The RGM data bank programme for storage and retrieval of geological collection data

M. Freudenthal

Freudenthal, M. The RGM data bank programme for storage and retrieval of geological collection data. — Scripta Geol., 31: 1-22, Leiden, July 1975.

The structure of the RGM registration system is discussed, with special attention to the costs for running this data bank, and for adding new information to it. The programme produces printed catalogues that offer a wide variety of entries into the collection. It is estimated that the entire registered collection of the museum can be transferred to the new system in about five years. Co-operation with other institutions may lead to a national data bank for geological collections.

M. Freudenthal, Rijksmuseum van Geologie en Mineralogie, Hooglandse Kerkgracht 17, Leiden, The Netherlands.

Introduction

The first ideas for creating a computerized registration system for the RGM collections were formulated about five years ago, in 1970. We then went through a theoretical phase of about two years, which we needed to outline the basic principles to which our data bank would have to answer. Construction of the programme started about three years ago. The data bank became operational in April 1974. Further optimalization is estimated to take one more year.

By now the basic programmes for constructing and updating the data bank,

and the most essential output programmes are ready and operational. Some more specialized output programmes, as well as the routines for deleting outdated information are ready in principle, but not yet fully adapted to the operational stage. Filling in these details may take some more time.

It is our intention to put the RGM data bank programme at the disposal of other geological institutions in the Netherlands, so that a central registration of all geological data in public or private collections in our country may be achieved, and a similar co-operation with institutions abroad is certainly among the possibilities. By means of some modifications, the programme is well suitable for the registration of collections from other fields of natural science, and related fields like for example archeology; it might even be adapted to the registration of libraries.

A first publication on this data bank appeared in 1972 (Germeraad, Freudenthal, van den Boogaard & Arps. A computer-based registration system for geological collections. Scripta Geol., 9, p. 1-12). Since then the technical development of the programmes has been mainly in the hands of the present author. My colleagues Drs G. E. de Groot and M. van den Boogaard undertook the task of formulating the scientific and curatorial specifications for the programme, and they supplied the punched cards for most of the 10 000 samples now stored in the data bank.

Aims of the RGM data bank programme

The RGM has a collection of about 200 000 catalogued samples of geological material, and a far greater amount of uncatalogued material. Most of the catalogued specimens are identified by numbers on themselves or labels in their boxes or both. Through the number, additional information can be retrieved in the master catalogue. The master catalogue generally does not give information about the storage location of the specimen and, if it does, this information is generally incorrect. Retrieval on the basis of selected criteria, like age, locality, classification, etc., is sometimes possible through the existing card-indexes, but there is no guarantee that this method will yield a complete answer, since entries into these card-indexes are far from complete.

From the above it is evident that, unless one is very familiar with the existing collections, it will hardly ever be possible to find all the material related to a wanted criterium: part of the collection is stored according to locality, part is stored according to classification, etc. The only solution that gives a complete entry to the collections on the basis of each possible criterium is a perfect system of indexing and cross-references. Practice has proven that keeping such a system up-to-date involves an unsurmountable amount of work, unless a computer is utilized.

Through a computerized registration system it is possible to keep each different kind of catalogue or card-index up-to-date by processing one single form of input information: the sample data are written on punched cards only once, and the computer programme transforms these data into the forms required by the different kinds of catalogues. In the same way, one single input form suffices to delete erroneous data from all the catalogues.

In a non-computerized system it may happen frequently that adding or deleting data to or from one of the catalogues is forgotten or postponed because of lack of time. Errors are also introduced because it is necessary to write the same

information several times, which is not always done correctly.

Such failures can be avoided by computerization. Furthermore a considerable amount of man-hours is gained, because it is no longer necessary to reproduce the same information manually as many times as there are different kinds of card-indexes and catalogues; the same goes for the labels that are laid in the boxes with the specimens; these can also be produced by the computer from the standard input form.

Once the decision was made that in our situation computerization was the best solution we had to deal with two problems:

1. Retrieval of data from a computerized registration system is only possible if these data are coded in some way, so that they may be recognized by the machine when requested. In practice this means that a method of coding must be found, that is easy to handle, even by people not familiar with computer techniques, and that the writing of coded information should not take more man-hours than writing the same information in a classical registration system.
2. Retrieval of data from a computerized system may be made in two ways: One basic data set contains all the information, and each time a retrieval request is made, the basic data set is scanned and the records fitting the retrieval request are selected and printed. Or, the basic input data are processed into a number of data sets, each of which is arranged according to different criteria. These processed data sets are kept up-to-date at every new input run and printed if necessary. Most retrieval requests can be answered by reading the ready prints of these data sets, and only in rare cases it is necessary to find the answer by means of a computer run.

Our solution to these two problems will be discussed in the next chapter.

Structure of the programme

Coding

Standardization of data is essential for every computer retrieval of information. Non-standardized information is not appropriate for machine retrieval. A common form of standardization is symbolic coding: each term is replaced by a numeric figure or a group of alphabetic and/or numeric characters. For various reasons computer treatment of such symbolically coded data is rapid and therefore cheap. However, a large amount of human work is required to prepare the data for input and, possibly even worse, error checking of coded input is tedious. In our first attempts for construction of the data bank we thought of a large degree of symbolic coding, but it soon became evident that this form of input would take too much time from the curators, and that it was therefore not feasible.

We then decided that symbolic coding should be reduced to an absolute minimum and that most of the information should not be coded, but just written in a standardized orthography. This means that computer processing of the input tends to be slower, but preparation of the input data is fast and efficient, so many hours of human work are saved; possibly this counts even more for error checking. Even the loss of time in computer processing of non-coded information is largely made up for, because output programmes do not demand any decoding.

So, most of the information in our data bank system is written in normal words and these words are treated as codes. The length of these code terms is not fixed, as it is in symbolic coding. However, we felt that for technical reasons we should put a maximum to it. This maximum is of no concern to the user who is preparing data bank input. He may use terms of any length, but he should keep in mind, that terms exceeding fifty characters are truncated.

Figure 1 shows that fifty characters alotted to a code term are largely sufficient to contain the characteristic part of the term. Nevertheless symbolic coding still seemed advantageous for some categories of data.

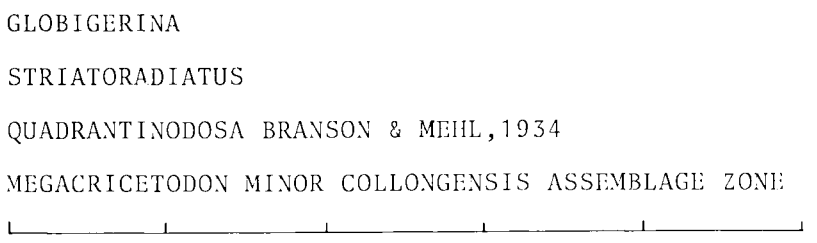

Fig. 1

One of the reasons for using symbolic codes may be, that the resulting code-list is so small that it may easily be remembered by heart, and consultation of the code-book is hardly ever necessary. This reason is valid for the following categories:

Age — A small code-list is used for the main units of chronostratigraphy. It contains about fifty items, so there is hardly any time lost in coding. The code-figures for age allow a certain amount of hierarchic sorting: the code for Cretaceous is 770, the code for Mesozoic is 700. A sort requesting Mesozoic samples will also recognize samples coded as Cretaceous, if the sort request analyses the first digit of the code only.

In our data bank input we use the age code twice: once for the oldest age estimated for the sample, and once for the youngest age considered for it. If a sample has been coded 730 as oldest age and 770 as youngest age, it will be selected in a sort request asking for 750, as 750 is in between the age limits stated for the sample. Such facilities would not be possible if age were not symbolically coded.

Main classification — For paleontology 'main classification' is roughly identical to phylum, for petrography it contains the most important levels of rock classification. The code-list contains about fifty items. The main reason for coding this category symbolically is that this enables us to subdivide other categories into subfiles. E.g. the file containing all species names could be very large; we divide this file into subfiles, one for each main group of classification, and these subfiles are of workable size. The code of the main classification is made part of the subfile number.

Hierarchic classification — A second reason for using symbolic codes is purely hierarchic. Hierarchic sort appeared to be an advantage in age-coding but it is essential for classification. Numeric coding of a hierarchic classification makes

it possible to sort data at any level of this classification by means of one single code-figure. In order to achieve the same facility without numeric coding it would be necessary to introduce a term at each level of classification that might be wanted in a sort request. In the code-figure 1234567890 the 12 may stand for an order, the 34 for a suborder, the 56 for a superfamily, the 78 for a family, and the 90 for a subfamily. The code-figure allows sorting at each of these levels, it replaces five uncoded terms. In this example the code-figure consists of five code-groups of two digits each. In another main group of classification, however, a different set-up may be chosen, for example, the classification code may consist of three code-groups of three, three, and four digits respectively, that may stand for subclass, order, and family. Minor changes in the classification concept within a group hardly affect the applicability of this system. If a major change in the concept occurs we introduce a new classification for the main group in question. The code-figures of this new classification are labeled with a digit that identifies the classification concept, e.g.:
b1234567890, in which b stands for a blank, is a code belonging to the first classification concept used within a specific main group. If a new concept is introduced the blank is replaced by a 0; the code 01234567890 may have the same meaning as the code mentioned before, but it may also have a completely different one. The system allows sixty different classification concepts within each main group, before a general overhaul is necessary. If the need is felt a conversion programme may translate the codes from the old concept into those of the new one; if not, sixty different classification concepts may be operational at the same time.

A third reason for using symbolic coding is, that some categories of data are mostly numeric in themselves. In such cases coding is no extra burden for the user of the system, and computer processing is efficient. Categories are:

Geographic co-ordinates — These are written in a compact way. If no exact locality is known the area of origin of the sample is coded by an area-map code-number. A retrieval requesting samples from a given area will not only recognize samples coded with the area-map number in question, but also samples coded by geographic co-ordinates within that area.

Collectors — Field numbers given by collectors are preceded by a three-letter code that identifies the collector.

Absolute age — These data will be written in a compact way, that contains some coding.

Furthermore we use a code to identify each institution co-operating in our data bank system, but this is of little concern to the user, and we use a code for the location where the sample is stored, which is optional.

In all other categories of data we use standardized terms instead of symbolic coding. This still requires that a list of accepted terms is kept up-to-date, but firstly such a list is much smaller than a code-book containing a translation of each term into a code and vice versa, and secondly the list of terms needs only to be consulted in case of doubt. This list is kept up-to-date by the machine and a printed copy of it is constantly available to every user.

Checking

All new input should be checked before it is fed into the data bank. In a not-computerized registration system all error checking has to be done by the curator, and this may be a time-consuming task. In a computerized system much of the checking may be done by the machine. Of course the computer cannot recognize whether the right information was given for a sample. E.g. if a sample is of Miocene age, but erroneously the curator codes Oligocene, this error can never be detected by automated checking, as long as both Miocene and Oligocene are valid terms in the system. This aspect will always remain the full responsibility of the user; but the machine can certainly recognize whether a term was written correctly.

We build up (on magnetic tape) a list of all terms used in the entire data bank. For every sample submitted for input all terms used in its description are compared with this list. If the input term is found in the list it is assumed to be correct. If it is not found, the term is printed and the user is requested to check whether it is a correct new term, or whether its absence is due to wrong spelling. So, he is freed from the task of checking all terms, and may confine himself to the checking of terms not yet present.

A second time-saving factor is the following: we submit samples for checking in groups of several hundreds, and try to make these groups as homogeneous as possible: groups of samples from a single locality, a collection of fossils of the same genus or species, etc. In this way many terms used in the description of the samples will be identical throughout the entire input data set, or at least throughout part of it. A new term that occurs in hundred samples of an input data set needs to be checked manually only once. So, here again the checking task is considerably reduced.

The following example illustrates the efficiency of our checking system: In an input data set of 350 samples, the system recognized 3450 terms that had to be checked. Machine checking of these terms recognized 3360 terms as present, and only the remaining 90 terms were printed for manual checking. In this print the terms are accompanied by the registration numbers of the samples in which they were used, so they can easily be found in the input cards. In a growing data bank the amount of manual checking will even decrease, since an evergrowing number of terms will be recognized as present.

After all errors have been removed the new input is loaded into the data bank, and at that moment the new terms are incorporated in the list of terms. This list is always available in printed form for consultation if one is not sure of the orthography of a term to be used in the description of new samples.

There are a few other checking methods in our system, that are fully machine-operated. One of them is whether the new input has been written according to system specifications. Each contravention is explained in printed form, and the user can easily correct these errors.

Another check is whether the registration numbers of new input data are new for the data bank; in the past it has happened from time to time, that a registration number has been used twice for different samples. The computer will recognize these errors and print the double numbers, so that they may be changed before the sample is loaded into the data bank.

```
RGM   173550 REF.:            COORD.:  NE 51 07 009 07 FIELD NR.: VDB159.5 CLAS.: 605        AGE.: 557-557
             POLYGNATHUS NORMALIS MILLER & YOUNGQUIST,1947
             FRASNIAN PALMATOLEPIS GIGAS Z. U
             WILDUNGEN BAD SCHMIDT Q GERMANY
             INV.:VAN DEN BOOGAARD,M.
             POST-SYMP. EXCURS. MARBURG, 1971. FIELD TRIP GUIDEBOOK, P.9

RGM   173551 REF.:            COORD.:  NE 51 07 009 07 FIELD NR.: VDB159.5 CLAS.: 605        AGE.: 557-557
             POLYGNATHUS DECOROSUS STAUFFER,1938
             FRASNIAN PALMATOLEPIS GIGAS Z. U
             WILDUNGEN BAD SCHMIDT Q GERMANY
             INV.:VAN DEN BOOGAARD,M.
             POST-SYMP. EXCURS. MARBURG, 1971. FIELD TRIP GUIDEBOOK, P.9

RGM   173552 REF.:            COORD.:  NE 51 07 009 07 FIELD NR.: VDB159.5 CLAS.: 605        AGE.: 557-557
             POLYGNATHUS PATHOLOGICAL BRUSH-LIKE FORM
             FRASNIAN PALMATOLEPIS GIGAS Z. U
             WILDUNGEN BAD SCHMIDT Q GERMANY
             INV.:VAN DEN BOOGAARD,M.
             POST-SYMP. EXCURS. MARBURG, 1971. FIELD TRIP GUIDEBOOK, P.9

RGM   173553 REF.:            COORD.:  NE 51 07 009 07 FIELD NR.: VDB159.5 CLAS.: 605        AGE.: 557-557
             ANCYRODELLA GIGAS YOUNGQUIST,1947
             FRASNIAN PALMATOLEPIS GIGAS Z. U
             WILDUNGEN BAD SCHMIDT Q GERMANY
             INV.:VAN DEN BOOGAARD,M.
             POST-SYMP. EXCURS. MARBURG, 1971. FIELD TRIP GUIDEBOOK, P.9

RGM   173554 REF.:            COORD.:  NE 51 07 009 07 FIELD NR.: VDB159.5 CLAS.: 605        AGE.: 557-557
             ANCYROGNATHUS ASYMMETRICUS ULRICH & BASSLER,1926
             FRASNIAN PALMATOLEPIS GIGAS Z. U
             WILDUNGEN BAD SCHMIDT Q GERMANY
             INV.:VAN DEN BOOGAARD,M.
             POST-SYMP. EXCURS. MARBURG, 1971. FIELD TRIP GUIDEBOOK, P.9

RGM   173555 REF.:            COORD.:  NE 51 07 009 07 FIELD NR.: VDB159.5 CLAS.: 605        AGE.: 557-557
             ANCYRODELLA CURVATA BRANSON & MEHL,1934
             FRASNIAN PALMATOLEPIS GIGAS Z. U
             WILDUNGEN BAD SCHMIDT Q GERMANY
             INV.:VAN DEN BOOGAARD,M.
             POST-SYMP. EXCURS. MARBURG, 1971. FIELD TRIP GUIDEBOOK, P.9

RGM   173556 REF.:            COORD.:  NE 51 07 009 07 FIELD NR.: VDB159.5 CLAS.: 605        AGE.: 557-557
             ANCYRODELLA SP.
             FRASNIAN PALMATOLEPIS GIGAS Z. U
             WILDUNGEN BAD SCHMIDT Q GERMANY
             INV.:VAN DEN BOOGAARD,M.
             POST-SYMP. EXCURS. MARBURG, 1971. FIELD TRIP GUIDEBOOK, P.9
```

Fig. 2. Sample page from the master catalogue.

Retrieval programmes

We have chosen a system that produces ready catalogues to answer the most frequent retrieval requests without the need of computer processing. It is estimated that far over 90% of the requests can be answered from these catalogues. Only for very special requests a computer run may be needed. In this way the costs of running the data bank are considerably reduced. Each sample is processed in the computer only once, all records for the standard catalogues are then produced, and renewed processing of the same sample is never necessary, unless the sample data need updating (e.g. revision of a species identification). In such a case all data pertaining to the sample are removed from the tapes, and after revision the sample is treated as a new sample for the data bank.

Basically our system produces four catalogues, that may be defined as 1) master catalogue, 2) sort file, 3) cross index catalogue sorted by age, 4) cross index catalogue sorted by classification.

Master catalogue — This catalogue contains *all* information for each sample and most of it is written out in normal understandable words, not in codes. Per sample the data are written in a standard sequence, e.g. first classification, than genus name, species name; on a next line stratigraphic information, and again on a next line locality data, etc. Figure 2 represents a sample page from this master catalogue. This example shows, that the data for each sample start with a line that contains all information that is numeric by nature, or that has been coded. We feel that the few codes used present no serious disadvantage to the user, as compared to the uncoded terms.

The tape containing the master catalogue is built up in numeric sequence of catalogue numbers. At every updating run new records are inserted in the proper numeric sequence. So, at every stage this tape is available to produce a complete print (in numeric sequence of registration numbers) of all the data for all samples present in the data bank. For practical reasons this catalogue is not entirely printed after each updating run. Generally a sample occupies 6 to 7 lines (including a blank line to separate it from the next sample). So, a data bank of 200 000 samples (the size of the RGM collections) would require over 1 200 000 lines (20 000 pages of printing paper) for the master catalogue. Evidently we do not want to print this amount after each updating run.

The programme that checks the new input produces, among other things, a printed copy of the form in which the new samples would appear in the master catalogue, and this print is kept as supplement to the latest printed generation of the master catalogue. After a number of updating runs it may become unpractical to work with these supplements, and then those parts of the catalogue that have been subject to updating are reprinted from the tape, whilst unchanged portions are not reprinted. If, for example, during a certain period updating has only affected registration numbers between 100 000 and 120 000, we may decide to reprint the master catalogue for 100 000 through 120 000, whilst lower and higher numbers are not printed. For these the latest print is still entirely up-to-date.

Sort file — This catalogue contains alphabetic lists of standardized terms arranged by categories. Examples of such categories are: local stratigraphy, locality, collector, etc. Each of these categories is designated by a file number: *local stratigraphy* bears file number 24, *locality names* bears file number 15, etc. These are main files. For some files we thought it practical to make a further subdivision,

FILE NUMBER IS 19870

ACTAEA	113135									
ATERGATIS	87987									
CALAPPA	11932									
CALAPPILIA	11739	11740								
CALLIANASSA	11722	11723	11724	11725	11726	11728	11729	11730	11731	11732
	11733	11734	11931	41642	60476	60571	60642	60682	60733	60783
	167619	167620	167621	167622	167623	167624	167625	167635	167636	
CHLORODIUS	11757									
COENOBITA	11735									
ETISUS	7359									
GALENE	11764									
HEPATISCUS	11741	167626								
LEUCOSIA	11748	11749	11750	11751	11752	113134	167629			
MACROPHTHALMUS	6259	11936	11937	11947	16227	87988	87989	87990	167646	167647
	167648	167649	167650	167651	167652	167653	167654			
MARTINOCARCINUS	11762	11763								
MEDAEUS	11755	11756	167630							

Fig. 3. Sample page from the sort file.

because otherwise the list of terms within the file would become so long, that this might hamper its consultation. For example, for paleontology we made a list of code-figures for the main groups of taxonomy (roughly a code-figure for each phylum). The file number for genus names is 19, the code-figure for Mammalia is 999; so all mammalian genus names in the sort file will be found under subfile number 19999. In the same way all mammalian species names will be in subfile 20999, etc. Within each file or subfile the terms are arranged in alphabetical sequence.

In a print of the sort file each term is followed by a list of the registration numbers of all samples in which the term occurs. So, if one wishes to retrieve all material belonging to the genus *Equus,* one has to search the term *Equus* in file 19999, in order to find the registration numbers of all *Equus* samples in the collection. An example of a sort file print is given in Figure 3.

The sort file may also be used to handle retrieval requests that ask for a combination of terms. If one wishes to know what samples in the collection pertain to *Equus robustus* and were found at the locality Tegelen, this request is fed to a programme that scans the sort file for the terms Tegelen, *Equus,* and *robustus,* and that copies the lists of registration numbers linked to these terms. These three lists are compared, and the matching numbers are copied into a new list. This final list is printed, or, if one wishes so, this list is fed to another programme that reads the master catalogue tape and copies the full information of all *Equus robustus* samples from Tegelen on printing paper.

The same records that are used to build up the sort file, are also used to construct the list of coded terms, that serves to check the orthography of terms when preparing the punched cards for input. In this list the registration numbers of the samples in which the terms occur are left out, which reduces its volume considerably as compared to the sort file (at present the sort file contains about 100 000 records, the code-list only 6000).

An important application of the code-list is, that it is used by the data bank programme to check new input. Before any information is entered into the data bank, all coded information contained in it is checked against the code-list. This check results in a checklist that contains all terms encountered in the new input, that were not yet present in the existing data bank. So the curator handling the new input can easily decide whether an unknown term is a valid new term indeed, or whether its recognition as unknown is due to an error in orthography or some other kind of input error.

Cross index catalogue sorted by age — A combined sort on several criteria, as described above, is a relatively expensive task. We came to the conclusion that costs could be reduced considerably if some of the most frequent retrieval requests could be dealt with by means of ever available cross-index catalogues, so that it would not be necessary to make a computer run for each demand. As the function of such a catalogue would be to give a quick reference to the samples without it being necessary to give full information for every sample, we selected a number of the most important criteria to be entered in it, such as: age, country, locality, local stratigraphy, phylum, genus, species, registration number, storage number, and several other ones. The terms used in the description of a sample are not written in full, but they are truncated at the number of characters that is estimated to be significant. Thus, for example, only the first 20 characters of a genus name are written, species names are truncated after the thirtieth character, and country

CATALOGUE OF AGES 557-557

173433	AUSTR	GROSZER PAL SECTION	FAMENNIAN	SPATHOGNAT	CONODON	GEN. INDET.	SP. INDET.
173434	AUSTR	GROSZER PAL SECTION	FAMENNIAN	SPATHOGNAT	CONODON	GEN. INDET.	SP. INDET.
173435	AUSTR	GROSZER PAL SECTION	FAMENNIAN	SPATHOGNAT	CONODON	GEN. INDET.	SP. INDET.
173436	AUSTR	GROSZER PAL SECTION	FAMENNIAN	SPATHOGNAT	CONODON	GEN. INDET.	SP. INDET.
173438	AUSTR	GROSZER PAL SECTION	FAMENNIAN	SPATHOGNAT	CONODON	SPATHOGNATHODUS	COSTATUS COSTATUS BRANSON,1
173437	AUSTR	GROSZER PAL SECTION	FAMENNIAN	SPATHOGNAT	CONODON	SPATHOGNATHODUS	COSTATUS SPINULICOSTATUS BR
173428	AUSTR	GROSZER PAL SECTION	FAMENNIAN	SPATHOGNAT	CONODON	SPATHOGNATHODUS	SUPREMUS ZIEGLER,1962
173429	AUSTR	GROSZER PAL SECTION	FAMENNIAN	SPATHOGNAT	CONODON	SPATHOGNATHODUS	SUPREMUS ZIEGLER,1962
173430	AUSTR	GROSZER PAL SECTION	FAMENNIAN	SPATHOGNAT	CONODON	SPATHOGNATHODUS	SUPREMUS ZIEGLER,1962
173431	AUSTR	GROSZER PAL SECTION	FAMENNIAN	SPATHOGNAT	CONODON	SPATHOGNATHODUS	SUPREMUS ZIEGLER,1962
173432	AUSTR	GROSZER PAL SECTION	FAMENNIAN	SPATHOGNAT	CONODON	SPATHOGNATHODUS	SUPREMUS ZIEGLER,1962
173408	GERMA	BALLBERG NEAR BALVE	FAMENNIAN	SCAPHIGNAT	CONODON	COELOCERODONTUS ?	SP.
173412	GERMA	BALLBERG NEAR BALVE	FAMENNIAN	SCAPHIGNAT	CONODON	FALCODUS	VARIABILIS SANNEMANN,1955
173427	GERMA	BALLBERG NEAR BALVE	FAMENNIAN	SCAPHIGNAT	CONODON	GEN. INDET.	SP. INDET.
173425	GERMA	BALLBERG NEAR BALVE	FAMENNIAN	SCAPHIGNAT	CONODON	HIBBARDELLA	PLANA HELMS,1959
173426	GERMA	BALLBERG NEAR BALVE	FAMENNIAN	SCAPHIGNAT	CONODON	HIBBARDELLA	SP.
173426	GERMA	BALLBERG NEAR BALVE	FAMENNIAN	SCAPHIGNAT	CONODON	HINDEODELLA	SP.
173415	GERMA	BALLBERG NEAR BALVE	FAMENNIAN	SCAPHIGNAT	CONODON	LIGONODINA	MONODENTATA BISCHOFF & ZIEG
173414	GERMA	BALLBERG NEAR BALVE	FAMENNIAN	SCAPHIGNAT	CONODON	LIGONODINA	PATHOLOGICAL BRUSH-LIKE FOR
173426	GERMA	BALLBERG NEAR BALVE	FAMENNIAN	SCAPHIGNAT	CONODON	LIGONODINA	SP.
173416	GERMA	BALLBERG NEAR BALVE	FAMENNIAN	SCAPHIGNAT	CONODON	NOTHOGNATHELLA	TYPICALIS BRANSON & MEHL,19
173423	GERMA	BALLBERG NEAR BALVE	FAMENNIAN	SCAPHIGNAT	CONODON	OZARKODINA	SP.
173411	GERMA	BALLBERG NEAR BALVE	FAMENNIAN	SCAPHIGNAT	CONODON	PALMATODELLA	DELICATULA BASSLER,1925
173424	GERMA	BALLBERG NEAR BALVE	FAMENNIAN	SCAPHIGNAT	CONODON	PALMATODELLA	UNCA SANNEMANN,1955
173407	GERMA	BALLBERG NEAR BALVE	FAMENNIAN	SCAPHIGNAT	CONODON	PALMATOLEPIS	GLABRA ELONGATA HOLMES,1928
173402	GERMA	BALLBERG NEAR BALVE	FAMENNIAN	SCAPHIGNAT	CONODON	PALMATOLEPIS	GRACILIS GRACILIS BRANSON &
173404	GERMA	BALLBERG NEAR BALVE	FAMENNIAN	SCAPHIGNAT	CONODON	PALMATOLEPIS	HELMST ZIEGLER,1962
173401	GERMA	BALLBERG NEAR BALVE	FAMENNIAN	SCAPHIGNAT	CONODON	PALMATOLEPIS	MAXIMA MUELLER,1956
173403	GERMA	BALLBERG NEAR BALVE	FAMENNIAN	SCAPHIGNAT	CONODON	PALMATOLEPIS	MINUTA SUBSP.NOV. ZIEGLER,1
173405	GERMA	BALLBERG NEAR BALVE	FAMENNIAN	SCAPHIGNAT	CONODON	PALMATOLEPIS	PERLOBATA SCHINDEWOLFI MUEL
173406	GERMA	BALLBERG NEAR BALVE	FAMENNIAN	SCAPHIGNAT	CONODON	PALMATOLEPIS	PERLOBATA SCHINDEWOLFI MUEL
173400	GERMA	BALLBERG NEAR BALVE	FAMENNIAN	SCAPHIGNAT	CONODON	PALMATOLEPIS	RUGOSA TRACHYTERA ZIEGLER,1
173421	GERMA	BALLBERG NEAR BALVE	FAMENNIAN	SCAPHIGNAT	CONODON	POLYGNATHUS	NODOCOSTATUS BRANSON & MEHL
173417	GERMA	BALLBERG NEAR BALVE	FAMENNIAN	SCAPHIGNAT	CONODON	POLYGNATHUS	SP.
173410	GERMA	BALLBERG NEAR BALVE	FAMENNIAN	SCAPHIGNAT	CONODON	PRIONIODINA	SMITHI STAUFFER,1938
173422	GERMA	BALLBERG NEAR BALVE	FAMENNIAN	SCAPHIGNAT	CONODON	SCAPHIGNATHUS	VELIFER ZIEGLER,1959
173413	GERMA	BALLBERG NEAR BALVE	FAMENNIAN	SCAPHIGNAT	CONODON	SCUTULA	SP.
173419	GERMA	BALLBERG NEAR BALVE	FAMENNIAN	SCAPHIGNAT	CONODON	SPATHOGNATHODUS	AMPLUS BRANSON & MEHL,1934
173418	GERMA	BALLBERG NEAR BALVE	FAMENNIAN	SCAPHIGNAT	CONODON	SPATHOGNATHODUS	STABILIS BRANSON & MEHL,193
173420	GERMA	BALLBERG NEAR BALVE	FAMENNIAN	SCAPHIGNAT	CONODON	SPATHOGNATHODUS	WERNERI ZIEGLER,1962
173409	GERMA	BALLBERG NEAR BALVE	FAMENNIAN	SCAPHIGNAT	CONODON	TRIPODELLUS	ROBUSTUS BISCHOFF,1957
173239	PORTU	POMARAO REGION, GRID	FAMENNIAN	POLYGNATHU	CONODON	ANGULODUS	WALRATHI HIBBARD,1927
173321	PORTU	POMARAO REGION, GRID	FAMENNIAN	POLYGNATHU	CONODON	ANGULODUS	WALRATHI HIBBARD,1927

Fig. 4. Sample page from the cross-index catalogue sorted by age.

names after the tenth. The records so formed contain a total of 243 characters that give information on 14 different categories. The data set containing these records can be arranged alphabetically according to different criteria, and two of these arrangments have been chosen as standard catalogues, one of which is the cross index catalogue arranged by age. Its structure is as follows:
all records are put in alphabetical order on the basis of the codes for chronostratigraphy. Within a group of chronostratigraphy (e.g. within the Devonian) the records are arranged alphabetically per country. Within each country the records are alphabetical by locality and this hierarchy goes on with the criteria phylum, genus and species. So, if one wishes to get information on all ostracodes from the Miocene of Dingden in Germany one takes the catalogue and easily finds the wanted information, alphabetically arranged by genus, with the species names alphabetical per genus.

From this alphabetically arranged data set on magnetic tape, one standard print is permanently available. It contains a selection of the data, and not the complete records of the data set, because the record length on tape is 243 whilst the line printer can only handle a maximum of 133 characters per line. In the standard print country names are truncated after the fifth character, although on tape 10 characters are available, stratigraphic information is truncated after the twentieth character although 30 characters are available on tape; other categories like organ and measurements are left out from the standard print.

Figure 4 represents a sample page from the standard print of the cross-index catalogue based on age. Any user requesting a different lay-out (giving more space to one category at the expense of another one, so that the total length of a line does not exceed 132 characters) can have such a print at his disposal at minimal costs.

Cross index catalogue sorted by taxonomy — This catalogue is based on the same records as described above, but the alphabetical arrangement is different. The first criterium is the main group of classification (in most cases this is a phylum, as far as paleontology is concerned). Within a phylum the records are arranged alphabetically by genus, and the hierarchy of alphabetical sorting proceeds with the categories species, country, and locality. A sample page of this catalogue is printed in Figure 5. Different print lay-outs may be produced cheaply.

The two cross-index catalogues are possibly the most important entrances to our collections. Is is estimated that, even though the data are truncated after a number of characters to make a column organization possible, they will practically satisfy all demands. If a basically different hierarchy of criteria for the alphabetical sort is required, such a data set may be produced from one of the standard cross-index data sets at reasonable costs. If the need is felt, such a third cross-index catalogue, once it has been made, may be introduced as a standard product in the basic data bank programme, and so be kept up-to-date at every new input.

Labels

The same punched cards that are used for loading a sample into the data bank can be used to produce labels that may be put into the boxes with the samples. Such labels contain a selection of the data present in the master catalogue. For a number of reasons the printing of labels was not incorporated in the standard load run:

167787	ACANTHASTREA	ECHINATA DANA,1848	INDON	SUNGAI GELINGSEH SAN	963	963	BALIKPAPAN LAYERS U.
167779	ACANTHOCYATHUS	MALAYICUS GERTH,1923	INDON	MUARA KOBUN SANGKULI	963	963	
167682	ACANTHOCYATHUS	SPINOSUS UMBGROVE,1950	INDON	KENDENG HILLS JAVA -	983	983	PUTJANGAN FM.
167683	ACANTHOCYATHUS	SPINOSUS UMBGROVE,1950	INDON	KENDENG HILLS JAVA	983	983	PUTJANGAN FM.
167684	ACANTHOCYATHUS	SPINOSUS UMBGROVE,1950	INDON	KENDENG HILLS JAVA	983	983	PUTJANGAN FM.
167685	ACANTHOCYATHUS	SPINOSUS UMBGROVE,1950	INDON	KENDENG HILLS JAVA	983	983	PUTJANGAN FM.
167686	ACANTHOCYATHUS	SPINOSUS UMBGROVE,1950	INDON	KENDENG HILLS JAVA	983	983	PUTJANGAN FM.
167687	ACANTHOCYATHUS	SPINOSUS UMBGROVE,1950	INDON	KENDENG HILLS JAVA	983	983	PUTJANGAN FM.
167688	ACANTHOCYATHUS	SPINOSUS UMBGROVE,1950	INDON	KENDENG HILLS JAVA	983	983	PUTJANGAN FM.
167689	ACANTHOCYATHUS	SPINOSUS UMBGROVE,1950	INDON	KENDENG HILLS JAVA	983	983	PUTJANGAN FM.
167862	ACRHELIA	HORRESCENS DANA,1848 SYN.CF	INDON	KEPULAUAN TOGIAN , S	999	999	
167862	ACRHELIA	SEBAE EDWARDS & HAIME,1849	INDON	KEPULAUAN TOGIAN , S	999	999	
167906	ACROPORA	CF. HEBES DANA,1846	INDON	BAAI VAN BATAVIA TEL	999	999	
125829	ACROPORA	DUNCANI REUSS,1866	INDON	NIAS	980	980	
167574	ACROPORA	FENNEMAI GERTH,1921	INDON	TJADASNGAMPAR JAVA	963	963	
167803	ACROPORA	FENNEMAI GERTH,1921 (SEE GE	INDON	GUNUNG BATU ANTICLIN	963	963	BALIKPAPAN LAYERS U.
167907	ACROPORA	PALIFERA LAMARCK,1801	INDON	BAAI VAN BATAVIA TEL	999	999	RHODOPHYCEAE FACIES
167852	ACROPORA	SP.	INDON	BAAI VAN BATAVIA TEL	999	999	
167905	ACROPORA	SP.	INDON	BAAI VAN BATAVIA TEL	999	999	
125828	ACROPORA	SP.	INDON	NIAS	980	980	
125830	ACROPORA	SP.	INDON	NIAS	980	980	
125831	ACROPORA	SP.	INDON	NIAS	980	980	
125838	ALVEOPORA	DAEDALEA FORSKAL,1775	INDON	NIAS	980	980	
167672	ALVEOPORA	POLYACANTHA REUSS,1866	INDON	GUNUNG LINGGAPADANG	967	967	
167673	ALVEOPORA	POLYACANTHA REUSS,1866	INDON	GUNUNG LINGGAPADANG	967	967	
167797	AMPHELIA	ALTERNANS GERTH,1923	INDON	GUNUNG BATU ANTICLIN	963	963	BALIKPAPAN LAYERS U.
167798	AMPHELIA	ALTERNANS GERTH,1923	INDON	SUNGAI GELINGSEH SAN	963	963	BALIKPAPAN LAYERS U.
143065	ANABACIA	SP.	ARGEN	PORTEZUELO ANCHO MEN	753	753	
143066	ANDENIPORA	LIASICA GERTH,1926	ARGEN	CATAN-LIL (N.OF), NE	753	753	
143067	ANDENIPORA	LIASICA GERTH,1926	ARGEN	CERRO LOTENA (10 KM	753	753	
167538	ANISOCOENIA	CRASSISEPTA REUSS,1866	INDON	JUNGHUHN O JAVA	963	963	
167544	ANISOCOENIA	SP. ACCORDING TO GERTH(?)	INDON	JUNGHUHN P JAVA	963	963	
167791	ANISOCOENIA	VARIABILIS GERTH,1923	INDON	GUNUNG BATUTA AT SUN	963	963	
167792	ANISOCOENIA	VARIABILIS GERTH,1923	INDON	GUNUNG BATUTA AT SUN	963	963	
167543	ANTHEMIPHYLLIA	PATELLA GERTH,1921	INDON	SEDAN JAVA	963	963	REMBANG BEDS
167541	ANTILLIA	INFUNDIBULIFORMIS GERTH,192	INDON	SONDE JAVA	967	967	SONDE BEDS
167542	ANTILLIA	ORIENTALIS GERTH,1921	INDON	NGEMBAK JAVA (BORING	963	963	
167784	ANTILLIA	ORIENTALIS GERTH,1923	INDON	GUNUNG BATU ANTICLIN	963	963	BALIKPAPAN LAYERS U.
143059	ASTROCOENIA	AFF. REGULARIS FROMENTEL,18	ARGEN	CERRO SALADA (10 KM	773	773	QUINTUCO FM. VALANG
143056	ASTROCOENIA	CF. TRIBOLETI KOBY,1897	ARGEN	CERRO LOTENA (5 KM E	773	773	APTIAN AGRIO FM.
143057	ASTROCOENIA	CF. TRIBOLETI KOBY,1897	ARGEN	SIERRA OF LA VACA MU	773	773	APTIAN AGRIO FM.
143055	ASTROCOENIA	COLLICULOSA TRAUTSCHOLD,188	ARGEN	NEUQUEN	773	773	
143056	ASTROCOENIA	COLLICULOSA TRAUTSCHOLD,188	ARGEN	NEUQUEN	773	773	
167563	BALANOPHYLLIA	OPPENHEIMI FELIX,1913	INDON	SONDE JAVA	967	967	SONDE BEDS
167564	BALANOPHYLLIA	OPPENHEIMI FELIX,1913 (UMBG	INDON	GARUNG JAVA	967	967	SONDE BEDS

Fig. 5. Sample page from the cross-index catalogue sorted by taxonomy.

One of these reasons is that for many samples labels are not at all required. No one will think of joining labels to slides with microfossils, or thin sections of rock samples.

Another reason is that a different kind of lay-out may be wanted for different collections. For one collection it may be important to state the collector's name on the label, for another collection, however, this may be of no use.

For these reasons the label programme was detached from the standard load run; every curator now has the choice to re-run his input cards for a label print run, and for every such run he may adapt the lay-out to the specific demands of the collection in question. The technical execution of this adaptation has been made so simple, that one need not have any knowledge of the programme language, and still can make the necessary alterations to achieve the lay-out wanted. If necessary photostatic reductions of the labels can be made in order to make them fit into the boxes.

Outdated information

As described above all information is checked before it is loaded into the data bank. This checking routine examines whether the new input is technically correct, viz. whether there are any errors in orthography, card sequence, file numbers, etc., but, of course, the computer can impossibly check whether the identification of a fossil is correct, or whether it actually was collected at the stated locality. Such errors may still occur in the checked input and they should be corrected if they are detected. Furthermore, changes may be necessary that are not due to erroneous input, but to changing scientific concepts, or more detailed study: a species name may be revised, or it may be necessary to add measurements or other scientific information to the data concerning a sample that had hitherto not been studied in detail.

In all these cases where data should be deleted or altered, we used the following standard procedure: we take the original card input from the stored card decks, process these cards with the normal loading programme, but then the resulting records for the various data sets that constitute our data bank are not merged with the data bank, as is done in a normal loading run; they are, on the contrary, deleted from the data bank. This procedure guarantees that no information pertaining to the deleted samples remains unobserved in one of the data sets.

After this deleting run, the cards are corrected as required and in the next input run they will be processed as if they constituted an entirely new sample.

In some cases the old data may simply be deleted, and need not be conserved in any form. For example a sample was provisionally entered into the data bank as Equidae gen. indet. sp. indet. After the sample has been studied it has to be entered as *Equus robustus*. There is no need to conserve the information gen. indet. sp. indet., so the old data are deleted and the new data are entered.

However, it may also occur that a sample was entered as *Equus caballus* stating that it was published under this name by Fox in 1969 in his thesis on fossil horses. The material has been restudied and it appears to belong to *Equus robustus*. It should now be entered under the new name, but the previously published identification should not be lost, as we should be able to find it under the previous name, when a scientist asks us for the *Equus caballus* material described by Fox. Our system makes it possible to enter the sample (after it has first been

```
RGM   125155 REF.:              AREA: 5140000001 FIELD NR.:          CLAS.: 605          AGE: 550-570
              CONODONTS   GNATHODUS SEMIGLABER BISCHOFF,1957
                   GNATHODUS CUNEIFORMIS MEHL & THOMAS,1947
                   GNATHODUS CF. CUNEIFORMIS MEHL & THOMAS,1947
                   GNATHODUS DELICATUS BRANSON & MEHL,1938
                   GNATHODUS PUNCTATUS COOPER,1939
                   POLYGNATHUS COMMUNIS COMMUNIS BRANSON & MEHL,1934
                   POLYGNATHUS LONGIPOSTICUS BRANSON & MEHL,1934
                   POLYGNATHUS CF. COMMUNIS COMMUNIS BRANSON & MEHL,1934
                   POLYGNATHUS INORNATUS BRANSON,1934
                   POLYGNATHUS COMMUNIS CARINA HASS,1959
                   SPATHOGNATHODUS COSTATUS COSTATUS BRANSON,1934
                   PSEUDOPOLYGNATHUS TRIANGULUS PINNATUS VOGES,1959
                   PSEUDOPOLYGNATHUS DENTILINEATUS BRANSON,1934
                   PSEUDOPOLYGNATHUS MULTISTRIATUS MEHL & THOMAS,1947
                   SIPHONODELLA OBSOLETA HASS,1959
                   SIPHONODELLA SP.
                   SIPHONODELLA SP.A VAN ADRICHEM BOOGAERT,1967
                   ELICTOGNATHUS LACERATUS BRANSON & MEHL,1934
              VISEAN ALBA FM. SCALIOGNATHUS ANCHORALIS Z.
              RIO ESLA AREA LEON SPAIN
              INV.:VAN ADRICHEM BOOGAERT 1957 DEVONIAN AND LOWER CARBONIFEROUS CONODONTS OF THE CANTABRIAN
              MOUNTAINS (SPAIN) AND THEIR STRATIGRAPHIC APPLICATION,THESIS,LEIDEN
              SAMPLE AG.6 SEE ARCHIVES NR.6120004

RGM   125156 REF.:              AREA: 5140000001 FIELD NR.:          CLAS.: 605          AGE: 550-570
              CONODONTS   GNATHODUS SEMIGLABER BISCHOFF,1957
                   GNATHODUS CUNEIFORMIS MEHL & THOMAS,1947
                   GNATHODUS DELICATUS BRANSON & MEHL,1938
                   GNATHODUS TYPICUS COOPER,1939
                   GNATHODUS PUNCTATUS COOPER,1939
                   POLYGNATHUS COMMUNIS COMMUNIS BRANSON & MEHL,1934
                   POLYGNATHUS CF. COMMUNIS COMMUNIS BRANSON & MEHL,1934
                   POLYGNATHUS INORNATUS BRANSON,1934
                   POLYGNATHUS COMMUNIS CARINA HASS,1959
                   SPATHOGNATHODUS COSTATUS SPINULICOSTATUS BRANSON,1934
                   PSEUDOPOLYGNATHUS TRIANGULUS PINNATUS VOGES,1959
                   PSEUDOPOLYGNATHUS DENTILINEATUS BRANSON,1934
                   SIPHONODELLA SP.
              TOURNAISIAN VISEAN ERMITA FM. BALEAS FM. ? SCALIOGNATHUS ANCHORALIS Z.
              RIO ESLA AREA LEON SPAIN
              INV.:VAN ADRICHEM BOOGAERT 1967 DEVONIAN AND LOWER CARBONIFEROUS CONODONTS OF THE CANTABRIAN
              MOUNTAINS (SPAIN) AND THEIR STRATIGRAPHIC APPLICATION,THESIS,LEIDEN
              SAMPLE OL.1 SEE ARCHIVES NR.6120004
```

Fig. 6. Example of the print of composite samples in the master catalogue.

completely deleted) as: *Equus robustus,* published by Bird, 1974, previously published as *Equus caballus* by Fox, 1969. In the various catalogues the sample can be found under *Equus caballus* as well as under *Equus robustus;* the master catalogue states that *caballus* is the old name and *robustus* the new one. For each sample any number of outdated names can be conserved in this way, if the need is felt.

Composite samples

The majority of the samples consists of one specimen to which a fossil name or a rock name may be attributed. However, in a large minority of the samples the situation is more complex: we may have a rock slab displaying various species of fossils, a rock sample may contain a number of minerals, that should all be registered, or a slide may contain a collection of various species of microfossils. In such cases each species name, rock name, or mineral name should be independently retrievable. We solved this problem in such a way, that any number of independent names may occur in the input data cards of a sample. There is no limit to the number of cards used for the description of a sample, apart from the limit that is imposed by the requirement that the data should be useful. Technically it is possible to incorporate in the input a full description of the locality in which a specimen was found, or even the complete text of the publication in which it was described, but such an approach takes large amounts of tape storage, and serves no practical purpose. So, in practice, the input is kept as short and compact as possible, and made as comprehensive and detailed as necessary.

Examples of these possibilities are given in Figure 6, where the full title of a publication is given, and where more than one species is listed under the same registration number.

Costs

This chapter can be divided into three parts that are more or less independent, viz. the costs of input, the costs of output, and other costs.

Costs of input

The costs for entering new samples into the data bank are related to two factors: the number of samples in the input data set, and the number of samples already present in the data bank.

The input data cards are analyzed by the programme and this results in records for each of the different catalogues supported by our registration system. This phase takes about 0.1 to 0.2 seconds of CPU-time per sample (CPU = central processing unit; CPU-time is the time during which the programme is active in the computer. It constitutes one of the main factors that determine the costs of a computer run). After completion of this step, the coded terms are sorted alphabetically and then checked against the existing list of coded terms. At present our data bank contains only 10 000 samples, resulting in a code-list of 6000 terms. Costs of the checking operation are almost neglectable. In a very large data bank

RGM 142901 FIELD NR: AGE:557-557 PTYCHOMALETOECHIA CF. GONTHIERI GOSSELET,1887 PL.1, FIG.3 CREMENES LST. FRASNIAN- FAMENNIAN PICO AGUASALIO ,5 KM SOUTH- EAST OF CREMENES;LEON SPAIN COL.:WESTBROEK,P. INV.: WESTBROEK,P. 1964,LEIDSE GEOL. MEDED.,30:243-252	RGM 142902 FIELD NR: AGE:557-557 PTYCHOMALETOECHIA CF. GONTHIERI GOSSELET,1887 PL.1, FIG.4 CREMENES LST. FRASNIAN- FAMENNIAN PICO AGUASALIO ,5 KM SOUTH- EAST OF CREMENES;LEON SPAIN COL.:WESTBROEK,P. INV.: WESTBROEK,P. 1964,LEIDSE GEOL. MEDED.,30:243-252	RGM 142903 FIELD NR: AGE:557-557 PTYCHOMALETOECHIA CF. GONTHIERI GOSSELET,1887 PL.1, FIG.5 CREMENES LST. FRASNIAN- FAMENNIAN PICO AGUASALIO ,5 KM SOUTH- EAST OF CREMENES;LEON SPAIN COL.:WESTBROEK,P. INV.: WESTBROEK,P. 1964,LEIDSE GEOL. MEDED.,30:243-252	RGM 142904 FIELD NR: AGE:557-557 PTYCHOMALETOECHIA CF. GONTHIERI GOSSELET,1887 PL.1, FIG.6 CREMENES LST. FRASNIAN- FAMENNIAN PICO AGUASALIO ,5 KM SOUTH- EAST OF CREMENES;LEON SPAIN COL.:WESTBROEK,P. INV.: WESTBROEK,P. 1964,LEIDSE GEOL. MEDED.,30:243-252
RGM 142905 FIELD NR: AGE:557-557 PTYCHOMALETOECHIA CF. GONTHIERI GOSSELET,1887 PL.2, FIG. CREMENES LST. FRASNIAN- FAMENNIAN PICO AGUASALIO ,5 KM SOUTH- EAST OF CREMENES;LEON SPAIN COL.:WESTBROEK,P. INV.: WESTBROEK,P. 1964,LEIDSE GEOL. MEDED.,30:243-252	RGM 142906 FIELD NR: AGE:557-557 PTYCHOMALETOECHIA CF. GONTHIERI GOSSELET,1887 PL.2, FIG.2 CREMENES LST. FRASNIAN- FAMENNIAN PICO AGUASALIO ,5 KM SOUTH- EAST OF CREMENES;LEON SPAIN COL.:WESTBROEK,P. INV.: WESTBROEK,P. 1964,LEIDSE GEOL. MEDED.,30:243-252	RGM 142907 FIELD NR: AGE:557-557 PTYCHOMALETOECHIA CF. GONTHIERI GOSSELET,1887 FOTO CREMENES LST. FRASNIAN- FAMENNIAN PICO AGUASALIO ,5 KM SOUTH- EAST OF CREMENES;LEON SPAIN COL.:WESTBROEK,P. INV.: WESTBROEK,P. 1964,LEIDSE GEOL. MEDED.,30:243-252	RGM 142908 FIELD NR: AGE:557-557 PTYCHOMALETOECHIA CF. GONTHIERI GOSSELET,1887 FOTO CREMENES LST. FRASNIAN- FAMENNIAN PICO AGUASALIO ,5 KM SOUTH- EAST OF CREMENES;LEON SPAIN COL.:WESTBROEK,P. INV.: WESTBROEK,P. 1964,LEIDSE GEOL. MEDED.,30:243-252
RGM 142909 FIELD NR: AGE:557-557 PTYCHOMALETOECHIA CF. GONTHIERI GOSSELET,1887 FOTO CREMENES LST. FRASNIAN- FAMENNIAN PICO AGUASALIO ,5 KM SOUTH- EAST OF CREMENES;LEON SPAIN COL.:WESTBROEK,P. INV.: WESTBROEK,P. 1964,LEIDSE GEOL. MEDED.,30:243-252.	RGM 142910 FIELD NR: AGE:557-557 PTYCHOMALETOECHIA CF. GONTHIERI GOSSELET,1887 FOTO CREMENES LST. FRASNIAN- FAMENNIAN PICO AGUASALIO ,5 KM SOUTH- EAST OF CREMENES;LEON SPAIN COL.:WESTBROEK,P. INV.: WESTBROEK,P. 1964,LEIDSE GEOL. MEDED.,30:243-252	RGM 142911 FIELD NR: AGE:557-557 PTYCHOMALETOECHIA CF. GONTHIERI GOSSELET,1887 FOTO CREMENES LST. FRASNIAN- FAMENNIAN PICO AGUASALIO ,5 KM SOUTH- EAST OF CREMENES;LEON SPAIN COL.:WESTBROEK,P. INV.: WESTBROEK,P. 1964,LEIDSE GEOL. MEDED.,30:243-252	RGM 142912 FIELD NR: AGE:557-557 PTYCHOMALETOECHIA CF. GONTHIERI GOSSELET,1887 FOTO CREMENES LST. FRASNIAN- FAMENNIAN PICO AGUASALIO ,5 KM SOUTH- EAST OF CREMENES;LEON SPAIN COL.:WESTBROEK,P. INV.: WESTBROEK,P. 1964,LEIDSE GEOL. MEDED.,30:243-252
RGM 142913 FIELD NR: AGE:557-557 PTYCHOMALETOECHIA CF. GONTHIERI GOSSELET,1887 FOTO CREMENES LST. FRASNIAN- FAMENNIAN PICO AGUASALIO ,5 KM SOUTH- EAST OF CREMENES;LEON SPAIN COL.:WESTBROEK,P. INV.: WESTBROEK,P. 1964,LEIDSE GEOL. MEDED.,30:243-252	RGM 142914 FIELD NR: AGE:557-557 PTYCHOMALETOECHIA CF. GONTHIERI GOSSELET,1887 FOTO CREMENES LST. FRASNIAN- FAMENNIAN PICO AGUASALIO ,5 KM SOUTH- EAST OF CREMENES;LEON SPAIN COL.:WESTBROEK,P. INV.: WESTBROEK,P. 1964,LEIDSE GEOL. MEDED.,30:243-252	RGM 142915 FIELD NR: AGE:557-557 PTYCHOMALETOECHIA CF. GONTHIERI GOSSELET,1887 FOTO CREMENES LST. FRASNIAN- FAMENNIAN PICO AGUASALIO ,5 KM SOUTH- EAST OF CREMENES;LEON SPAIN COL.:WESTBROEK,P. INV.: WESTBROEK,P. 1964,LEIDSE GEOL. MEDED.,30:243-252	RGM 142916 FIELD NR: AGE:557-557 PTYCHOMALETOECHIA CF. GONTHIERI GOSSELET,1887 FOTO CREMENES LST. FRASNIAN- FAMENNIAN PICO AGUASALIO ,5 KM SOUTH- EAST OF CREMENES;LEON SPAIN COL.:WESTBROEK,P. INV.: WESTBROEK,P. 1964,LEIDSE GEOL. MEDED.,30:243-252

Fig. 7. Examples of labels produced by the computer.

CPU-time for this checking may increase, but we estimate that it will not easily exceed ten seconds.

The input data are then screened to check whether they contain registration numbers that were already allotted to other samples previously entered into the data bank. The time involved in this check is mainly dependent upon the size of the data bank, but will not exceed ten seconds of CPU-time.

The master catalogue is printed at a speed of about 300 samples per second CPU-time.

Sorting the cross-index catalogue by age and printing it, takes about 1 second per 500 samples, and the same goes for the cross-index catalogue sorted by classification. As an input data set generally contains up to 1000 samples, these operations are not very expensive. The printed results are used for manual checking, and then kept as supplements to the existing catalogues.

These are the steps that constitute a complete check run of a new input data set. The data bank contains about 10 000 samples at this moment, and a check run takes about 1 minute of CPU-time for 300 samples. We estimate that in a data bank of 200 000 samples the same check run would take about 5% more CPU-time.

A part of the CPU-time is needed for the basic programme functions, which is illustrated by the fact that processing only one sample would take about 20 seconds. This programme overhead represents a cost factor that is independent of the size of the input data set. Costs per sample decrease if larger data sets are submitted for input: a set of 700 samples took 115 seconds, and a set of 1100 took 155 seconds.

CPU-time is the most important factor in the costs of a computer run. So, it is essential that the amount of time for a run should not increase too much in a growing data bank. Other factors are the number of *read* or *write* instructions, the number of times a new tape has to be set up, the amount of printing paper, etc. Each computer institution may put different values on these factors. In our present situation a check run costs about Hfl. 90.— for 300 samples, Hfl. 160.— for 700 samples, and Hfl. 230.— for 1100 samples. Costs per sample decrease along a curved line with the increasing size of the input data set. Little gain is to be expected from data sets of more than 1000 samples, so we try to submit 1000 samples at a time for checking and loading.

After completion of the check run, the user reads the resulting print to detect eventual errors in the input. In practice there are always some samples that contain erroneous data. If most of the samples contain errors it may be wise to re-run the entire input data set after the necessary corrections have been made. In this case each of the samples in the input is twice submitted for checking, and costs for this set will be doubled. However, in most cases we find only a few percent of erroneous samples in the input data set, and in this situation we have a special programme available that deletes all records pertaining to faulty samples from the tape where the results of the check run are stored. Such a deletion programme takes only a few seconds of CPU-time. Costs are about Hfl. 20.— to 40.—.

The remaining faultless results are then loaded into the data bank tapes. For this loading we use very simple programmes that have no other function than merging two alphabetically arranged data sets into one great data set identically arranged.

At present these merging operations take between 1.5 second for the smallest

and 10 seconds for the largest data set. The total load run takes 30 seconds of CPU-time or about Hfl. 30.— plus another Hfl. 30.— for technical operations like tape set-up, etc. In a data bank of 100 000 samples this would amount to 300 seconds or Hfl. 300.— plus Hfl. 30.— for tape set-up.

We are now developing a programme that will roughly cut CPU-time for these merging operations by 50%, but felt it not appropriate, in view of the low costs so far, to give to this development a high priority.

Thus, the total costs of data bank input are constituted from the three parts described above: check run, deletion of errors (if any), and load run. The costs of the check run depend mainly on the size of the input data set, and only to a very small extent upon the size of the existing data bank; the costs of the deletion run are so low that they will hardly influence total costs; the costs of the load run depend mainly on the size of the data bank, and only to a small extent on the size of the input set, as the latter is generally small in comparison with the existing data (and tends to become relatively smaller, as the data bank keeps growing).

At this moment we are loading each input data set after checking, which brings total costs (check, delete, and load) at about Hfl. 300.— for 1000 samples, or Hfl. 0.30 per sample. This same procedure in a data bank of 100 000 samples would amount to about Hfl. 0.60 per sample. There are several solutions to avoid this increase in costs, one of which is the following: a data set of 1000 samples is checked, erroneous samples, if any, are deleted, and the results are stored on tape. A next data set of 1000 samples is treated in the same way, and so on, until a total of 5000 samples is ready for loading. Then all these are added to the data bank in a single load run. This would amount to about Hfl. 1000.— for five check runs, plus Hfl. 500.— for one load run, including a reasonable amount of overhead. The total costs per sample would then be Hfl. 0.30 in a data bank of 100 000 samples, and this would be Hfl. 0.40 in a data bank of 200 000 samples, etc.

At these costs the user will have at his diposal almost any kind of catalogue or cross-index he may want. Production of labels to be put in the boxes with the specimens is left out from the basic processing, as described before. A special programme produces these labels, if wanted, at a cost of Hfl. 0.15 per sample, and this figure is going to be reduced considerably through programme optimalization.

A special case of input costs is updating of data for a sample already present in the data bank. The standard procedure is to remove all data pertaining to the sample from the data bank tapes, correct the sample data, and submit it for check and load as a new sample. Generally the number of samples in a deleting run will be small, so, costs depend entirely on the size of the data bank, and weigh heavy upon each of the deleted samples; they are estimated to be equal to the costs of a load run, which means that they amount to Hfl. 0.60 at present and Hfl. 3.30 per sample in a data bank of 100 000 samples, if the deleting run contains 100 erroneous samples. If only one sample were submitted these figures would be Hfl. 60.— and Hfl. 330.— respectively, so it is evident that we should save up the samples that need updating, instead of processing each one as soon as it is signalled.

Furthermore we think, that these deleting runs will not occur very frequently, so they will not have too much influence on the total running costs of the data bank. Once a sample has been deleted, and its data have been corrected, the

costs for reloading it are the same as for every new sample.

In a number of cases updating may be simplified: if the erroneous information affects only one of the catalogues, and has no influence (or is irrelevant) to the other ones, it may be possible to execute the corrections in this single catalogue without deleting and reloading the entire sample. This might reduce costs considerably.

All in all, deleting costs are not so specifically related to the number of deleted samples. They are better regarded as general running costs for the data bank, and their total amount is determined by the frequency of the deleting runs, so, by the quality of the data bank information and by organizational efficiency. They be best expressed as costs per year; it is not yet possible to make an estimate.

Cost of output

Once a sample has been entered into the data bank, the costs of the different kinds of output are almost neglectable. Printing a copy of the master catalogue costs about Hfl. 1.— per 100 lines (a sample may generally take 6 or 7 lines; a composite sample like a slide containing twenty species may take about 25 lines, i.c. 1 line per species plus some lines for stratigraphy, locality, etc.). Costs for printing one of the cross-index catalogues are at about the same level. Printing the sort file is slightly more expensive.

A great advantage of our system is, that most of the requests to the data bank can be answered by catalogues already existing in print. Only in rare cases it may be necessary to make a special computer run to answer a special question.

Other costs

The figures mentioned so far include processing costs only. No account has been made for costs of personnel. Such costs include mainly three factors:

1. Man-hours involved in writing, testing, and improving the programmes.
2. Man-hours for operating the data bank, preparing programmes for special output requests, submitting data for input, maintenance of the programmes, and error tracing in case of failure.
3. Man-hours for making the punch documents and punched cards from the sample data.

The first category should be considered as a long-term investment, together with the actual computer costs for the test runs. Institutions co-operating in our data bank system will have to pay an overhead on the basic input and output costs, to cover part of this investment, and they will also have to cover part of the operating costs.

Preparing sample data for input, is a task that is required by every registration system, computerized or not. We decided to have the punched cards produced by the curator making the registration. Instead of writing the data on a typewriter, he writes it on a punching-machine. This constitutes a considerable gain of time, once he has acquired some experience, and the risk of introducing errors while copying information is eliminated. The punching machine offers the facility to duplicate automatically those data that are identical in a number of consecutive samples.

All in all we estimate that our computerized registration system presents a considerable gain of time that otherwise would be spent in copying information,

keeping several card-indexes up-to-date, making labels to go with the samples, etc. So, the curator will save time in the registration of new samples, and we hope this time will be used to transfer other samples from the old registration system to the new one. We estimate that in this way it should be possible to handle our entire collection of 200 000 samples in about five years, assuming that the old data are copied without any revision whatsoever.

Technical details

Programme language

The checking programme is written in Fortran IV, and compiled with the G compiler; we are now recompiling with the H compiler, which will reduce processing costs through optimalization. Some parts of the checking programme and most of the output routines are written in Assembler language, or we are using standard IBM utilities. It is planned to translate in the future the most time-consuming parts of the input programme from Fortran into Assembler language, in order to achieve a further reduction of processing time and costs.

Hardware

We have at our disposal an IBM 370/158 with time-sharing facilities, installed at the *Centraal Rekeninstituut* of the University of Leiden. A standard input run takes 256 K of hard core, three tape drives for 2400 ft tapes, and about 25 cylinders of a 3330 disk, to be used for temporary data sets. With these facilities it is possible to process over 1000 samples in each run (a larger number of samples per run could be processed by increasing the space on the 3330 disk proportionally).

Through adaptation of the programme it is possible to work in a 168 K area, but this would increase processing costs. As our machine operates under virtual storage, the amount of core required is of no concern, and we chose the most efficient organization with 256 K (a larger amount of core is possible, but would not increase efficiency).

The data bank information resides on eight magnetic tapes. During each updating run the contents of these eight tapes are copied onto eight parallel tapes, and in the meantime merged with the new input. So, at every moment there are two generations of each data set available, and updating switches from one generation to the other. The set of tapes that serves as input for one updating run will be the output set for the next run. If an updating run fails for some reason, it may be corrected and repeated, as the input generation remains available unaltered.

Furthermore the contents of the latest generation is dumped from time to time on a third set of eight tapes, so that a hardware failure during an updating run will not make it necessary to reprocess the entire data bank. If so required, we may restart the data bank from the last dumped generation. This, plus a tape containing the programme modules and a tape containing the processed results of the check run, brings the total amount of tapes in use at 26. At this moment the

tapes are written with a density of 1600 BPI (bits per inch). At this density one set of tapes is sufficient to contain a data bank of 70 000 samples. When the data bank exceeds this figure, it will be necessary to divide some of the data sets over more than one tape. However, it is expected that, long before we reach this limit of 70 000, the tapes will be written with a density of 6250 BPI, and at this density we expect that the tapes now in use can contain the entire information of our present collection of 200 000 samples.

Co-operation with other institutions

Apart from registering our own collections, we plan to make our programme available to other institutions wishing to benefit from it. For each co-operating institution we will set up an entirely separated data bank, that will use its own tapes and will not be merged with our collection data. It will, however, use the same list of standard terms, so that the description of data will be fully compatible among these data banks. In this way we might register the collections of all universities, provincial and municipal musea, and also of private collectors in our country, thus constituting a national registration system for all geological collections.

When this kind of co-operation has been established, we will be able to give detailed information on all geological material in the Netherlands. On special request it will be possible to make a combined retrieval through all the data banks we support, and have the results merged according to any wanted criteria.

There are no fundamental objections to co-operation with foreign countries, but in such a case there may be technical problems, such as transportation of punched cards and printed results, regular contacts with the co-operating institution, etc. Anyway such a co-operation will not be hampered by language problems, as we decided to use English terminologies for all standardized information, apart from those data that are typically linked to local situations, like for example local stratigraphy and local stratigraphic names; in these cases we use the terms used in the country of origin of the sample, instead of trying to translate them into English.

Manuscript received 28 February 1975.