26 October 1956, Volume 124, Number 3226 SCIENCE

Data Processing for Scientists

Karl F. Heumann

It is the thesis of this article that scientists can expect, and indeed can demand, more assistance from machines in the matter of scientific data handling (*1*). This will be available through the use of techniques which have had a recent, rapid development in business and accounting applications.

Two of these techniques are considered here in their relation to science: integrated data processing (IDP) and electronic data processing (EDP). My plan is first to provide a short background in the concepts of IDP and EDP through a consideration of the reasons for their development in business. This is followed by several case-histories of successful applications and by some information on the machines involved. Finally, I want to connect this to the related problem of scientific information storage and use by discussing current work in classification and indexing, closing with some recent work on computers that points toward increased use in the future.

The scope of the discussion is limited for a number of reasons. There has been a flood of books and articles on this subject which have, by the way, not been adequately handled by bibliographic means. The only extensive bibliography (*2*), for instance, has neither a subject nor an author arrangement. Preparation of an authoritative bibliography on data processing containing adequate indexes would be a timely venture.

Another limitation of the subject matter must be understood. No mention is made here of operations research, linear or other programming, game theory, information theory, and other new disciplines that touch the matter at some point. The use of these disciplines is mentioned in several of the books noted in "References and Notes."

Dr. Heumann is on the staff of the Chemical Abstracts Service, Ohio State University, Columbus.

Definition and Origin of IDP

Early in 1954, the American Management Association held a conference on "Integrating the office for electronics," which described the pioneer work of the U. S. Steel Corporation in office mechanization. Apparently, it was here that the phrase "integrated data processing" was first publicly used. A published report (*3*) of the conference gives a statement about the elements of IDP. They are two: (i) Creation of a document is accompanied by recording of the data in mechanical form. (ii) Subsequent processing of the data is done mechanically in an integrated system.

Closely connected with this original statement is the concept of a "common language" (*3*, p. 9). This refers to a means by which IDP can be carried out, in this case a group of machines that can interchange data through the use of the same impulse-code for each symbol involved. The U.S. Steel program decided on the five-channel punched paper tape (Fig. 1), already used with telecommunications equipment, as the carrier for this common language, and it has been widely adopted. More is said on this later.

Unfortunately for a scientific audience, *data* here is used to mean almost any business information whatever, such as that involved in accounting, sales, or payroll handling. We shall come back to a more familiar use of it later on. *Processing* refers to any subsequent use of the information or "data," and in this context it is used to mean processing by way of some machine.

Electronic data processing (*4*, *5*), by contrast, does not have a sharp beginning but developed through the use of large-scale computers, in experimental trials. It is evident that the two concepts merge where any electronic components are used in an integrated system. The advantages, of course, are the speed, flexibility, and capacity of such devices. It, nevertheless, remains true that it is possible to use an electronic data processor without having an integrated system.

It seems reasonable to refer to work in this area as *data processing*, and to reserve *automation* for the handling of materials by mechanical control, although the terms have been used interchangeably (*6*).

Let us consider why an industry would scrap long-used office procedures for a method that is often initially of little economic advantage and raises important management and personnel problems. I believe the answer is in the future value that a company can hope to gain by having its paperwork under tight control, with a concurrent reduction in its clerical force. The implications that this rather trite statement has for science and its problems should concern us all. Do we, at present, have good bibliographic control over the documents that we produce? Consider government research reports or atomic energy literature. Are we committed to a situation in which an ever-larger number of our dwindling supply of scientists act as "clerical workers," through failure to apply scientific methods to science's own record? The example of another area of human activity—namely, business—can show us a pertinent model here for recasting our own procedures.

Business Applications

What are some of the applications that have proved successful?

Suppose we look at the pioneer effort, that of U.S. Steel (*3*). This company found that its old procedures for processing of data about its procurement, manufacture, and sale of products were falling behind the requirements of management for prompt information. The redesign of parts of the system would not remedy this situation. Partial mechaniza-

tion led only to an increase in repetitive keyboard entries of the same data.

Fortunately the suppliers of office equipment were receptive to a new approach, and in the past few years they have brought out a variety of machines, such as adding machines, electric typewriters, printers, and bookkeeping machines, which either accept or create the five-channel common-language tape. As an example, this has enabled the U.S. Steel Corporation to prepare, for each Monday morning, an accurate inventory, as of the previous Friday afternoon, of up to 200 models in each of 130 different locations, with a variety of data on each individual model. This, of course, is a single link in a highly integrated chain of procedures.

Another instructive example exists in the program now used by the Aluminum Company of America for handling orders (*7*). An integrated system takes order tapes, which have been prepared in more than 60 sales offices, and transmits them through a teletypewriter switching center at the rate of 1000 per day to 24 different plants. The five-channel punched paper tape again carries the common language. Initial recording on a tape-cutting typewriter starts a series of repetitions which ultimately produces duplicates for proof copies, sales orders, salesman's copy, production planning copies, and even shipping labels, besides numerous file copies.

Such tight control of the information flow was achieved by intensive study and redesign of forms and procedures, once the concept of integration was accepted. Several years of work by many people were required to reach the point where successful operation could begin.

A third system, planned with utilization of an electronic data-processing machine in mind, has been developed by the Chesapeake and Ohio Railroad (*8*, pp. 74–122). It is referred to as "The one-shot process," which symbolizes the goal of a single keyboard entry of information.

In 1954, the C. and O. used more than a billion pieces of paper, chief among them being millions of freight waybills. The information on this form was minutely analyzed, by a special methods research team, with regard to the source and future use of each individual entry. It became apparent that extensive duplication existed in later retyping of the same information. In combination with other forms, such as car orders and wheel reports, it was possible to plan an integrated system, using the five-channel tape, a teletypewriter network, and a large-scale, general-purpose computer. The input was decentralized so as to include any place where data originated, and the communications network carried it to the central computer. The output goals were as follows: (i) current digested results for management decisions, (ii) exceptions for investigation, and (iii) detailed listings for reference.

Fig. 2. The Programatic Flexowriter, an important part of many IDP systems.

A company official, E. L. Morrison, has said (*8*, p. 96): "The ultimate objective of the communications network is the provision, eventually, of a major input for the computer. The design and development work that has gone into the network, the format in which the information is placed, the distribution of that information and the necessary machine coding have been developed to be compatible with this computer use. Thus we have achieved a translation of the information from clerical documents to machine language, performing it by key stroke, in all the varied locations over the railroad. In effect, we have decentralized the key-punching function. The teletypewriter network brings the information all together, in form for processing in the computer.

"The computer can produce, much more quickly and economically than existing methods, many of our current operating reports for management. These reports, by and large, were developed as being the most refined tools that could be produced under a completely manual- and, in some cases, a punched-card-application. But they lack the sophistication which appears necessary for modern management in a progressive industry. We look forward to computer ouptuts contributing to a more informed management judgment in a greater variety of areas, and at an earlier date after the fact, than has ever been previously possible."

Many other, and equally successful, accounting and clerical applications of this general approach already exist.

Some Implications for Science

I have given these three examples because I think they have lessons for the use of similar methods in scientific information handling, lessons for groups such as all the science departments of a university, the research department of a business, or a government laboratory. These places now produce scientific information and are paying for its handling in one way or another.

If such a procedure were considered, one could look forward to steps such as these: (i) There will be a long period of preparation, including a job of getting the concept across to some person or group that can say "go ahead," and support will then be provided. (ii) Forms and procedures will probably require extensive revision or replacement. Incidentally, despite its legal value, what could be more archaic than the handwritten recording of experimental results in the standard laboratory notebook? (iii) It will be of maximum value if the system is completely integrated, with some specific end-use in mind. Fortunately, machines for this goal are already well developed. (iv) Not only will the output be rapid, and of improved accuracy, but it will be possible to provide answers to generic questions not hitherto possible.

Take the man who wishes to test 100 chemical compounds for their effects on 12 species of plants. With replications and a variety of dosages, he can expect to deal with several thousand results of even one technique of application. At present he might report only the successful compounds, but even if a journal were to print all the data, they would usually still be considered as a single bibliographic unit, a "paper." Here is where the item-by-item technique of electronic data processing shows its advantage. The

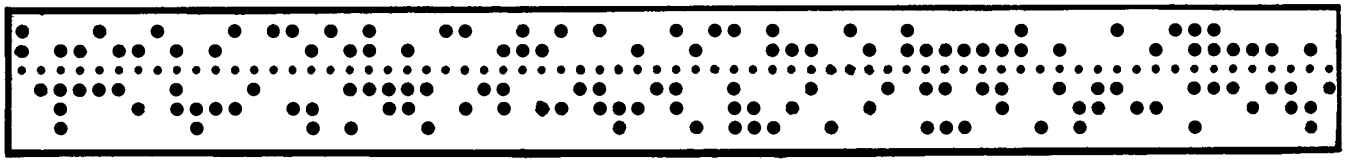

Fig. 1. Five-channel punched paper tape, the "common-language" medium in most integrated data processing. A sprocket channel is near the center.

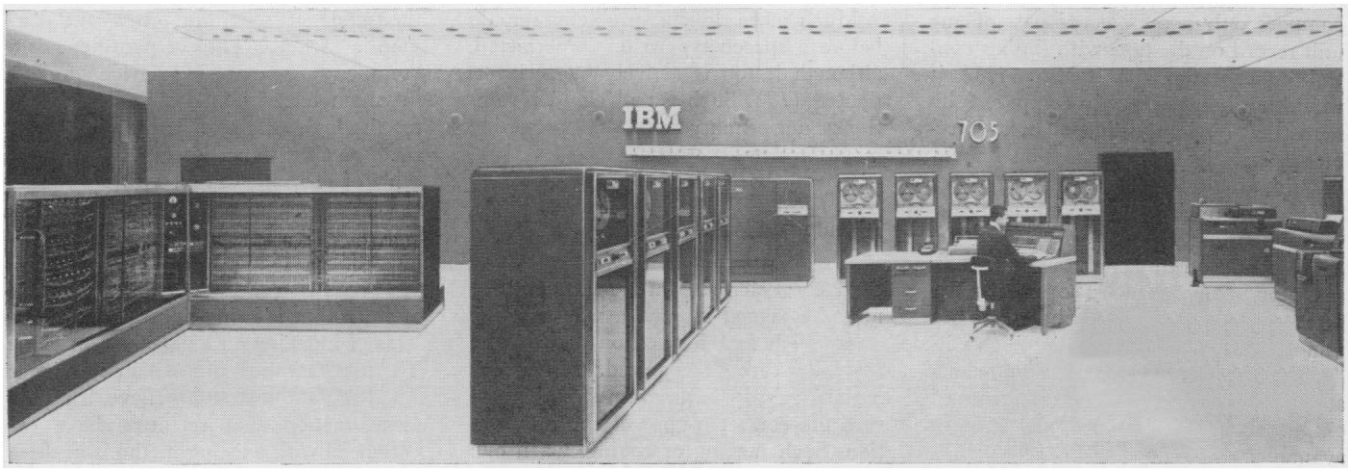

Fig. 3. The IBM type 705 EDPM, a large-scale general-purpose data processor.

complete set of results can be scanned rapidly, arranged, reordered, and analyzed for correlations.

Similar problems in data handling exist in a series of wind-tunnel experiments where the number of variables is large and the speed of analysis could be critical (see later).

I believe any scientist of even moderate acquaintance knows of caches of unused scientific data that would be raw material for such a procedure as the afore-described one. If the Chesapeake and Ohio Railroad considers itself a statistical factory (for the Interstate Commerce Commission), it would not be unrealistic to consider a research laboratory to be an information factory and to plan accordingly.

Some Machines for IDP

It will not be possible to note all the important mechanical devices that are of interest in connection with data processing. Fortunately, several recent books enable one to survey these developments in detail (*4, 9–11*).

In considering integrated data processing, the original keyboard step is generally required to create two copies. One copy is a "graphic" form in which the result of keystrokes is readable—that is, visible—but which may consist of any symbols whatever, for instance, chemical formulas on a sheet of paper. The other copy, in "coded" form, is the same information in some common language, perhaps the five-channel punched-paper tape (Fig. 1), ordinarily used for further machine processing of the information. Each transverse row of holes in such a tape contains in coded form one symbol, such as the letter *R*, or an instruction to the machine.

A widely used input device is the Flexowriter (*12*), an electric typewriter equipped with a tape punch and reader, and the capacity for reproducing tape (Fig. 2). Operation of the machine produces a typewritten sheet of paper (the graphic form) and a by-product tape (the coded form) containing all, or a selected part, of the information. Punching is almost completely automatic. Errors are easily signaled and controlled.

It is also possible to operate a typewriter and punch a standard tabulating card at the same time, or to type and at the same time enter the information onto magnetic tape. Oddly enough, little has been done with cards that carry magnetic spots or strips. Recently, checks have been coded and sorted magnetically (*13*), and there have been several other experiments (*8*, p. 135; *14*).

An ingenious variant is the punching of a standard tabulating card with the usual tabulating punches and also with the five-channel punching along one or several edges (*9*, pp. 54a and 54b). The final card not only carries a code by which it can be sorted but also contains in coded form a means of reproducing text.

Machines also exist which can interconvert between cards and tape. In fact, it is probably safe to say that any standard keyboard device can now also be found in a "common-language" version or can easily be turned into one. With the development of higher-capacity tapes (six-, seven-, and eight-channel forms already exist) almost any device that operates from or into a keyboard could be assimilated into an IDP system.

Larger-Scale EDP Machines

Kozmetsky and Kircher point out (*10*, p. 94) that the significant electronic counterpart to IDP machines is just now beginning to emerge. First came the true computer, the "scientific" computer, which required only moderate input and output but computed at high rates of speed. For applications such as payroll or accounting procedures, the second type, a "general-purpose" computer was developed which could handle large volumes of input and output, in addition to altering each unit record in some way, on usual runs. The third type, and latest to arrive, is the one that seems destined to be most useful for scientific data processing. This "inventory" computer would also be expected to deal with large volumes of input and output data and to perform some logical operations, but in any one run it would refer to only a small part of all the items in the file. These characteristics fit in well with the usual procedures for storage and retrieval of scientific information.

It will suffice here if we consider a large-scale computer and a medium-size computer for electronic data processing, with the understanding that these are representative of groups of similar devices (*4, 9–11*).

IBM Type 705 EDPM

The type 705 electronic data-processing machine is a high-speed, general-purpose device of advanced design (*15*) (Fig. 3). The principal processing and input medium is magnetic oxide-coated plastic tape. A tape reel 2400 feet long, equivalent to from 25,000 to 50,000 punched cards, depending on arrangement, can be read into the system in about 6.5 minutes. This magnetic input tape can be prepared from punched cards or from punched-paper tape, with information in alphabetic, numeric, or symbolic characters. Internally the system utilizes magnetic core memory and, optionally, magnetic drum storage for high-speed access and processing. Adequate logical operations are provided, and programming can supply almost any imaginable complexity.

A special device makes it necessary to

compare only a preselected part of any unit record on the tape with similar control data in the 705 storage. Thus it would not be necessary to compare fully items that obviously did not fit requirements. In scientific information searching, this would increase the speed of the operation, provided that coding and classification or indexing were properly done. More is said on this later.

Output onto magnetic tape is at the same rate as input, 15,000 characters per minute. A device is available which will "print" data at the rate of 1000 60-character lines per minute. Other forms of output, such as punched cards, are also available.

It is obvious that this system has the characteristics that were specified for search and retrieval devices. It is an expensive data processor, but some measure of its value in business applications can be learned from the fact that more than 150 of these systems were on order at mid-year, 1956, and a score had been delivered.

UNIVAC File-Computer

A medium-size, general-purpose computer that is also of potential interest in scientific data processing is the UNIVAC File-Computer (*16*). This system has a more flexible input, in that it will accept data directly from electric typewriters and other key-actuated machines in addition to magnetic tape, perforated paper tape, or punched cards. Input speeds will vary accordingly, but magnetic tape is still the fastest input, about 6000 characters per second.

All internal storage is on a variety of magnetic drums, assuring fast access time, given proper coding. Two methods of programming are available: stored program and external panel control. Output may be in the form of magnetic or perforated tapes, punched cards, or various printers. Although this machine is slower and smaller than the 705 and is correspondingly cheaper, it does possess the desiderata for literature searching that are mentioned in foregoing paragraphs.

These and other devices are in existence now and could be used today for the purposes I have discussed. It seems likely that they will increase in speed and capacity as new models appear, in a way similar to past development. But what of the future for new types?

A machine for reading printed matter directly off the page is currently under development in several laboratories (*17*). This would enable us to deal efficiently with the vast printed record of science by eliminating more keyboard work before printed material was ready for data processing.

Much more difficult problems remain before a speech typewriter is perfected, although it too is actively being investigated (*18*). Such a tool could revolutionize our primary methods of preparing scientific data.

Indexing and Classification

Inventory entries on a magnetic tape must be tagged in such a way that the ones which match question data in the computer are "recognized." Once the matching is done, program steps can assure the correct output of the information. Such tagging or coding, when one is dealing with scientific information, immediately raises questions of classification and indexing. These, properly, are not a part of IDP or EDP, but it may be well to indicate some current work that seems likely to be of value.

Scientists in this country have not, in general, been satisfied with large-scale classification or marshaling schemes, such as the Universal Decimal Classification or the Library of Congress Classification. Instead, efforts to extend bibliographic control to scientific materials have largely turned to indexing techniques.

The main limitation of the latter is the fixed array imposed by the alphabet on an index. The white pages of a telephone directory are a good example. Another limitation, that of space, ordinarily precludes all permutations of a multiple subject heading from being entered. For instance, "Electromagnetic-waves-propagation-equations" can be arranged in 63 other orders, many of them significant, and in the majority of lists only one permutation would be used.

A technique for dealing with these problems has been proposed and used as the Uniterm system of coordinate indexing (*19*). Briefly, the method consists in assigning to each document or unit record a unique serial number and characterizing the contents of the document by means of words, the uniterms. The document number is entered on each card bearing one of the uniterms; in our example the number for a document would be entered on four separate cards. To select documents dealing with these four topics, it would be necessary only to scan the four cards for common document numbers. In actual installations, all this matching is now done by hand, I believe, but note that the technique is one that would lend itself easily to machine processing. As such it should be investigated for scientific applications of EDP.

Another system, which starts from different premises, is largely the work of J. W. Perry and his associates (*20*). His concern begins with the problems of language, and he has minutely analyzed thousands of scientific terms into "semantic factors." Thus, a pyrometer is a "temperature-measuring device," as is a thermometer. This analysis and a concern for the logic of classes led him more directly to a consideration of machine handling of scientific data. His final product in the system, an encoded, telegraphic-style abstract, becomes an item particularly amenable to machine manipulation.

These two examples are representative of work that is being done on the problems of indexing and classification (*21*). The fact that most of it occurs in science and technology stems, I believe, from the dissatisfaction that scientists have experienced with existing methods in these areas.

Some Related Current Work

I know of only one integrated system at present that utilizes the approach under discussion in dealing with scientific data. It exists at the Lewis Flight Propulsion Laboratory, Cleveland, Ohio (*22*). A wind tunnel is fitted with instruments which take readings of such parameters as speed, thrust, temperature, and fuel flow. The digital encoder translates these signals to binary code (in which the number 2 is the base) and records them on magnetic tape. From this the data can be reduced to usable form immediately or printed out for future study and reference.

With the large number of EDP machines going to industry for accounting work, it should not be long before some company with a research department finds that research reports can be "written" on the computer in a matter of minutes, and that, at the same time, storage of the data for future interrogation can be accomplished.

There are already a few other pioneer examples of computer uses near enough to our subject to be instructive. A book containing 1 million random digits has recently appeared (*23*). This was "composed" by a computer at the Rand Corporation specifically programmed to generate random numbers.

A series of 20,810 self-demarcating code words has been generated on an electronic data processor (*24*). These words, from *BAB* to *ZUZ* and from *BAAB* to *ZUUZ*, have the property that no two words in conjunction can make another word in the system by chance combination of letters. They thus eliminate the necessity for "word-stop" marks and thereby increase coding efficiency.

A most imaginative use of computers, similar to the foregoing example, has been the coining of "drugless names" by the thousands for Chas. Pfizer & Co. (*25*). A computer, again programmed in a special way, combined syllables to make

names for future drugs, thus neatly avoiding a tedious and formerly unsystematic task.

The Revised Standard Version of the Bible obviously requires a new concordance, and it was the happy thought of Reverend John W. Ellison that this be made with the aid of UNIVAC (*26*). Every word of the new Bible was entered, with its context, onto four reels of magnetic tape. By proper programming, the machine eliminated 132 frequently used short words (thus reducing the number of entries from 800,000 to 350,000) and then rearranged the words alphabetically. The output included the context and book, chapter, and verse.

The foregoing four examples have had outputs that would correspond to preparing compilations of scientific data and thus illustrate only that part of the process.

A meaningful and pertinent use of data handling for answering questions has recently been described by two chemists from the Dow Chemical Company (*27*). They have attacked an old problem in chemical literature searching: how to select chemicals having parts of structures in common. For instance, it may be required to select from an "inventory" of chemicals, stored on tape, all those containing two nitro groups, or those with three or more rings, or even chemicals having groups in a specified orientation to one another.

Using a specially designed code for chemical structure, and with the aid of a general-purpose, stored-program digital computer, Opler and Norton have been able to program a search on 1000 compounds that takes only a few seconds to complete. A manual on this program has appeared (*28*). The code for this experiment is of interest because it is derived from more general topological solutions, which have a bearing in searching circuit diagrams, maps, and the like (*29*).

It may also be of interest to record here that a mathematical model for integrated data systems has been proposed (*10*, p. 275).

Summary

This brief survey of integrated and electronic data processing has touched on such matters as the origin of the concepts, their use in business, machines that are available, indexing problems, and, finally, some scientific uses that surely foreshadow further development. The purpose of this has been to present for the consideration of scientists a point of view and some techniques which have had a phenomenal growth in the business world and to suggest that these are worth consideration in scientific data-handling problems (*30*).

To close, let me quote from William Bamert on the experience of the C. and O. Railroad once more (*8*, p. 121): "Frankly, we have been asked whether we weren't planning for Utopia—the implication being that everyone except starry-eyed visionaries knows that Utopia is unattainable. Our answer is that of course we are! Has anyone yet discovered a better way to *begin* program planning of this nature? Our feeling is that compromise comes early enough in the normal order of things."

References and Notes

1. I wish to thank Earl L. Green for his help in discussing this paper, but the errors, whether of omission or commission, are my own. The views expressed here are not necessarily those of the Chemical Abstracts Service or of the American Chemical Society.
2. H. F. Klingman, Ed., *Electronics in Business* (Controllership Foundation, New York, 1955).
3. E. Marting, Ed., *A New Approach to Office Mechanization: Integrated Data Processing through Common Language Machines* (American Management Assoc., New York, 1954).
4. Haskins and Sells, *Data Processing by Electronics* (Haskins and Sells, New York 1955).
5. R. G. Canning, *Electronic Data Processing for Business and Industry* (Wiley, New York, 1956).
6. Moore Business Forms, Inc., *Automated Data Processing* (Niagara Falls, N.Y., undated).
7. Standard Register Co., *Paperwork Simplification*, No. 36 (4th quarter, 1954).
8. M. J. Dooher, Ed., *Electronic Data Processing in Industry* (American Management Assoc., New York, 1955).
9. R. H. Brown, *Office Automation/Integrated and Electronic Data Processing* (Automation Consultants, Inc., New York, 1955).
10. G. Kozmetsky and P. Kircher, *Electronic Computers and Management Control* (McGraw-Hill, New York, 1956).
11. M. P. Doss, Ed., *Information Processing Equipment* (Reinhold, New York, 1955), later chapters.
12. A product of Commercial Controls Corp., Rochester, N.Y.
13. Stanford Research Institute, *Research for Industry* 7, No. 9 (Oct. 1955).
14. British Pat. Specification 708,780, assigned to Compagnie des Machines BULL.
15. A product of International Business Machines Corp., New York, N.Y.
16. A product of Sperry Rand Corp., New York, N.Y.
17. L. N. Ridenour *et al.*, *Bibliography in an Age of Science* (University of Illinois Press, Urbana, 1951), pp. 55–56.
18. E. C. Berkeley, *Computers and Automation* 5, Nos. 3, 9 (Mar. 1956).
19. M. Taube *et al.*, *Studies in Coordinate Indexing* (Documentation, Inc., Washington, 1953–); vols. I–III have appeared.
20. J. W. Perry, A. Kent, M. M. Berry, *Machine Literature Searching* (Western Reserve Press; Interscience, New York, 1956), annotated bibliography.
21. Current work is often reported on in journals such as *Am. Documentation*, *J. Documentation*, and *Special Libraries*.
22. Anon., *Ind. Laboratories* 7, No. 9, 48 (Sept. 1956).
23. Rand Corp., *A Million Random Digits, with 100,000 Normal Deviates* (Free Press, Glencoe, Ill., 1955).
24. H. P. Luhn, *Self-Demarcating Code Words* (International Business Machines Corp., New York, ed. 2, 1956).
25. Anon., *Chem. Eng. News* 34, 774 (1956).
26. W. R. McCulley, *Systems Magazine* 20, No. 2, 22 (Mar.–Apr. 1956).
27. A. Opler and T. R. Norton, *Chem. Eng. News* 34, 2812 (1956).
28. ———, *A Manual for Programming Computers for Use with a Mechanized System for Searching Organic Compounds* (Dow Chemical Co., Pittsburg, Calif., 1956).
29. A. Opler, *A Topological Application of Computing Machines*, presented at the Western Joint Computer Conference, San Francisco, 8 Feb. 1956.
30. After this article was set in type, the following new book appeared: R. N. Anthony, Ed., *Proceedings Automatic Data Processing Conference* (Harvard Univ. Press, Boston, 1956).

Cryogenic Instrumentation

J. G. Daunt

Progress in low-temperature technology has been associated with the development of methods of producing lower and lower temperatures. Milestones in this progress have been the successive achievements of the large-scale liquefaction of the so-called "permanent" gases, in particular air, hydrogen, and helium. It is now well over half a century that liquid air, as well as its important components liquid oxygen and liquid nitrogen, has been available. In this time the technology of liquid air temperatures has formed the basis for a multipurpose large-scale industry. Many plants operate today to produce liquid oxygen at rates of 120 tons per day (*1*), and the commercial needs for these low-temperature products continue to increase.

Production and Transportation of Low-Temperature Refrigerants

The development of the production of liquid hydrogen and liquid helium on a commercial basis, however, is relatively

Dr. Daunt is professor of physics at Ohio State University, Columbus.