A computer-based registration system for geological collections

J. H. Germeraad, M. Freudenthal, M. van den Boogaard,
C. E. S. Arps

Germeraad, J. H. et al. A computer-based registration system for geological collections. - Scripta Geol., 9: 1-12, Leiden, June 1972.

The new computer-based registration system, a project of the National Museum of Geology and Mineralogy in the Netherlands, will considerably increase the accessibility of the Museum collection. This greater access is realized by computerisation of the data in great detail, so that an almost unlimited number of operations to select and sort the data is possible. A flexible design of annotating information, mainly in plain words, saves considerable time. The fast mechanical data processing permits the efficient preparation of catalogues which contain selected information about the geological collection; as an additional benefit labels may be produced at low costs. The RGM system disposes of the burden of classical registration books and card-indexes, without any essential quality being lost. It combines a minimum of man-hours with an optimal flexibility in storage and retrieval of data, and an acceptable employment of computer time and equipment.

The system may also serve for the registration of collections of other institutions, in geology as well as zoology and botany.

J. H. Germeraad, M. Freudenthal, M. van den Boogaard and C. E. S. Arps, Rijksmuseum van Geologie en Mineralogie, Hooglandse Kerkgracht 17, Leiden, The Netherlands.

Abbreviations used:
RGM = Rijksmuseum van Geologie en Mineralogie (National Museum of Geology and Mineralogy)
CRI = Centraal Reken Instituut (University Computer Centre, Leiden)
CPU = Central Processing Unit

Introduction

The RGM collection contains over one million samples; internationally this collection rates as one of fair to medium size. In general the information to be recorded varies considerably and this does not facilitate the problems of registration, storage and retrieval. The new RGM registration system matches a simple application with a complex internal structure which need not concern the user.

Acknowledgements – The system announced in this article is the result of the teamwork of many members of the RGM. The authors received most substantial assistance from: Miss G. E. de Groot and Messrs P. J. Hoedemaeker, A. W. Janssen, J. P. Krijnen and C. F. Winkler Prins.

General aspects of collections

Collections of natural objects are characterized by a number of properties that are of great importance to accessibility, storage and retrieval. Although some of the following paragraphs report on specific RGM details, the quintessence is thought to be of general interest.

REGISTRATION

The specific data on any one sample to be registered may be classified in a large number of basic categories, according to: 1) acquisition, e.g. name of collector or donator, 2) systematic group, such as mineralogy, petrology, palaeontology, and subdivisions, 3) geographical origin, 4) stratigraphical position, 5) absolute age, 6) size of the specimen, 7) state of the sample: a fossil may be an almost complete individual, or one or more organs, foot prints, burrows, etc.; a rock sample may be present in its original state or as mechanical or chemical residue, 8) degree of examination, ranging from superficial field determination to detailed laboratory analysis, 9) documentation: field notes, reports, publications, 10) status of the specimen, e.g. holotype.

The annotation in the above categories contains dissimilar designations for the description of more or less equivalent data. The following types may be distinguished: 1) linguistic: due to a variety of collectors several languages are used in the catalogue of the RGM, as for example sandstone, Sandstein, or zandsteen, 2) terminological: both a scientific and a vernacular name, or even more than one of each, may be in use to describe a certain specimen, for example granulite – Weiszstein, crinoidal limestone – petit granit, 3) systematic: there exists a great number of synonyms, and partly or entirely equivalent scientific names, such as *Nummulites* – *Camerina,* tetrahedrite – Fahlerz or panabase, arenite – sandstone, 4) chronostratigraphical: at each level of the subdivision some terms overlap partly or entirely the terms at other hierarchical levels, for example Miocene – Burdigalian, 5) lithostratigraphical and biostratigraphical: there is an extremely large number of parallel subdivisions, and a term in one subdivision may cover one from another subdivision: e.g. Guasare Formation – *Turritella mortoni* Zone.

STORAGE

In the past the RGM samples have been stored in separate units according to several criteria: systematic position, geography, stratigraphy, collector, and size of specimen. These different arrangements made both storage and retrieval complex. To maintain the growing units as entities, free space was reserved between the units in anticipation of new material. When more material came in than was expected, shifting in the storage room became necessary. Or, conversely, when new material did not turn up, the extra space was reserved in vain and the investment wasted. Any solution will have to meet the exigencies of retrieval.

RETRIEVAL

In large collections, retrieval of specific samples is an ever recurring problem. In fact the time involved in finding a desired sample determines to a large extent the value of the entire collection. It is evident that good retrieval facilities include first of all a detailed and up-to-date registration of the actual location of every sample, either in the storage-rooms or in the exhibition, exchanged with other institutions or lent out.

The necessity of a detailed registration system on behalf of retrieval may be shown in the following example: one scientist may be interested in examining the entire fossil content of a specific stratigraphical level; another, however, wishes to see the material of only one systematic group, but throughout its entire stratigraphical succession. The former would be quite content if the collection were arranged stratigraphically, whereas the latter would prefer a systematic arrangement. Both are bound to get an incomplete answer if parts of the collection are arranged according to different criteria. The only solution is a sufficient number of differently arranged card-indexes coupled with the perfect registration of the location of the specimen. However, keeping the card-indexes of an expanding collection up-to-date demands a considerable amount of time and becomes almost impossible when the storage system is repeatedly altered by various persons in different ways.

Today it is common knowledge that a computer-based system for registration is the only solution for many of the problems discussed. The RGM has established a system with two principle advantages, viz. it permits retrieval on complex, intricate requirements, and it introduces a storage system in which no spare room has to be reserved and shifting of samples is avoided.

The retrieval system does not preclude the use of other storage systems, preferred by participating institutions.

General facts of computer-based registration systems

In any system the most time-consuming part of the whole procedure determines its efficiency and this part may form a bottleneck that causes the system to become technically or financially inexecutable. It is assumed that the first steps in the registration of data, i.e. both the establishment of all the tables with code symbols, and the annotation of the coded data on so-called punchdocuments, may be that bottleneck. With the enormous number of data to be handled, only a very short time for the coding of each sample is acceptable. The problems involved in the coding of data will now be discussed in detail.

CODING SYSTEMS

Two methods of coding may be applied, one by means of the words we all use in our daily work, and the other by means of symbols, like the cyphers used in arithmetic. Both methods have their advantages and disadvantages.

Coding in words – Advantages are: the scientist knows the meaning of the term; he is also familiar with the spelling, so that errors can easily be detected. Disadvantages are: the spelling should remain unchanged throughout the years; the terms may be of considerable length; the various terms differ in length; quite acceptable synonyms may easily pass undetected.

Coding in symbols – The term is substituted by one or more cyphers and/or letters. Advantages are: the spelling is constant; the length of the symbol is small; this length can be kept constant in more than one category; the decoding of the symbols permits the conversion into many terminologies and languages; more or less synonymous terms fall into one and the same symbol code. Disadvantages are: the great number of terms makes it virtually impossible for the codifier to know all the codes by heart; consequently he is obliged to consult frequently a large number of voluminous tables (as an illustration: a small subdiscipline like angiospermous pollen descriptions requires over two hundred code tables); errors are hard to find and if, by any chance, they show up in the decoded text, their correction demands renewed consultation of the tables. Evidently all this is a cumbersome routine.

The time-consuming aspect probably outweighs all factors which are in favour of coding in symbols; the latter is only acceptable for short tables, which need not be consulted frequently, as they can easily be remembered. However, the tables of many categories will contain thousands of terms, and therefore the coding with symbols in these categories is not feasible.

VERIFICATION OF INPUT DATA

A second time-consuming aspect is the verification of every coded item before it finally enters the data bank. It will be clear that in a symbol-coded system the number of annotation errors is likely to be larger than in a term-coded system, thus symbol-coding requires more checking. Whenever possible this should be done by the computer.

In a symbol-coded system two types of errors can be traced mechanically. First the code symbols that have not been punched in their proper columns, and second the code symbols that have been entered upon the cards, but do not exist in the code tables. Effective checking of the first type of errors requires a rigid format (i.e. an inflexible layout of the punched card). Any later changes in this format, resulting from an extension of the system, may also have consequences for the data bank.

However, a third type of error, the erroneous use of an existing code number, cannot be detected by the computer, but can only be traced by man. In a term-coded system the first type of error does not occur, as the format is entirely flexible. An even more important advantage of this system is, that the machine can test the accuracy of the terms used. This is achieved as follows. Each term to be entered in the data bank should occur in a previously established file of terms. These files, one for each category of data, are made only once. The machine compares each term of the input data with the corresponding file, and gives a message whenever a term is lacking. If the absence is not due to an error in the spelling, the term is a really new one and should be added to the file (updating), a simple procedure. The input will then be accepted in the next machine run. An extension of the system will merely require the establishment of new files, without consequences for the existing data bank.

CONVERSION TO OTHER DATA BANKS

It is obvious that more differentiated systems of codification can easily be converted into less differentiated ones, but not the other way round. Quite conceivably, the scientific institution that has a more differentiated system will not be eager to spend much effort in any exchange of data from which it can expect little benefit.

Design of the RGM registration system

After several years of research a system has been designed that combines both types of coding: symbols and words. These are used for different categories in a procedure of manual and mechanical data handling (e.g. the writing of punchdocuments and the computer processing respectively).

PUNCHED CARDS

The punched cards used are the standard eighty-column cards. Symbolic codes – cyphers and/or letters – are brought together on a single card which, as it receives the special punch 1 in the first column, is called the -1- card.

Term-codes are annotated on one or more following cards which receive a 2-punch in the first column and are all called -2- cards. There is no limit to the number of -2- cards in the registration of a sample.

The -1- cards contain codes of three groups of data:

a) all data that are already symbolic, such as registration number, storage number, field number, geographical co-ordinates and absolute age;

b) data easily coded by symbols as they occur in categories which have few items, such as the code-numbers of the chronological subdivision, and the institution in which the sample is kept or was previously registered;

c) data concerning hierarchical classifications. Selecting and sorting on all levels of the hierarchical subdivision is very easy by means of code-numbers. For example, the systematic levels in the palaeontological classification, from phylum to tribus, may be coded. (Genus and species are recorded by name in the -2- cards and therefore not incorporated in this code). The classification levels chosen for one phylum need not be equivalent with those selected for another phylum, e.g. in Mollusca the lowest level chosen may be the family, in Mammalia the subfamily. The only criterion in this respect is the usufulness of the classification. The hierarchy in the classification of rocks and minerals is similarly coded.

The classification code numbers permit the retrieval on all hierarchical levels of these groups. To achieve the same facilities with term-coded classifications, all possible terms on all levels of this hierarchy have to be annotated. It is evident that in this case symbol-coding is to be preferred, in spite of the cumbersome consultation of the necessary code tables. Apart from saving time in the end, it also saves space in the computer processing system.

However, classifications are neither unique nor permanent, as specialists tend to be influenced by fashion, and also by the emphasis they themselves put on certain aspects. Thus provisions for the changing of code tables and the subsequent updating of the data bank are required. To this purpose a special signal is introduced, that precedes the classification code-number and indicates its classification concept. The classification code-number consists of two parts, one for the higher systematic levels, and one for the lower. The annotation on the punched cards of the former is compulsory for all samples. On the other hand the application of the code-numbers for the lower systematic levels is optional.

The -2- cards contain all term-coded information. The terms of some categories have already been annotated numerically on the -1- cards. This double coding offers the possibility to check mechanically the code-errors in the -1- cards, a check that otherwise ought to be performed by man.

At present twelve categories of term-coded data are distinguished: sample type; locality name; collector; donator; preservation; first name (e.g. genus or rock name); second name (e.g. species name); local stratigraphy; absolute age; status (type-specimen, published or unpublished, etc.); organ, habit, constituents, texture, structure; facies, paragenesis, petrogenesis; and in addition uncoded 'other information'.

This number of categories is not limited by any means. Nor is there any restriction to the number of terms used within one category (e.g. genus name, subgenus name, and synonyms if wanted). If the information on a sample is incomplete, no spaces have to be left open for the missing categories. There is no need for a fixed position of the categories on the -2- cards. For each sample the most practical sequence may be chosen in order to save time.

The essential device that permits this open flexible arrangement is the use of a number of 'stop-symbols', each with its own meaning and value. They serve to separate the data of the different categories, and the different terms within one category, and also to distinguish the terms from the 'subterms'. The subterms serve to detail the information of the terms. The combined use of terms and subterms results in a maximum of sorting possibilities at different levels. For example the annotation /Jurassic/beta/lower/ can cope with selections that either ask for "Jurassic" only, or "Jurassic beta", or "lower part of Jurassic beta". Or: /limestone/ micritic/skeletal/ gives access to the collection by means of selections that ask for "limestone" only, "micritic limestone", "skeletal limestone", or "skeletal micritic limestone". Likewise the arrangement /*Cardium*/*Cerastoderma*/ within the genus group (called 'first name' in the RGM system), permits the retrieval of all specimens listed as *Cardium,* of all specimens of the genus *Cerastoderma,* and also of all specimens of the subgenus *Cerastoderma* within the genus *Cardium.* Similarly species and subspecies names can be annotated, to which the name of the author and the year of publication may be added as subterms.

Within each category an unlimited amount of non-coded information may be given after the last stop-symbol. Such information is not meant for sorting or selecting, and its spelling is completely free. Data which do not fit into any of the twelve categories mentioned can likewise be stored as 'other information' (see above). In this way it is possible to store in our registration system virtually all data. For example it may be important to introduce at full length the geological description of a locality where samples were collected. The recording of any amount of irrelevant information is technically possible, but should be avoided for practical reasons.

PREPARATION OF THE PUNCHDOCUMENTS

Each line of a punchdocument serves to plot the data to be punched on a single eighty-column card. Identical data in identical columns may be duplicated rapidly on the punching machine. To get the utmost benefit from this device, the samples are arranged in groups that have a maximum of identical information in common. This recurring information has then to be entered on the punchdocuments in identical columns, as follows:

a) the data for a group of -1- cards on successive lines of a punchdocument, so that many symbol-coded data can be marked for automatic duplication,
b) the data for a set of first -2- cards on another punchdocument, on which are annotated those terms that are identical in many successive samples; here again the repetition on the punchdocument saves time in writing and punching,
c) the data for a number of second -2- cards on yet another punchdocument, and so on, until all identical data have been annotated. Finally the dissimilar data are entered on the next punchdocument and/or following ones. It should be stressed that this way of registering the data on the -2- cards is the fastest procedure known. Its speed surpasses that of any ingeniously designed symbol-code system by far.
d) all cards begin with the registration number so that the operational sequence of the cards can be achieved by the card-sorting machine.

AREA MAPS

The geographical co-ordinates have been chosen as numerical codes for welldefined sample localities. Whenever merely a general region is recorded, a rather common designation in old collections, the co-ordinate code is replaced by the code number of an area map that stands for this specific region. These area maps may cover any geographical area, independent of political frontiers. They may be used in mechanical data processing, provided they are defined by a code number, a name, the outlines of the area, one meridian, one parallel and a specified scale (fig. 1). All co-ordinates within that area are computed mechanically and recorded under the number of the

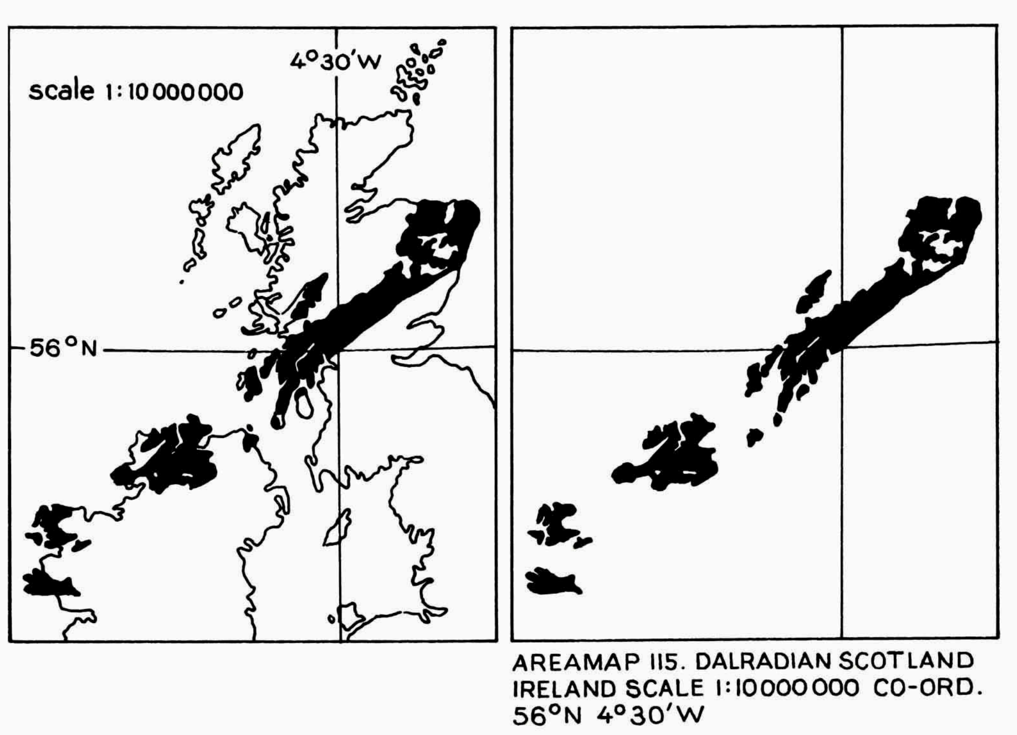
scale 1:10000000
4°30'W
56°N
AREAMAP 115. DALRADIAN SCOTLAND IRELAND SCALE 1:10000000 CO-ORD. 56°N 4°30'W

Fig. 1. Left: map of the geographical distribution of the Dalradian Series in Scotland and Ireland. Right: area map of the Dalradian, specially prepared for processing by computer.

area map. This set-up offers a most versatile device to meet the various requirements for retrieval on a geographical base. It is possible to select all samples from a specific area, not only those actually coded with the number of that area, but also those from localities within that area, which are defined in the data bank by co-ordinates.

SOME TECHNICAL DETAILS ABOUT THE PROCESSING SYSTEM

The processing is split into two almost independent parts, i.e. file processing and data bank processing. As the available computer IBM-360/65 is a big machine, it is possible to combine both parts in a single machine run (see fig. 2). In smaller machines they may be processed separately.

File processing – In a complete job three sets of cards are to be processed:

1) A deck of file-cards, containing the checked new terms that have to be incorporated in the existing term-files. It is preceded by a special signal card initiating the processing. A signal card at the end of the deck announces the next set.

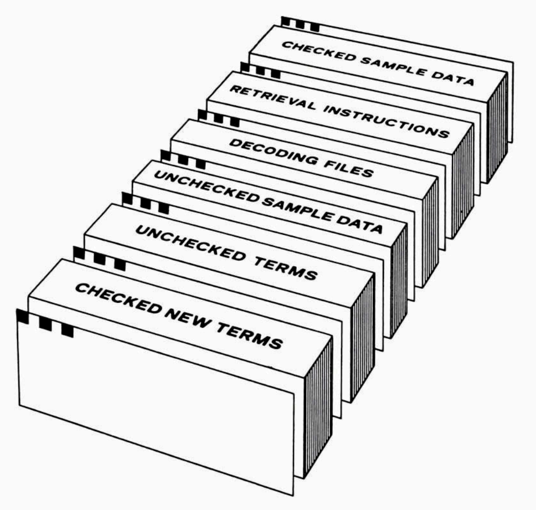

Fig. 2. Sequence of the card decks for processing geological data. The signal cards that initiate each step and terminate the machine run are marked with black squares.

2) A deck of file-cards, with the terms that have not yet been compared with the existing files. Whenever a term is not encountered in the files, it is printed with a special message. The end-card of this deck initiates the processing of the next.
3) A deck of cards which contain unchecked sample data. Although in general most codes and terms in these data already occur in the existing files, the data have to be screened for codes and terms which are not yet present in the files. These codes and terms are printed together with the registration number of the sample.

Data bank processing – A complete job comprises two parts: updating of the data bank and retrieval of data. However, either may be executed without the other. Whenever a retrieval is required, the relevant code-files have to be read, so that the symbols encountered in the data bank can be converted into readable words. Thereupon the cards with the retrieval instructions are read.

The old data bank is updated with the checked new sample data by reading one card at a time and one tape-record at a time. After comparison the data of the card can enter the data bank in their proper place. Any correction of an existing record in the data bank can simultaneously be accomplished.

A retrieval may consist of one or more selections and/or sorts. A selection simply assembles specific information to be printed in the sequence that exists in the data bank. A sort, on the other hand, results in a rearrangement of the selected data in a different sequence. At the moment that the data of either a card or a record are to enter the new data bank tape, they are compared with the items of the retrieval requests. The new sample data, if present, are printed immediately afterwards. All data of selections and/or sorts are saved on separate disks, from which they are retrieved and subsequently printed.

CATALOGUES

The RGM plans to print periodically a 'General Catalogue' that contains all information of each specimen. Other catalogues will be produced by means of special retrieval programmes which arrange the data according to stratigraphy, geography, systematics, etc. Their contents will be limited to the basic information on the samples. In fact these catalogues replace the former card-indexes. For most purposes they will be sufficient, but if they do not give the information wanted, adequate catalogues can be specially processed as a matter of course.

LABELS

The mechanical data processing permits the simultaneous preparation of labels. The advantage of this procedure is, that much time is saved and no errors are introduced by manual conversion of catalogue information to labels. Furthermore it will be possible to keep all labels updated.

SAFETY PRECAUTIONS

The entry of new data is preceded by a printout for final checking and approval. This printout will be the property of the contributor. In special cases the data can be protected against abuse by a built-in automatic safe-guard preventing unauthorized data retrieval. As an added precaution the contributor may request the complete registration to be printed in an entirely private numerical code (for origin, age, etc.). It may be emphasized that a plain printout of the data bank tapes or disks merely shows very complex internal code numbers. These cannot be decoded without knowledge of the keys of the system.

FEASIBILITY

Many attempts to computerize a large amount of information have failed, because they involved too much money or man-power. This in spite of the fact that the mechanical data processing is so cheap that man cannot possibly compete with it. Generally these mishaps may be imputed to underestimating the amount of time involved in annotating the data for input. Again, the way of recording may not have been thoroughly investigated. If the average time to register the data of a single specimen would be for example fifteen minutes, then the manner of annotation is obviously unsound. We consider that the average time spent on the transfer of data from the old system to the new should not surpass three minutes per sample. We expect that about fifty thousand samples of our backlog can be handled annually, in addition to the processing of the incoming samples. Meanwhile it will be quite possible to process a large amount of information from other institutions.

We assume that the RGM registration system will be operational by the end of 1972.

Co-operation with other institutions

After several years of preparatory work, the RGM has started with the establishment of its data bank. Other institutions in the Netherlands have already shown interest in participation. The board of the data bank will advise them how to comply with the requirements of the RGM system. If the measures recommended can be effectuated, participation can be realized. For the time being we reckon with the following categories of contacts, each having its particular benefits and obligations:

1) Co-operators s.str. having their own data banks: generous exchange based on reciprocity. 2) Participants, whose data are incorporated in the RGM data bank: service at cost price. 3) Clients who have contributed nothing to the RGM data bank: the client will be charged the data processing costs plus a surcharge of maybe 300% as a contribution to the establishing costs. 4) Buyers of the system, exclusive of the data (e.g. botanical or zoological museums): the programme-system is for sale at establishing cost, not at duplication-of-tape price.

Potential co-operators are informed that the exchange of data bank information depends on reciprocal convertibility. The RGM system may be characterized as highly differentiated at all levels and most flexible. Therefore only co-operation with those data banks that have a codification at a comparable level can be taken into consideration.

COSTS

It is not easy to give exact figures concerning the financial aspects of the use of the data bank, as the requests may differ considerably. In general a machine run for selecting data may be expected to take several minutes of the CPU.

The mechanical data processing takes place at the CRI. Currently the users outside the Leiden University are charged with the CPU-time only, for the time being at a rate of Dfl. 3600.— per hour. In the near future the rent of tapes and the costs of printing paper, and perhaps 14% tax must be considered too. Moreover, in a more distant future, all users may be charged with the following 'minute' costs (figures are IBM rates dated August 1970):

tape access	per 'read' or 'write'	Dfl.	0.028
disk use	per input or output	,,	0.030
printing	per line print	,,	0.011
card reading	per 'card read'	,,	0.011

These cheap operational steps occur so frequently during machine processing that their total costs may exceed those of the CPU time involved.

It will be clear that any co-operation will have to be preceded by thorough discussions, in order to avoid disappointment on either side.

RGM LETTERS ON REGISTRATION

Information about the RGM system will be published in bulletins which may serve as manuals for co-operators and participants. These papers will contain technical details on the construction of code-files, the procedure of writing punchdocuments and on any other question that may arise.

Persons or institutions interested may apply to receive the RGM letters on registration. Applications should be addressed to:

Dr M. van den Boogaard
Rijksmuseum van Geologie en Mineralogie
Hooglandse Kerkgracht 17
Leiden, the Netherlands.

Literature

Berner, H. et al., 1971. Data storage and processing in geological mapping, I. Field data sheet; II. Data file. – Geol. Fören. Förhandl., 93, 544: 85-101; 93, 547: 693-705.

Cogeodata, 1971. Geological data files, Results of an international inquiry. – I.U.G.S. Doc. 31.

Creighton, R. & R. King, 1969. The Smithsonian Institution information retrieval (SIIR) system for biological and petrological data. – Smithsonian Inst. Inf. Systems Innovations, 1: 1-25.

Crovelto, T. J. & R. D. Macdonald, 1970. Index of EDP-IR projects in systematics. – Taxon, 19: 63-76.

Hall, A. V., 1972. Computer-based data banking for taxonomic collections. – Taxon, 21, 1: 13-25.

Harrison, R. K. & P. A. Sabine, 1970. A petrological-mineralogical code for computer use. – Inst. Geol. Sci., Nat. Environment Res. Counsil, Rept. 70, 6.

Hughes, N. F., 1970. Remedy for the general data handling failure of palaeontology. – Systematics Assoc. Spec. Vol. 3. "Data Processing in biology and geology", edited by J. L. Cutbill: 321–330.

National Advisory Committee on Research in the Geological Sciences, 1967a. Storage and retrieval of geological data in Canada. – Geol. Survey Canada Paper, 66-43: 1-98.

——, 1967b. A national system for storage and retrieval of geological data in Canada. – Geol. Survey Canada: 1-175.

Penn, I. E., 1971. A system for the storage, retrieval and analysis of numerical data in palaeontology. – Palaeontology, 14, 1: 154-159.

Rensberger, J. M. & W. B. N. Berry, 1967. An automated system for retrieval of museum data. – Curator, 10, 4: 297-317.

Robinson, S. C., 1970. A review of data processing in the earth sciences in Canada. – Math. Geol., 2, 4: 377-397.

Wynne-Edwards, H. R. et al., 1970. Computerized geological mapping in the Grenville Province, Quebec. – Canadian Jour. Earth Sci., 7: 1357-1373.

Manuscript received 21 April 1972.