KfK 4562
März 1989

Proposal to Make our Information System more Effective

E. L. Haase
Institut für Nukleare Festkörperphysik

Kernforschungszentrum Karlsruhe

KERNFORSCHUNGSZENTRUM KARLSRUHE

Institut für Nukleare Festkörperphysik

KfK 4562

Proposal to Make our Information System more Effective

E.L.Haase

Kernforschungszentrum Karlsruhe GmbH, Karlsruhe

Kernforschungszentrum Karlsruhe GmbH
Postfach 3640, 7500 Karlsruhe 1

ISSN 0303-4003

PROPOSAL TO MAKE OUR INFORMATION SYSTEM MORE EFFECTIVE

This paper discusses the necessity and some of the possibilities to fundamentally reform our information system. A self-improving, data bank based, closed loop information handling system is proposed. It satisfies the needs for universality and speed of access, updatability and transparancy. The drastically reduced cycle period is novel, where it can be brought to function. It is proposed to test it and the hard- and software in the more mature area of A15 materials. After necessary corrections, it is then proposed to apply it in the new area of high T_c superconducting materials, where the conventional information system has almost broken down. With minor modifications, the method is applicable to all fields of rational knowledge. The method can also be adopted to build up updatable text books by many authors, data files and transparent information files of interdisciplinary character with universal access, to quickly handle the information needs in science, technology and modern society.

VORSCHLÄGE UNSER INFORMATIONSSYSTEM EFFEKTIVER ZU GESTALTEN

In dieser Arbeit werden die Notwendigkeit und einige Möglichkeiten diskutiert, unser Informationssystem fundamental zu reformieren. Es wird ein selbstverbesserndes, datenbankgestütztes Informationssystem mit Rückkopplung vorgeschlagen. Es erfüllt die Forderungen nach Schnelligkeit und Universalität des Zugriffs, Transparenz und die Möglichkeit, die Informationen auf dem neuesten Stand zu halten. Die drastisch reduzierte Zeit für einen Informationsdurchgang ist neu, jedenfalls dort wo die Methode zum Funktionieren gebracht werden kann. Es wird vorgeschlagen, die Methode und die Hart- und Software in dem reiferen Gebiet der A15 Materialien zu testen. Nach notwendigen Korrekturen wird dann vorgeschlagen, sie auf dem Gebiet der Hoch T_c supraleitenden Materialien einzusetzen, wo das konventionelle Informationssystem fast zusammengebrochen ist. Mit kleineren Korrekturen ist die Methode auf allen Gebieten rationellen Wissens anwendbar. Sie kann auch angepaßt werden, um Textbücher vieler Autoren auf den neuesten Stand zu bringen und zu halten. Man kann damit transparente Datensammlungen mit interdisziplinärem Charakter und universellem Zugriff schaffen, um die Informationsbedürfnisse in Wissenschaft, Technik und der modernen Gesellschaft schnell zu erfüllen.

Introduction

What is the sense of scientific work ? It is not just to write proposals, many publications and annual reports. It is generally to gain knowledge and understanding. It is the task to order the information available to our senses. More specific it is to collect information, for instance by carrying out experiments, to order the multitude of results of measurements or observations and then to interpret them with the aid of models, simulations and theories. Ultimately the theories reflect and constitute our understanding of nature [1].

The key instrument to implement the handling and transmission of knowledge is an appropriate information system, that permits quick access to any specific information.

Some years ago it was pointed out that, in many areas of endeavour, the available knowledge doubles every five years. This fact however has the negative implication that an individual scientist usually can keep himself informed on a geometrically decreasing ever smaller special area. In thirty years this amounts to a factor of roughly 64. Thus there is the serious danger that the overview gets lost, without the scientists involved being even aware of it. This is a serious problem, as this narrows the mental horizon.

What is then needed are a few scientists, who take the time to gain and keep an overview over a larger area and inform the specialists with reviews and books. Their responsibility is to collect and order the information in the larger area, to point out needed new experiments, the appropriateness of models and to obtain a deeper understanding.

The author has been working in the area of superconducting materials for the past twelve years. Recently he has spent several months to sort the data on Nb_3Si and then Nb_3Ge. This has been a most frustrating experience, but has lead to a partial clarification in both areas. Only the insights gained have motivated him to pursue. There is urgent need to filter and review the published information rapidly and make it accessible to everybody who needs it (1). Without up to date review literature, it has become too laborious to inform oneself in adjacent areas of knowledge, and hence this is rarely done.

The discovery of the new high T_C superconducting materials has made it evident that the information situation in that field of solid state physics should no longer be handled just with conventional conferences and publications. Crudely estimated, there are about 10.000 papers in about two years. As much of this is parallel work and the communication of ideas and results hardly works, this has become most inefficient. In the authors opinion the information handling in that new area has almost broken down.

The analysis of this present field is difficult, but we can learn from the past. For 15 years the field of the A15 superconducting materials has been at the forefront of superconductivity, with Nb_3Ge having the then highest T_C. It is proposed to apply the proposed method to that more mature field and in doing so learn how to use the method in the new field and in other areas. The A15 class of materials has been studied using the most modern methods of analysis. Combining the information of 2-3 thousand papers does bring out the fascinating physics of that class, which so far has not been accessible generally. This will then provide the understanding needed on how to effectively proceed with the new materials and what experiments and analyses are needed. Once analysed and combined, the information on the new materials will lead to a much better understanding of their physics.

Except for the scientists who do the "filtering", this should not require more than a few days work for previous researchers in that field. If this is not done soon, the available knowledge in the minds and the notebooks of the researchers will be lost forever and the literature will stay in its present state.

So far we have primarily collected knowledge on the A15 materials and, in partial fields, it has been interpreted with models. However we have in large areas not done the essential review and "filtering" of irrelevant information. We have not done the ordering with hypotheses and there has been little in the way of theoretical interpretation.

The aim of the present paper is to start a discussion on how to improve our present information system with the aid of modern technology. The proposed system should be considered as a starting base. The release date is supposed to indicate that the conjunction of ideas presented in this paper is dynamic rather than static.

Primary Literature and Conferences

In the field of the high T_c materials lasting since two years, the literature situation is most unsatisfactory. The flood of roughly 10.000 original papers, appearing with delay of 3-9 months, prevents one to keep oneself informed in more than a small area and to cite much of the relevant literature. Many papers report on the highest or lowest value of specific physical properties. They are outdated as soon as a better value is reached. Much work is done in parallel and the communication between such groups seems to work rather poorly .

The large number of conferences and meetings provides some information exchange, but in a way, they more hinder than stimulate careful work. In the hectic of the times between conferences experiments are often not carefully planned, nor are the samples characterized by all available methods nor is there time for careful analysis. Many measurements of other physical properties are done on samples that have not been fully optimized and they will have to be repeated, when such samples finally become available. There is too little systematic work (1).

The development of a multitude of new experimental methods in solid state physics has led to a tremendous amount of detailed data on complex systems. Often very little information on physical properties can be extracted from them, as long as the simplest systems are not thoroughly understood. The challenge is to simulate such data with models and thus extract the physical properties of matter, where possible by combining information from different methods.

The FIZ Karlsruhe [2] is undertaking a pilot study collecting abstracts of preprints on high T_c materials submitted to them. They are available via keyword controlled data bank literature searches. This is an important step in the right direction, but not wholly satisfactory, as the abstracts are only incompletely covered.

No sense is seen in placing the primary literature into data banks. The sheer amount of largely unstructured information prevents an efficient use. What is needed as a first step are stored and updatable reviews with pictures and tables of smaller areas, accessible to anybody, who wants to inform himself. These reviews need to be cross referenced to other such areas, as they become existent on data banks or computer centers. The information contained in the reviews would then constitute a starting base for data files, as conceived for the A15 materials. However it is to be noted that such reviews have a different structure than the file system proposed later on.

Abstracts of A15 papers are available from data banks via keyword searches. In future papers, the authors should take care to provide the appropriate keywords both in the abstracts and as separate strings. Unless the right keywords are provided, their work will not be accessible via data bank searches. Present lists of such keyword for each area should be made available to authors.

Original papers on Nb_3Ge (about 700) and Nb_3Si (about 100) showed a large spread of data. It took months to clarify the causes for discrepancies in the T_c , lattice parameter and atomic percent correlations and other properties and to work out hypotheses for the physics underlying the observed phenomena. These hypotheses need to be tested. It is planned to provide drafts of both papers on data files for a while as starting bases, to set up data files on the physical properties of both systems.

It would be most inefficient to repeat the trial and error steps, that were done over many years in the A15 field.

Whatever the storage medium, there ist no substitute for original papers. Only the experimental details and the chains of reasoning give a feeling for the reliability of data and the degree of certainty of a given conclusion. In the secondary literature and the proposed data files, only the data and the conclusions should be stored, perhaps with an indication of their reliability. Additionally, the proposed system facilitates the input of unpublished data, updated conclusions, when more information becomes available, suggestions for interpretation, hypotheses etc. Such information should be submitted in written form or as a data file and should be documented in a paper file with name and adress of the contributor(s). It would be desirable that original papers or contributions refer to specific information in documents that they make obsolete.

Secondary literature

For the high T_C materials secondary literature is rather scarce. Such literature is essential to scientists entering the field and to gain an overview for people not actively working in the field (1). Primarily the manuscripts of invited talks in proceedings do partially serve as review articles. There are few books and they mostly consist of collections of more carefully researched original papers. Much of the information in these books will be outdated soon (1). Many review talks no longer appear in written form and are thus not documented nor available afterwards to non-participants. There is an urgent need to provide updated reviews on smaller fields like single crystals, thin films, bulk samples or wires on data banks for general access to speed up the information retrieval.

For the A15 materials, apart from three early articles [3-5], the researchers involved, with to the authors knowledge only four exceptions [6], were unwilling to undertake the ardous task. The most extensive is the Habilitationsarbeit of Flückiger [7]. There also is the review paper by Muller [6]. For the area of thin films also the review arcicles on Nb_3Ge [8] and Nb_3Si [8] will be both published and made available as data files. All four works can serve as starting information on A15 data files.

Criteria for an (Improved) Information System

Before details are considered, it is essential to list necessary and desirable properties. Placing information into data banks requires explicit criteria. Generally the criteria must conform with the scientific tradition.

1. The primary literature and contributions must be accessible within a matter of days.

2. The secondary information must be universally accessible, but not alterably, with as little time delay as possible. The following criteria refer to the secondary literature or data files.

3. The time required for one cycle of critical review should be substantially reduced.

4. The secondary information should be as compact as reasonable.

5. There must be a structure system to the files (full text data base) and there must be files, that contain the directories for ordered access to the files. The information must be structured in sections of no more than a few pages to facilitate retrieval in written form. Once the data are available in banks, computer searches can be made. Si-

milar to the abstract searches, keywords and descriptors should be provided to speed up the searches. The possibility to search some files directly can provide fast answers even to more complicated "questions".

6. The information handling and "filtering" should be as transparent as possible and should be checked and, if needed, be corrected. Written rules must be laid down for this.

7. Unlike journals and printed books, the information should be updatable and a way must be found to implement this effectively.

8. The system must handle text, tables black, and white drawings and black and white photographs. Colour is technically feasable, but requires much storage space for good resolution. The presently available text symbols are somewhat restricted, but can readily be added on. Complicated formulae can also be handled as pictures with a scanner.

9. There should be references to <u>all</u> relevant original literature and to the names and adresses of researchers who contribute unpublished information, contradictions, hypotheses with tests etc.

10. There should be cross references to related data files and banks.

11. The system must be realizable at reasonable cost, easy to use and use available hardware and makable software. This implies the use of small computers like the Atari or MacIntosh both as terminals and perhaps even as data banks.

12. The self-interest of authors and contributors must effectively cause them to submit their original papers and contributions to (the) contact person(s). Therefore it is neccessary to cite them in the data files. This also hopefully will cause them to critically think about what they are submitting. They might be asked to formulate entries to the data files and point out specific information, which their contribution makes obsolete.

13. The burden of being a contact person, reviewing the primary literature and the contributions, must be limited by distributing the work onto many shoulders. For each area this should preferable be one person, but with coordination. Theirs is the responsibility that the system works within their domain. <u>Only</u> they should be able to enter new information and update or modify existing one.

14. Updated information should be marked with a release number (correlated with date) to mark new information for fast sorting and access.

15. It would be desirable to have files of the most relevant raw data with (or without) error bars. At least this should be discussed.

16. It would be desirable to store the names of researchers active on given problems, to enable communication and too much duplication.

17. It would be desirable to have files pointing out open problems and needed experiments.

The proposed system

All of the stated criteria can be satisfied with the available human potential and with the present hard- and software. There are some good and easy to use text editing programs that can handle pictures. There are high resolution scanners to digitize black and white drawings and photographs into condensed data files and it is possible to edit these. There are open questions about an international format for integrated text, data and pictures, but this is no fundamental problem and probably can be settled shortly. References 9 and 10 discuss the international standartisation of procedures and formats. Attempts are on the way to get this implemented practically. Inexpensive and easy to use small computers like the Atari ST series (used by the author) or the Apple Macintosh or the IBM-PC can readily be hooked up to data banks or large computers linked via BITNET, EARN or other data channels. Conceivably Telefax could be used. Transmission speed is a problem.

A schematic drawing of the proposed system is shown in figure 1. The information flow is mainly from left to right. Most important components are the "filters" of incoming information to compact the knowledge to the essentials and to reject poor, duplicated, or irrelevant information. The advance of knowledge has elements of criticism, feedback, convergence and cross links, which must be implemented in any improved systems , for it to work effectively with a short time "delay", or, in the language of physics or electronics, time constant. Criticism, arising on reading the files, should be adressed to the contact persons acting as "filters", both for the initial file and the growing set of data files. The proposed feedback is conceived in analogy to physical or electronic feedback systems, which normally perform extremely satisfactorily . Feedback is neccessary to update and improve the quality and quantity of information in the files. The feedback loop is shown with a thicker line. The feedback is implemented via a closed loop with human "filter". This should preferably be one contact person for limited area, to whom also new primary literature, and written contributions and criticisms are adressed.

Feedback has existed previously, but with a time constant of years. This could be reduced to a matter of weeks. The sense of this diagram is to point out essential features of the proposed system, and to cause thought. It is not complete and in no way an overall description of the proposed system.

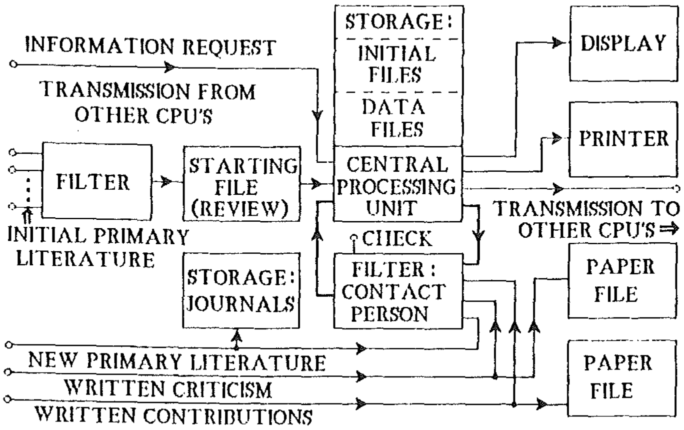

Fig. 1. Schematic drawing of one station of the proposed information system

This "filter" is the weakest link in the whole system. It should be a senior scientist with experience in the given or at least a related field. This contact person also must act rationally according to the established and laid down rules of scientific tradition. There must be as part of the system informed individuals of the scientific community, that check his performance and that can be appealed to, when an author has arguments that his contribution is not integrated correctly (2). If that person does act on other than rational grounds, there is serious danger that that section diverges like a mathematical series. This danger particularly arises, when such a method is applied to areas beyond exact science, where serious harm can be the consequence.
(c.f. stock market, Oct. 1987).

Again it is neccessary to point out the analogy to physical or electronic closed loop systems. Such systems immediately get out of control, until the "filter" (PID) is adjusted properly. Then they perform most satisfactorily, at least until serious external interference occurs. Such interferences also usually can be taken care of by "shielding".

There is no question that it is essential to pay close attention to the choice and performance of the filter and feedback "function" (1). This will again be discussed later on.

Pilot Project on Nb_3Ge, Nb_3Si and Generally the A15 Materials

It is proposed to test out the procedure on the more mature field of the A15 materials, essentially to check out the hard- and software and to gain experience with the "filtering" and the data file structure. This should take no more than a few weeks, at least until the need for additional experiments becomes clear. It would be unrealistic to wait until the method is tested, without some action in the immature field of the new high T_C materials. It would be useful to set up updatable review files in limited areas, which then can be used as starting information for data files on this subject.

The decision on how to act has to be taken by the scientific community. It is proposed to set up an international commitee of a few senior people in the field of superconductivity, who would find and appoint then (a few) contact persons (responsible for the content) for the A15 field and check, monitor and correct the structure of the emerging data files. It is proposed to start with both the Nb_3Ge and the Nb_3Si [8] systems. This offers the opportunity to develop cross references, as an understanding of these systems requires the knowledge of both.

Starting files for both systems are available and will be published for documentation shortly. They are and need not be complete, as they will be outdated very soon, if the proposal is accepted. In these drafts, the emphasis is on ideas and hypothesis, rather on completeness of data and references. These files will be purged when no longer needed.

Then, after the structure and form of content has been modified to an agreed form, it is proposed to tackle the other A15 materials. The review article by Muller [6] and the Habilitationsarbeit of Flückiger [7] are excellent starting bases. When the task is well on the way, the "filtering" and the structure should again be checked. Then we should be ready to tackle the new high T_C materials and many other problems.

The structure of the data files

These thoughts are presented to show up some of the problems that arise. At this stage, they cannot be final and they need to be corrected as the system is realized and improved. In implementing the system librarians and people experienced in setting up computer based data banks need to be consulted. The book on B1 materials by Toth [11] and the Gmelin series [12] have influenced these thoughts. Many features, that were not possible in the traditional literature, can now be done and we should implement and use these as far as they are sensible. Rather than having just files of physical properties of given systems, it is proposed to provide files giving help with the system, files providing a directory of files and an organisation diagram, files giving common figures and tables, files giving raw data, references, lists of contributors, open problems with the names of scientists working in the field, proposals for new experiments, models and theories. Keywords and descriptors should be provided to speed up computer searches. One could also provide overviews on a given subject, but that would be the second step.

Placing files into storage media does force one to give such files a structure. Data files on (magnetic) storage (disks) controlled by an computer have a hierachical structure on the upper levels and can, but need not, have one on the lower levels. For ease of access , it appears sensible to give the files the same hierachical structure, which the information to be stored has. The field of solid state physics does have a hierachical structure. A possible structure of the information and the proposed files of solid state physics are shown in figure 2 as an example. Arbitrary systems of other fields of knowledge can be substituted for the given ones.

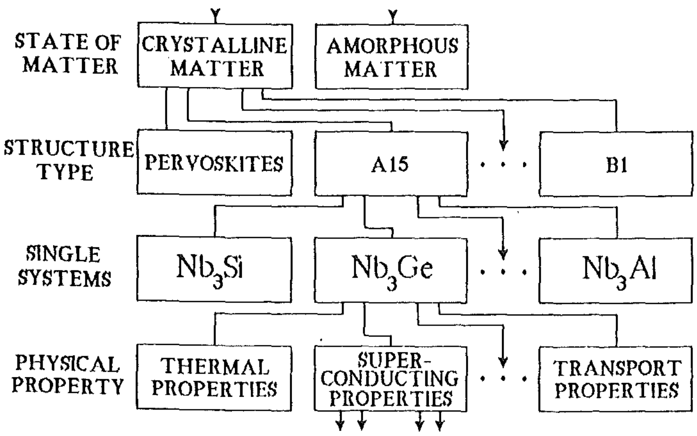

Fig. 2. Example for a hierachical physics and file structure

This diagram shows the "position" of Nb_3Ge, Nb_3Si and generally of the class of A15 crystallographic structure type. Only one hierachical sequence is indicated by lines. It is at the lowest level, where most of the information exists. Going upwards , there are common properties of the various structure types and forms of matter, but this information, storable in overview files with common tables and figures, takes much less space. The diagram does give an idea of the task ahead , if this method should be applied to other single systems in solid state physics, and then gradually going up further levels to the other states of matter, to other fields of physics, to other fields of science and to all fields of rational knowledge. Of course the huge amount of information not fitting into this classification also has to be dealt with, in separate but cross linked systems. Somehow such files have to be integrated with the numerous already existing data banks. A commercial database on high T_c materials has just been set up [13].

The structure presented is appealing from the standpoint of a physicist, as it emphazises the similarity of the systems considered. The question of structure however is of secundary significance, as the possibility to carry out computer searches permits one to sort for any desired structure.

The situation is different for a single system. There the structure is only partly hierachical. Figure 3 shows a possible organisation for an arbitrary system. The arrows give a first indication of the relationships. As the theory of a given property can be very similar for a given structure type or even state of matter, the given position is only symbolic and extents upwards in the levels.

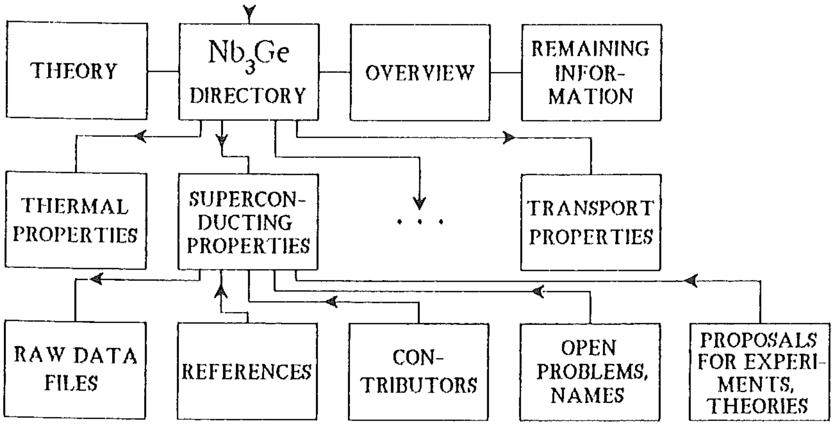

Fig. 3. Possible organisation of date files for a single system

For the actual realisation of the hard- and software there are several possibilities. An international standard both for the file structure and the file format would be the best solution, perhaps along the line of the ISO recommendations [9]. It would be desirable that this standard is identical with, or at least convertible to, the formats used in the small computers employed as terminals. The author is presently informing himself on acceptable formats. The DCA format for text is widely used and converted. The actual implementation of the method in the area of physics is probably best done by an organisation like the FIZ Karlsruhe [2] (3).

The Gmelin series [12] is presently being placed into a factual data base. It has a somewhat different structure, organised according to chemical elements. As eventually the information in that series and in the proposed system will have to grow together, attention has to be paid, that the stuctures are compatible.

It is felt that the pilot study should be carried out soon, with a sensible format, that is convertible. If it should be decided to change the data standard, this is then anytime possible with fairly simple conversion programs. We should not delay the pilot study until an international format is achieved.

Some of the ideas presented are not final. This is not serious, as an updated version of this paper will be available on storage for a while. Access and details will be found in the file 'IAK046.RATIO.HELP' on the IBM 3090 computing centre of the Kernforschungszentrum Karlsruhe.

In developing the ideas in this paper, it is amazing to observe that the structure of the existing systems determined almost every detail. There are no essentially new ideas in this paper. It is the combination of elementary knowledge of from many disciplines that constitutes the new concept. This paper including the figures was composed on an Atari MEGA ST2 computer using the STAD and SIGNUM 2 programs. This entire file has a length of 96.6 kB. Fig. 1 has 8.4 kB and a resolution of 80 dots/cm. The ASCII text file has 35.3 kB.

Summary, Comments and Conclusions

A self-improving data bank based closed loop system is proposed to handle at first the information of the community of scientists working in the field of superconductivity. With minor modifications it is applicable to all fields of rational knowledge. To work effectively, it relies on the self-interest of authors to have the essentials of their work integrated into the collective information files. It also depends on the willingness of more experienced and rational individuals in the relevant field to act as contact persons , who carry out the responsible task of information filtering and ordering. By sharing this work load between many shoulders, this task becomes managable. They also will be rewarded with an overview of the original literature and the contributions, and should be able to make suggestions for further experiments and improvements. All of this should improve the quality of physics in the areas discused.

The scientific community should motivate, recognize and perhaps honor the contact persons , as the performance of the proposed system crucially depends on them. They could be asked for instance to give invited talks to share the insights they gain. Reimbursing their costs to attend such meetings would be one possibility {3}.

Additionally or alternately, paying the contact persons, like the authors of the Gmelin series, will motivate, if the amount is adequate. That might not apply to all experienced scientists {3}. As the amount would only be a negligible fraction of the expenditure for research, it could do much to make research more efficient. If the scientific community does want such an information system, it will have to pay the price for it in honor or in money.

The system does satisfy the criteria of transparency, universality and speed of access and updatability; that is the possibility to grow and be modified as new data, new interpretations and criticism arise. The transparency and ease of access will do much to facilitate deeper understanding by combining various pieces of information and interpreting them with the aid of models and simulations.

The method can also be adopted to build up updatable text books by many authors, data files and information files of interdisciplinary character to quickly handle the information needs in science, technology and modern society.

Acknowledgements

The author wants to thank O. Meyer for the tolerance to let him pursue what was felt to be important. He is grateful to A. Hoss for discussions and help with the computer drawings. He acknowledges discussions on the technical aspects with H. Hanak, J. Thilges and M. Stenzel.

References

1. Planck, M.: Sinn und Grenzen der exakten Wissenschaft. Leipzig. J.A. Barth Verlag: 1942
   Planck, M.: Scientific Autobiography and other Papers. p.80. New York. Philosophical Library 1949
2. Fachinformationszentrum Karlsruhe, D-7500 Karlsruhe. HTC-File. W.Lück . Telefax: 07247/ 824639. This project was initiated by M. Dietrich and supported by W. Klose.
3. Dew-Hughes, D.: Cryogenics **15**, 435 (1975)
4. Izyumov, Yu.A., Kurmaev, Z. Z.: Sov.Phys. Usp.**17**, 356 (1975)
5. Izyumov, Yu.A., Kurmaev, Z. Z.: Sov. Phys. Usp.**19**, 26 (1976)
6. Muller, J.: Reports on Progress in Physics **43**, 641 (1980)
7. Flückiger, R.: Atomic Ordering, Phase Stability and Superconductivity in Bulk and Filamentary A15 Type Compounds. Habilitationsarbeit.
   Physics Institute of the University of Geneva 1986.
   Also KfK Report 4204. Kernforschungszentrum Karlsruhe 1987
8. Haase, E.L. and Khubeis, I., to be published
9. ISO Standard : ISO 8613: Part 2: Office Document Architecture (ODA),
   Part 4 : Office Document Interchange Format. International Organisation for Standardization, Geneva 1989
   Also: NJE (Network Job Entry) Format for BITNET (IBM)
10. Flasche, U., Scheller, A.: Comput. and Graphics **10**, 119 (1986)
11. Toth, L.E.: Transition Metal Carbides and Nitrides.
    Refractory Materials. 7, Margrave, J.L.: ed.: Academic Press . New York 1971
12. Gmelins Handbuch der anorganischen Chemie . Nb [B1] 389, and other volumes, Verlag Chemie. Weinheim 1970
13. Superconductivity World Database, Arbitech Communications Inc., 4010 Moorpark Avenue, San Jose, CA 95117, USA

Contributors

(1) Haase, J.: Mainzerstr. 26, D-7514 Eggenstein-Leopoldshafen 2, FRG
(2) Schommers, W,: INFP. Kernforschungszentrum Karlsruhe, D-7500 Karlsruhe, FRG
(3) Buckel, W. : Physikalisches Institut, Universität Karlsruhe, D-7500 Karlsruhe, FRG