THE AMERICAN MINERALOGIST, VOL. 50, OCTOBER, 1965

DATA PROCESSING: A CHALLENGE TO GEOLOGISTS

S. C. ROBINSON, *Geological Survey of Canada.*

ABSTRACT

Geology has been handicapped more than any other science by the difficulty of comparing essential data. Concepts and theories are normally supported by maps and reports which are themselves the products of integration and interpretation of primary data. The techniques of data processing offer an opportunity to geologists to make data from field and laboratory available to their colleagues. These techniques can be used for descriptive data only if the observations, classifications and terminology are entirely objective and consistent.

Studies of the earth employ facets of virtually all other physical sciences and for this reason most new techniques developed in those sciences find a use in some branch of geology. Some of these techniques—such as use of the mass spectrometer for age determination or of the electron probe for study of the precise distribution of elements in rocks and ores—have made possible major advances in the science. In this context, application of data processing techniques to storage and retrieval of geological data may appear to have pragmatic rather than fundamental significance. However a recent study of the feasibility of their application on a national scale in Canada suggests that they might have an almost revolutionary effect on the evolution of the science.

This revolution, if achieved, would be due to systematic and consistent gathering of geological data and their storage in a machine-processable medium, so that they may be available as a basis on which to develop and test geological concepts. Such data would provide an integrating force which would permit different specialists in earth sciences to compare their data and to bring them together in coherent classifications, theories and, ultimately, laws. Of even greater importance these data would provide generalists in geology with a common ground on which to debate and resolve conflicting ideas.

A majority of geological data are observations made in the field and even data from the laboratory are, in general, useful only if they are related to field observations. Because of this dependence upon field observations, it is only rarely that the data on which a geologist develops his ideas and concepts can be observed by another. Moreover, the necessity for interrelating so many variables in making observations in the field has led geologists to record only interpreted or integrated information rather than the data on which it is based. It is only rarely for example that a geologist records the data that lead him to describe an outcrop as granite.

Data processing techniques employing either portable punches or

spread sheets make it possible for the geologist to record the actual data of geology without depriving him of the advantage of recording his interpretation of those data in his notebook. The principal advantage here is that geologists could have access to a large volume of objective facts free from subjective integration or interpretation by the observer.

The second advantage, which is more difficult to achieve, would be freedom from subjectivity in the selection of facts that are to be recorded. Methods of recording data for machine processing require a definite format and thus it is possible to introduce uniformity in the selection of data to be stored. If this is achieved it becomes possible to test hypotheses developed in one area against data from another. This should lead to an orderly evolution of geological theory by retention of those aspects of hypotheses that are broadly applicable as tested by data consistently recorded and rejection of those aspects that are only locally true. This would be a great advance over the all-too-frequent controversy between protagonists of hypotheses that have each proved to be locally sound but which for lack of objective data cannot be scientifically reconciled.

A third major advantage is the opportunity to employ statistical procedures, firstly to evaluation of internal consistency of geological data and secondly to establish the validity of hypotheses. Many tests of fit have been devised, and the more sophisticated procedures of regression analysis, such as trend surface analysis, offer methods that may assist geologists to differentiate between the effects of different geological events and of similar events or processes that took place at different times. Current criticism of these promising methods is due largely to their application where data are either insufficient, are inconsistently recorded, or are incomplete. Given consistent collection of data from many geological environments, these methods should provide geologists with new tools of great potential.

Data collected for specific projects are usually of importance for many other studies. In the case of government geological services, the principal products are maps and reports which represent professional integration and interpretation of field and laboratory data for specific purposes. These data if stored in a machine-processable medium could be made available to the public as an additional service. Any officer of a geological survey could cite numerous examples or requests from mining and oil companies for actual data to supplement information provided in publications. Moreover, insofar as development of geology as a science is concerned, availability of these data to scientists in universities and other research organizations is of even greater importance for two reasons. Firstly, data on which to base or test new concepts would be available in sufficient volume to meet statistical requirements, and secondly,

students and professors alike would be able to devote more time to developing ideas and less to the relatively routine processes of assembling data.

The principal method of research in geology today still has much in common with Chamberlain's system of multiple hypotheses. However as a result of the increase in the number of variables that must be considered by modern geologists, it is usually difficult for any one scientist to find more than a single hypothesis that adequately explains his facts. If, however, when this hypothesis is published, the facts are available in machine-processable form, the experience and scientific imagination of other geologists may be brought to bear on the problem and alternative concepts and hypotheses based on the same data may be developed. This should result in a more orderly development of the science of geology, than is at present possible when choices must be made between apparently conflicting hypotheses based on different sets of facts from diffent areas.[1]

The problems to be solved in attempts to apply data processing methods to storage and retrieval of geological data are so formidable that many geologists who have studied the matter doubt that application on a broad front is feasible. Briefly, application of data processing procedures to descriptive data requires that those data must be stated in terms approaching mathematical precision. Because geology is noted for its wealth of specialized terminology and of differing classifications this will be difficult to achieve. Classifications and names of rocks for example may be based on such different considerations as genesis, chemical or mineral composition, inferred processes, texture and grain size, size and shape of the geological entity concerned, and other factors. Obviously it will be essential to standardize terms used in descriptive data and to do this we must first decide what we can accept as data.

Probably the most useful criterion of what should be included in any file of geological data is consistency in use. Tests made by one company in which several geologists independently recorded observations on outcrops and core in an area of sedimentary rocks indicated clearly that there was acceptable consistency in the use of only 2 or 3 rock names. This reflects the results of trying to record information or processed data

[1] The problem of bringing to bear the opinions of different geologists and other specialists on a single set of facts has long been recognized. In recent years geologists in Canada in cooperation with two mining companies have made comprehensive studies of two orebodies as they were mined. Dr. J. A. C. Fortescue of the Geological Survey of Canada has proposed setting aside specific areas across Canada that could serve as "bench marks" for development of geological ideas.

instead of the basic data themselves. It indicates that observations to be recorded must include such items as grain size, texture, mineral content, shape and attitude of the rocks and other factors on which a dominant percentage of geologists (90 per cent or more) could be expected to agree. Obviously the inclusion of information that cannot be consistently recorded and universally understood by geologists will defeat the purpose of storing geological data.

A second criterion is that data to be recorded shall be what the statistician refers to as natural populations. As one of my confrères, Dr. F. P. Agterberg observes, "Computations based on rocks classified by some artificial scheme may easily lead to results which solely reflect on *a priori* suppositions. The usage of certain traditional terms in cataloguing data will be unavoidable, but users of the file should realize that these terms are a necessary evil. . . . " Obviously it would detract from the usefulness of the stored data if we attempted to dispense with distinctions that are so commonly used in Geology as to be almost universally accepted. For example, it would be extremely difficult to develop a usable classification of rocks if we did not divide them into igneous, metamorphic and sedimentary groups. To do this we must make provision for a fourth group into which we place rocks whose genetic classification is in doubt. Where it is necessary to use such interpretative distinctions, scientists using the data may choose to disregard them and to draw from the file only those facts that are unequivocally defined, such as descriptions based on mineral and elemental content, grain size, texture, etc. Thus although it would be preferable to restrict contents of the files to data completely free from interpretation, some compromise must be made in the interests of utility.

In the Canadian study consideration was given to the feasibility of devising a single system for recording data of all those disciplines that together comprise the earth sciences. It was unanimously concluded that such a system could not fulfil the requirements of all the disciplines without becoming so complex and ponderous as to defeat its own purpose. Instead it was recommended that machine-processable files be established for each of about 30 subjects and that these be linked by certain components that would be common to all. Moreover contents of these files would be linked through a general index.

Components that would be common to all files are:

1. A common reference numbering system
2. A common system of geographical location
3. A common system for classification and coding of descriptive terms.

Some thought was given to the possibility of using coordinates of geo-

graphic location as reference numbers but despite its advantages this proposal was found to be impractical because if the number were to be sufficiently unique it would have too many digits, and because it is often difficult in the field to locate a position precisely by coordinates. A sequential system of numbering provides the shortest reference numbers but some central authority is essential to issue the numbers. Probably a system combining sequential numbering within each organization with a code number indicative of the organization is the best solution.

There is a surprising disparity of views on the best method of indicating geographic location. In Canada the choice lies between use of latitude and longitude and the rectangular grid of the Universal Transverse Mercator system and in either case location to the nearest metre seems desirable. Location in the vertical dimension may only be necessary where stratigraphic, geophysical or other factors require it.

Classification and coding of descriptive terms is the real challenge to modern geologists. Successful use of data processing depends upon our willingness to select and stick to one classification for all rocks and for other descriptive terms. To be effective the classification must be capable of expressing descriptions with adequate precision. For example, it is likely that the best classification of rocks will describe them in terms of texture, mineral content, etc. expressed preferably in quantitative terms. Above all the classification and terms used must be based on facts for which consistency in use can be assured.

If use of data storage is to attain its maximum use, it is desirable that there be broad, preferably international, agreement on the three factors discussed above, on choice of subject files and even on card format. To achieve this uniformity will be exceptionally difficult when it is recognized that at present more data are punched into cards for single, uncoordinated projects then are punched for the kind of archival storage envisioned in this paper.

Uniformity should be sought first amongst geologists who have interests in common. It is likely for example that those in government and universities will be more interested in exchanging data than those in companies whose data are usually confidential. Already geologists interested in exchanging geochemical data have begun discussions internationally under the leadership of Dr. D. F. Davidson of the United States Geological Survey's Bureau of Geochemical Census. Other international files are evolving in two fields of geochronology, and there is widespread interest in many other geological disciplines. It is obvious that the more geological data files are developed independently, the more difficult and expensive it will be to adopt a common system. There

is therefore a real advantage in early standardization of as many facets of a data storage system as possible.

The application of data processing poses a real and immediate challenge to all those concerned with the future of Earth Sciences. Geology as a science could achieve a modern renaissance through uniform and consistent application of data processing techniques to storage and retrieval of its data. The decision to do so would involve a large degree of regimentation in the selection and presentation of geological data but at the same time would confer an even greater degree of freedom to exercise scientific imagination and initiative.