to improve the quality, reliability, processing, management and accessibility of data of importance to science and technology.

In the area of enhanced electronic communications and the world-wide development of information systems, electronic publishing and the Internet offer powerful tools for the dissemination of all type of scientific information. This is now made available in electronic form, not only from computerised databanks, but also from primary sources (journals, proceedings, theses and reports). It has definitely increased the information flux available by orders! However, because of the multitude of existing data of interest to materials science and technology and the variety of modes of presentation, computer-assisted extraction of numerical values of structural data, physico-chemical properties and kinetic characteristics from primary sources is

as difficult as before. As a consequence, the collection of these data, the assessment of their quality in specialised data centres, the publication of handbooks and other printed or electronic secondary sources (compilations of selected data) or tertiary sources (collections of carefully evaluated and recommended data), the storage in data banks, and the dissemination of these data to end users (educational institutions and basic scientific and applied research centres), still remain tedious and expensive operations.

Misery of databases

Motto: 'The spirit indeed is willing, but the flesh is weak' [Matthew 26:41]

The total amount of knowledge, collected in databases of interest for materials science, is impressive. On the other hand , the incompleteness of this collection is alarming. The 11 million reactions covered by the BCF&R database constitute only a negligible fraction of the total number of 200 000 000 000 000 binary reactions between 19 million already-registered compounds, not even considering tertiary reactions, etc. In other words, lots of substances are known, but little is known of how these substances react with each other! Compounds A and B may be familiar, but we have not yet found whether mixing them under appropriate conditions will produce a medicine, or a kind of fuel, etc. We cannot even imagine how to handle such a large database containing information on 10^{14} reactions. The number of substances registered grows by more than a million compounds annually, so the incompleteness of our knowledge of individual compounds increases rapidly.

Table 2 presents the numbers of reference spectra in the largest databases for the identification of substances by various spectroscopic techniques. From these data it is clear that there are no spectra available for the overwhelming majority of the19 million registered compounds, by means of which these substances could be identified. In other words, the great majority of substances 'known' today cannot be identified because we have no spectra for this purpose. The number of reference spectra grows by several thousand spectra annually, which is much slower than the pace of the registration of new compounds. It follows that the incompleteness of our knowledge quickly expands in this region, too. The fraction of the total number of registered substances, which can be recognised, also declines.

Table 2 The number of reference spectra for identification of substances by various techniques

UV-VIS	500 000
IR	160 000
MS	140 000
XRD	106 000
electron diffraction	103 000
13C-NMR	99 000
1H-NMR	48 000

The sizes of the world's largest collections of structural data are given in Table 3. From this table it is obvious that (three-dimensional, metrical) data on structures are not available for most of the 19 million registered substances. The number of published structures does not exceed several thousand in a year, far fewer than the number of compounds registered annually. Again the incompleteness of our knowledge grows quickly.

Table 3 The number of structures stored in the listed databases*

CSD	160 000
ICSD	43 000
CRYSTMET	60 000
Crystal Data	270 000

*CSD=Cambridge Structural Database (organics)
ICSD=Inorganic Crystal Structure Database (Karlsruhe)
CRYSTMET=National Research Council of Canada; Metals Crystallographic Data File
Crystal Data=database produced by the US National Institute of Standards and Technology

Statistical analysis demonstrates that 75% of substances are mentioned in the literature only once. A mere 1% of compounds produce almost 45% of the literature references. The great majority of synthesised compounds have never been utilised. Thus, not only our knowledge, but also our employment of newly registered compounds is incomplete. The abundance of substances, which we 'know', but which nobody has attempted to use for any purpose, represents a great potential for the future of materials science and technology.

As one example, portland cement and derived concrete, is the most widely used, yet by the materials science community most ignored, industrial material. The world-wide production of concrete exceeds that of steel by a factor of ten in tonnage and by more than a factor of 30 in volume. The reasons for the popularity of concrete are multifold: its components are available in almost every corner of the word; the cost of production is low compared with other engineering construction materials; it can be cast at ambient temperature to produce complex shapes; exhibits excellent resistance *vs.* water; and, although relatively weak in tension, readily lends itself to reinforcement. Concrete is also one of the oldest man-made composite materials. Ancient people noted the extraordinary properties of mixtures of certain volcano ashes with water and lime, over two thousand years ago and several structures were built, such as Roman Colosseum, that are still standing today, demonstrating the inherent durability of concrete-based materials. Recently, even the main disadvantage of concrete, its low strength, has been overcome by reducing the size of pores down to few micrometres by, e.g., tailored mixing and classified granulometry of cement powder fractions. The newly designed materials have became known as macro-defect-free concretes [6]. Tested in bending, they show a strength of more than 150 megapascals and, in that respect, become comparable with aluminium. The chemical reactions involved in concrete formation proceed to only a small percentage of completion due to the quick saturation of grain surface reactivity and resultant formation of firm intergranular contacts. With this improvement, concrete is even able to be used for spring

fabrication! How many unemployed possibilities are hidden among the huge number of registered compounds known, the majority of which have never been investigated in detail?

Materials databases expand steadily and quickly, becoming more and more difficult to comprehend. Man perceives serially and the speed with which he receives information is small. It is estimated that an average researcher reads 200 papers annually. This is much less than the one million papers published in the sixty thousand scholarly journals throughout the world. Even if a person could read the abstracts (about 2.5 min each) of the 559 009 papers on chemistry processed during the last year by Chemical Abstracts Service, in order to optimise the selection of those 200 papers, it would take him 23 292 h, which is more than two and half years. Fortunately there are other ways of making priority selections.

One can trust the search for information to computers which will quickly locate it by title, keywords, authors or citations, using complicated algorithms. The possibility of looking for atypical papers, which may bring unusual solutions, beyond the frame of the algorithms used, is, however, lost. Such papers may be very important and valuable. Most of the great discoveries made in any domain of science, including thermochemistry and materials science, during recent years, have developed out of uncommon concepts: quasicrystals, fullerenes, low-dimensional (quantum) electronics and optoelectronics, non- and nano-crystalline metals, high-temperature oxide superconductors, macro-defect-free cements, etc.

The intellectual treasure contained in scientific papers is great and any projection of this body of knowledge to simplify the search as performed by computers may lead to irreplaceable losses. People will rediscover, again and again, things that were already described in old and forgotten papers which they were not able to find. This rediscovered knowledge will be published in new papers, which, again, will not fully succeed in passing into the hands of those who could make use of them. The unwelcome result is steadily and quickly growing databases. The best way to make some data inaccessible, is to file them in a large database. Large databases can act like astronomical black holes in the information domain.

The steadily growing databases attract large numbers of scientists away from their active labour, but also give jobs to new specialists engaged in information and data assessment. Scientists may spend more and more time in searching ever more numerous and extensive databases. This allows them to be acquainted with the (sometimes limitless) results of the often extensive work of other scientists. On the other hand, this consumes the time which they could otherwise use in their own research work and, owing to this, they are prevented from making use of the results of the work of the other scientists. Gradually the flow of easily available information may impact on even youngsters and students, providing them with an effortless world of irrationality developed through games, perpetual browsing the Internet, trips to virtual reality, etc., although certainly not underestimating its significant educational aspects (encyclopaedia, languages, etc.).

If the aim of Science is the pursuit of truth, then the pursuit of information may divert people from Science (and curiously thus from the truth, too). If knowing the

truth makes a man free [*John 8:32*], the search for data may thus enslave him (eternally fastening his eyes to nothing more than the newborn light of information: a computer display).

Future vision

Motto: 'It is the spirit that quickeneth; the flesh profiteth nothing [John 6:63]

What is the way out of this situation? How can we make better use of the knowledge stored in steadily growing databases? An inspirational solution to this problem was foreshadowed in 1938 by H. G. Wells, who described an ideal organisation of scientific knowledge that he called the 'World Brain' [7]. Wells appreciated the immense and ever-increasing wealth of knowledge being generated during his time. While he acknowledged the efforts of librarians, bibliographers and other scientists dealing with the categorising and earmarking of literature, he felt that indexing alone was not sufficient to fully exploit this knowledge base. The alternative he envisioned was a dynamic 'clearing-house of the mind', a universal encyclopaedia that would not just catalogue, but also correlate, ideas within the scientific literature. The World Brain concept has been applied by E. Garfield 1978, who became a founder of the Institute for Scientific Information (ISI), of the ISI's citation databases [8] and, in particular, co-citation analysis [9]. The references that publishing researchers cite, establish direct links between papers in the mass of scholarly literature. They constitute a complex network of ideas that researchers themselves have connected, associated and organised. In effect, citations symbolise how the 'collective mind' of Science structures and organises the literature. Co-citation analysis proved to be a unique method for studying the cognitive structure of Science. Combined with single-link clustering and multidimensional scaling techniques, co-citation analysis has been used by ISI to map the structure of specialised research areas, as well as Science as a whole [10, 11].

Co-citation analysis involves tracking pairs of papers that are cited together in the source article indexed in the ISI's databases. When the same pairs of papers are co-cited with other papers by many authors, clusters of research begin to form. The co-cited or 'core' papers in the same clusters tend to share some common theme, theoretical, or methodological, or both. By examining the titles of the citing papers that generate these clusters, we get an approximate idea of their cognitive content. That is, the citing author provides the words and phrases to describe what the current research is about. The latter is an important distinction, depending on the age of the core papers. By applying multidimensional scaling methods, the co-citation links between papers can be graphically or numerically depicted by maps indicating their connectivity, possibly to be done directly through hyperlinks in the near future. By extension, links between clusters can also be identified and mapped. This occurs when authors co-cite papers contained in the different clusters. Thus, the co-citation structure of research areas can be mapped at successive levels of detail, from particular topics and subspecialties to less-explicit science in general.

Now, we propose that the numerical databases of interest to materials science be related to the ISI's bibliographic databases. Each paper bearing the data under consideration cites and is cited by other papers, which determine its coordinates in the (bibliographic) map of (materials) science. In this way, definite data (a definite point in data space) is related to a definite point in bibliographic space (image of these data in bibliographic space). The correlation between data (objects, points in data space) is expressed (capable of locating) as correlations between their images in bibliographic space (which is a well-approved technique developed and routinely performed by ISI [12]).

* * *

The authors are grateful for the financial support of the Grant Agency of the Czech Republic, No. 106/97/0589. We appreciate valuable discussions with our friends at and within the seminars held at the Institute of Fundamental Studies (directed by Prof. Z. Pinc) and at the Centre for Theoretical Studies (directed by Prof. I. Havel) both belonging to the Charles University in Prague.

References

1 J. Fiala and T. Havlík , Vesmir (Prague), 66 (1987) 395.
2 J. Horejší, Vesmír (Prague), 67 (1988) 5.
3 J. Šesták, ICTAC News, 31/2 (1998) 166.
4 J. Fiala, Archiwum Nauki o Materialach, 12 (1991) 85.
5 R. E. Maizell, How to Find Chemical Information, Wiley, New York 1987.
6 J. D. Birchall, Sci. Am., 248 (1983) 104.
7 H. G. Wells, World Brain, Doubleday, New York 1938.
8 J. Fiala, Thermochim. Acta, 110 (1987) 11.
9 E. Garfield, Citation Indexing – its Theory and Application in Science, Technology and Humanities, Wiley, New York 1979.
10 H. Small and E. Garfield, J. Informat. Sci., 11 (1985)147.
11 H. Small, Scientometrics, 26 (1993) 5.
12 E. Garfield, Current Contents of the Physical, Chemical and Earth Sciences, 34 (1994) 5.