This is a preprint of a paper intended for publication in a journal or proceedings. Since changes may be made before publication, this preprint is made available with the understanding that it will not be cited or reproduced without the permission of the author.

UCRL - 74884
PREPRINT

Conf-730719-2

LAWRENCE LIVERMORE LABORATORY
University of California/Livermore, California

DISCIPLINE INDEPENDENT ASPECTS OF
DATA BANK ORGANIZATION

Robert J. Howerton

July 17, 1973

NOTICE

This report was prepared as an account of work sponsored by the United States Government. Neither the United States nor the United States Atomic Energy Commission, nor any of their employees, nor any of their contractors, subcontractors, or their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness or usefulness of any information, apparatus, product or process disclosed, or represents that its use would not infringe privately owned rights.

This paper was prepared for presentation at
The CODATA Symposium on Man-Machine Communication
for Scientific Data Handling
Freiburg, Federal Republic of Germany
July 22-27, 1973

MASTER

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

DISCIPLINE INDEPENDENT ASPECTS OF

DATA BANK ORGANIZATION^{*}

Robert J. Howerton

Lawrence Livermore Laboratory, University of California
Livermore, California 94550

Before discussing any characteristics of data bank organization, I would like to establish the frame of reference with respect to which I shall be directing these remarks. A data bank as I will be referring to it will usually be associated with the experimental numerical data of some discipline together with an annotated bibliography from which the numerical data were extracted. I shall assume that these data are retrievable from the bank according to user specification. There will be certain variables and certain parameters characteristic of the discipline such that a customer or user of the data bank can specify ranges of any of these variables or parameters and retrieve exactly the data he wishes to have, arranged by an order he specifies either in hard copy or machineable form and including the pertinent bibliographic data.

This definition of a data bank can be contrasted with simpler forms of collections in which the data associated with a reference are stored together with the bibliographic information and the only possible form of retrieval is by data set arranged in an a priori defined format, or a data bank may be simply an annotated bibliography. Still another form

^{*}This work was performed under the auspices of the U. S. Atomic Energy Commission.

is a collection of evaluated data for which no explicit bibliography exists.

One of the major uses of a bank of experimental data is to provide input information for the creation of evaluated data sets. For many scientific disciplines there are associated engineering disciplines that require sets of numerical data for computer calculations that are complete in the sense of spanning specified ranges for specific independent variables and also in the sense of including a prespecified set of properties of materials. If all such properties have been determined by measurement over the required range of the proper independent variables, a qualified scientist can then produce a set of evaluated data for the material being considered. Usually measurements have not been made for all properties in all ranges of the independent variable and the scientist performing the evaluation must choose between conflicting experiments, fill in gaps using established theory if a theory exists and finally resort to systematics and educated guessing. The end product of the evaluation activity, when collected together with other evaluations can be called an evaluated data bank. Once again, the organization of such a collection is not the main subject of my discussion. The thrust of the remarks I will be making is directed toward the organization of banks of experimental data. I shall be delighted to discuss the methodology associated with evaluated data but time precludes making that discussion now. I would only comment that, in general, the evaluated and experimental data should not be mixed together in a single collection.

There are certain common factors in the process of organizing scientific information that are independent of the discipline or sub-discipline with which the information is associated. For the purpose of this discussion I shall define scientific information as being the totality of facts belonging to unambiguously defined categories and to which logical order can be assigned. The logical order may be any reproducible method which is characterized by the absence of direct recourse to intuition. An example is an arrangement by intrinsic numeric value associated with a hierarchical scheme. Another example in an arrangement in simple alphabetic order.

The facts appropriate to experimental scientific information can be divided into two categories. First, is the bibliographic information and second, the numeric data resulting from measurement. For ease in handling and efficiency of computer usage, these two facets of scientific information should be handled separately in a computer but linked by an appropriate identifier, usually a reference accession number. With such a linkage the two files of information can not only be dealt with quite separately but can also be recombined at will in whatever arrangement is deemed desirable, either for further computer manipulation or for producing hard copy displays. The reference accession number can probably best be assigned to the bibliographic information in strict order of the integers. There is a distinct advantage to including the year of the reference. A combination of year and accession number, I shall refer to as the reference number. Fairly often, one

wishes to retrieve from the data file only measurements made since a specific year. While the bibliographic information contains the date of the reference, specification of the year with the data is a desirable redundancy since the year measured then becomes just one of the fields for retrieval specification.

Let us now consider the bibliographic aspects of a data file. It is probably a sine qua non of the bibliography that it be annotated in a free form. That is, the scientist who enters a reference as part of the bibliography associated with a data bank wishes to enter certain comments about the experiment reported in that reference. We have experimented with complete formalization of bibliographic entry by encoding journals or report series, method used, institution, etc. Our experience has been that a much lesser degree of formalization is more useful in the long run. The reason is that the encoding is more susceptible to clerical error than is the straightforward entry of accepted journal designation in the case of journal entry. Within our experience we have found that, while formalization of methods, standards and institutions is attractive intellectually, we have never had a request for retrieval of "all experiments using method A" or "all reports of work from Argonne National Laboratory." In addition to the straightforward method of entering the bibliographic data in the form one would find it in a journal reference, free text comments and see-also references both properly flagged can be entered with a minimum of effort on the part of the scientist. The flags we use are extremely simple: 1 is for the bibliographic information; 2 is for free text comments; and,

3 is for see-also references. Aside from considerations of possible clerical errors, etc., the method I have just described has the distinct advantage of being completely discipline independent. An early version of our code for producing an annotated bibliography has been implemented at three different institutions and for, at least, five different disciplines. This has saved a considerable amount of programming time. Our current codes can update our bibliography of 4187 primary and secondary references and produce three listings, one by accession number, one by reference and an author index in about 30 seconds of computer time on a CDC 7600. Let me hasten to add that the more highly structured annotated bibliography can likewise be converted from one discipline to another by simple re-assignment of the meaning of the encoding system. Thus, the first two common aspects of data bank organization are the production of an annotated bibliography and the linking of the bibliography to the data by a reference number.

One of the other papers to be given at this symposium deals with hierarchical systems for classifying and storing numerical data. It is certainly a common aspect of numerical data bank organization that some type of hierarchical system should be developed for the discipline being considered and I will only make the point that most disciplines are self-organizing to a greater or lesser extent. I shall say nothing about how data should be organized for any specific field nor anything about the methods of extraction of data from the references in which they appear. Let us just assume that for the data appropriate to some

field, a method of organization exists and that these data have found their way into the bowels of a computer. Likewise let us assume that the properly encoded information about the reference is available in mechanized form. The main question which arises is what should be done with these data once they are available. Much and perhaps too much has been written about "real time" or "on-line" availability of such information. Obviously, in those cases in which a computer is used to control other activities such as the re-entry of a space vehicle, there is need for "real time" response. It is clear however that the popularization of the "real time" concept has lead to over-extrapolation of its virtues.

Rather than talking about the data or references of a specific field of science, please allow me to illustrate my point with sets of data associated with common experience. Probably the best existing data retrieval system in the United States is that of the telephone system. Within about one minute one can obtain the telephone number of any individual in the country. This is not done by the interrogation of a computer, but rather by consulting, through a formalized process, an appropriate person who has a list of telephone numbers for specific geographic area. This leads then to the next principle of data bank organization. Specialized lists or graphical displays organized in accordance with the immediate or anticipated needs of the requestor should be made available. For brevity, I shall refer to this aspect as the "Telephone Book" method. This principle is efficient for an information analysis center because all the input is generated at the

center. This is to be contrasted with the on-line interaction needs of an airlines reservation system because the input for such a system comes from many stations. I can, however, think of a situation in the geo-physical area in which on-line warning, if not interrogation, would be desired. This is for early warning of an impending Tsunami wave following an earthquake. Such a situation is, however, not really in the subject matter of this symposium.

The next common factor of data bank organization is the production of indexes both to the data and to the bibliographic information. I have discussed previously the three types of indexes found to be useful for the bibliographic area. How many indexes should be produced for the measured data is strictly a function of the discipline and the adopted hierarchy for that discipline. In my own area there are 15 defined fields within the description of a datum. It would, therefore, be possible to create any number up to 15 factorial permuted indexes to the data. We have found that four indexes with specified permutations of the data fields are completely adequate. I shall add that without these indexes in telephone book form, I would not be able to answer about 70% of the telephone requests I receive. Another 20% are for specific numerical data for which I must consult my telephone book form of the data file and the final 10% are for specific retrievals where the retrieval criteria can be for any range of values of the variables in the 15 fields I have mentioned. It is only for this latter 10% that we have need for recourse to a computer retrieval.

The next common aspect of data bank organization has to do with the delineation of the field for which the data are to be organized. One must, of course, define his method of encoding data to be appropriate to the data he envisions to be stored in the bank. Quite often the field with which he starts is closely related to another field or fields. It is clearly a mistake to try to organize a hierarchical system to encompass all knowledge. On the other hand, it can be equally disastrous to define an area or discipline such that other closely related areas cannot be joined to the area being considered. Thus one might say that the method of organization should be hospitable to expansion. The practical point I would make is that the default option should never be used. I shall use, as an example a field with which I have some familiarity. I apologize for particularizing to a field with which some of you are not familiar. The general field of experimental nuclear physics is composed of two main parts: first is nuclear structure physics and second is nuclear reaction physics. Now, if a system is to be created for neutron reaction physics one might not assign a designation for the incident particle because it is always a neutron. If the system is to be extended for other incident particles, they can be assigned non-zero designations and all is still fine with an implicit zero designating an incident neutron. But if it later becomes desirable to include nuclear structure information for which there is no incident particle, a problem arises. The designer of the system should have provided a non-zero value (non-default option) for the incident neutron ab initio.

The final point I would make is one which needed to be made 10 or 12 years ago but which probably does not need to be made now except in the sense of completeness of this discussion. There is quite obviously a great difference between the form of the information that is put into the computer and that which is produced by the computer for people to peruse. Most input is encoded using a numerical equivalent to the property being described. Because the user of the output is not, in general, familiar with the encoding system, the output must be couched in the terminology of the discipline whether in the form of printouts or graphical displays. In the latter case axes should be labelled in the appropriate language with the appropriate terms and with appropriate units. In short, output from the system should be designed for the convenience of the user, not the convenience of the computer.

In summary, I have tried to make the following points:

1. Numerical data and bibliographic reference data should be linked by a reference accession number.
2. Most disciplines are to a greater or lesser degree self-organizing.
3. Telephone book types of listings or graphical displays should be produced routinely rather than resorting to on-line interaction.
4. Selected permuted indexes to both the bibliographic information and the numerical data files should be produced routinely.

5. Default options should be avoided when designing an encoding system.
6. Output modes should be designed to be convenient to the user.

REFERENCES

1. D.E. Cullen, R.J. Howerton and S.T. Perkins, "An Integrated System for Production of Neutronics and Photonics Calculational Constants, Vol. 1 (Rev. 2), ECSIL, A System for Storage, Retrieval, and Display of Experimental Neutron Data," Lawrence Livermore Laboratory, UCRL-50400 (in press).
2. R. J. Howerton, S.T. Perkins, C.P. Altamirano, K.L. Hill, "An Integrated System for Production of Neutronics and Photonics Calculational Constants, Vol. 2, A Bibliography of the Experimental Data of Neutron-Induced Reactions," Lawrence Livermore Laboratory, UCRL-50400 (1970).
3. R.J. Howerton, S.T. Perkins, C.P. Altamirano, K.L. Hill, "An Integrated System for Production of Neutronics and Photonics Calculational Constants, Vol. 3, An Index to the Experimental Data of Neutron-Induced Reactions," Lawrence Livermore Laboratory, UCRL-50400 (1970).
4. R.J. Howerton, R.J. Doyas, T.C. Michels, S.T. Perkins, "An Integrated System for Production of Neutronics and Photonics Calculational Constants, Vol. 4, Evaluated Nuclear Cross Section Library," Lawrence Livermore Laboratory, UCRL-50400 (1971).
5. R.J. Doyas, T.C. Michels, S.T. Perkins, R.J. Howerton, "An Integrated System for Production of Neutronics and Photonics Calculational Constants, Vol. 5, CLYDE, a Code for the Production of Calculational Constants from Nuclear Data," Lawrence Livermore Laboratory, UCRL-50400 (1971).
6. E.F. Plechaty and J.R. Terrall, "An Integrated System for Production of Neutronics and Photonics Calculational Constants, Vol. 6, Photon Cross Sections 1 keV to 100 MeV," Lawrence Livermore Laboratory, UCRL-50400 (1968).

7. R. J. Howerton, "An Integrated System for Production of Neutronics and Photonics Calculational Constants, Vol. 9, Thresholds of Nuclear Reactions Induced by Neutrons, Photons, Protons, Deuterons, Tritons, and Alpha Particles," Lawrence Livermore Laboratory, UCRL-50400 (1970).
8. S.T. Perkins and R.J. Howerton, "An Integrated System for Production of Neutronics and Photonics Calculational Constants, Vol. 10, Tabulated Experimental Data for Neutron-Induced Reactions," Lawrence Livermore Laboratory, UCRL-50400 (1971).
9. M. Gyulassy, R.J. Howerton, S.T. Perkins, "An Integrated System for Production of Neutronics and Photonics Calculational Constants, Vol. 11, Experimental Data, Indexes, and Techniques of Obtaining a Selected Set of Neutron Resonance Parameters," Lawrence Livermore Laboratory, UCRL-50400 (1972).
10. S.T. Perkins and M. Gyulassy, "An Integrated System for Production of Neutronics and Photonics Calculational Constants, Vol. 12, An Atlas of Resolved Neutron Resonance Parameters," Lawrence Livermore Laboratory, UCRL-50400 (1972).
11. M. Gyulassy and S.T. Perkins, "An Integrated System for Production of Neutronics and Photonics Calculational Constants, Vol. 13, An Atlas of Unresolved Neutron Resonance Parameters," Lawrence Livermore Laboratory, UCRL-50400 (1972).
12. T.F. Budinger, "Clinical and Research Quantitative Nuclear Medicine System," pp. 501-555, Medical Radioisotopes Scintigraphy 1972, International Atomic Energy Agency, Vienna (1973).