PROC. N.S. INST. SCI. (1989)
Volume 39, Part 3, pp. 79-98

ASSEMBLY AND UTILIZATION OF A DATABASE OF FUNGAL CULTURES HELD IN CANADIAN COLLECTIONS[†].

D. BREWER, K SCHWARTZENTRUBER AND A. TAYLOR*
National Research Council of Canada,
Institute for Marine Biosciences, Halifax, B3H 3Z1

A. EVANS
Dalhousie University,
Computer Center, Halifax, B3H 4H8

Descriptions of holdings of live cultures in ten Canadian collections of fungi have been collated and assembled as an indexed database. The data include the binomial names of the organisms, the date and place of their isolations, the substrate on which they were found, their mode of preservation, human pathology and ability to produce toxic metabolites. Each entry on the file has a unique accession number which indicates the collection where the organism is to be found. This file is indexed with respect to accession number and organism name and can therefore be searched efficiently for a particular organism, genus or species using simple commands. Access to the system, assembled on the Dalhousie University computer, has been freely available since 1987. At the moment it is not on line but is available to the public through the services of the Librarian of the Institute of Science. In four years we have had no evidence of abuse of the system by the public.

Les descriptions de dix collections canadiennes de cultures vivantes de mycètes ont été collationées et assemblées en base de données indexée. Les données comprennent les noms binominaux des organisms, la date et le lieu de leur isolement, la couche inférieure où ils ont été découverts, la manière de préservation, la pathologie humaine, et la capacité de produire des métabolites toxiques. Chaque inscription dans le fichier a un numero unique d'accession qui indique la collection où l'organisme est logé. Ce fichier est indexé quant au numero d'accession ainsi qu'au nom de l'organism; on peut alors le fouiller, avec des ordres minimes, pour un organisme particulier, un genre ou une espèce. L'accès à ce système, disposé sur l'ordinatuer à l'Université Dalhousie, est disponible sans frais depuis 1987. A l'heure actuelle, le système n'est pas branché, mais il est disponible par l'entremise des services de la bibliothécaire de l'Institute des Sciences. Pendant quatre années on a eu aucune évidence d'abus de ce système par le public.

Introduction

The Latin binomial names for living organisms, compiled since 1895 in the Index Kewensis are internationally accepted (among other things) as the code for retrieval of information about self-reproductive systems whether maintained as preserved specimens or as collections of viable species. Obviously such collections have many uses and are often assembled by students interested in groups of organisms having particular morphological or physiological characteristics. As Ainsworth (1961) has pointed out, such collections retain their value long after the reasons for their assembly have been submerged in printed scientific archives and after the death of their curators. There has, therefore, been a considerable effort to conserve these collections in institutions such as zoological and botanical gardens. These institutions have, with restrictions, permitted students access to the collections, but the value of such access depends on the holdings being readily known. Attempts to meet this need by publication of catalogues of specimens have been common, but have universally suffered from

* Author to whom correspondence should be addressed; present address: 14, Canterbury St., Dartmouth, Nova Scotia, B2Y 1S7.

† NRCC No. 33018

inaccuracy due to constant change in holdings. There have been many attempts to correct this deficiency by using mechanical search and retrieval systems. In theory, the ease of editing such data should make the systems up to date. In addition the availability of cheap personal computers and long distance telephone charges should make these databases available to a wide scientific audience.

One of the first attempts to implement this kind of database was that of Martin and his colleagues (Martin, 1964; Quadling and Martin, 1968; Simpson, et al., 1971). The database they assembled consisted of details of bacteria and some fungi held in about 60 collections of these organisms in Canada. They had the idea to organize the data in so-called fields i.e. blocks of specified numbers of bytes within a designated record assigned to each organism. Six fields were defined as follows: 1. Binomial name, authority, and its accession number in that collection. 2. the source of the organism, i.e. who deposited the organism in the collection. 3. where and from what the organism was originally isolated. 4. the methods of preservation in the collection. 5. 'special' features of the organism e.g. its pathogenicity. 6. references to the literature (if any). This system was used by bacteriologists in Canada and elsewhere but its main use was in compiling catalogues of organisms held in various collections and lists of names and addresses of such collections. As such it probably represented the first example of automated printing of microbiological data. Full details of search and editing programmes, files of synonyms of binomial names and other utilities have been published (Simpson et al., 1970). Its success also stimulated several industrial organizations holding collections of commercially valuable micro-organisms to automate retrieval of data about their collections in this manner.

It has collapsed because of lack of support, inaccessibility to working microbiologists, and the advent of more comprehensive databases. Two of these will be mentioned. The first operates, at least nominally, under the aegis of the American Type Culture Collection (ATCC) and is based on the work of Colwell and her colleagues (Rogosa et al., 1971, 1986; Philpot et al., 1982). This database was designed to be expandable i.e. as more information about particular species became available the field could be expanded without major modification in the search and retrieval programming. It was also designed to be much more comprehensive than the Martin database. This led to a system of converting information into a numerical six digit code, presumably convertible into machine readable code by scanning techniques. About 43 sections have been defined: 14 are broadly morphological and 12 strictly biochemical, though there is some duplication between sections. The first two sections of the code and section 16 contained information used in the Martin database but in more detail. For example the first section (Fig 1) contained the name and address of the person supplying the information, the genus, sub-genus and species epithet, a 'microbe code' of 12 two digit abbreviations for algae, fungi (& yeasts!), protozoa, bacteria (and rickettsia), mycoplasmas, virus (animal, insect!, bacterial & plant) and 'others'. In addition 20 questions (Fig 2) regarding strain characteristics are posed, the answers to many being apparent to a microbiologist from the preceding data. In addition to the 3 sections mentioned a further 40 sections had been designated by 1986, and the enterprise had become an attempt to codify the microbiological literature without a rational basis. Much of the data recorded was of limited longevity. For example 3 large sections (19, 35 and 40) are devoted to growth inhibition by some arbitrarily chosen antibiotics many of which have lost or will lose all clinical interest. Despite these and other criticisms the database is of great value and is available to the public through the BT Tymnet company (Customer support telephone number - 1-301-881-9020). Access to this source requires a $72 joining fee and thereafter user fees of about

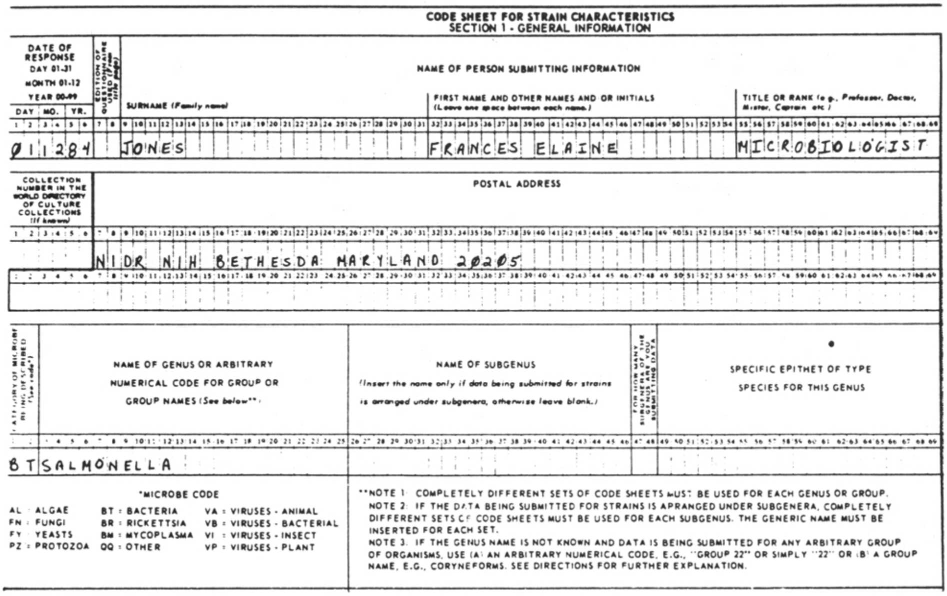

Fig 1 General information Coding Sheet for strain characteristics used in the Rogosa-Krichevsky-Coldwell system

$50 h⁻¹ are imposed. There are additional charges for downloading information e.g. by electronic mail and for storage to permit retrieved information to be accessible in subsequent sessions.

The second attempt to construct a microbiological database that will be discussed is that initiated by the biotechnology action programme of the Commission of the European Economic Community. The idea was to construct a network of databases each set of data representing a major culture collection in Europe (possibly restricted

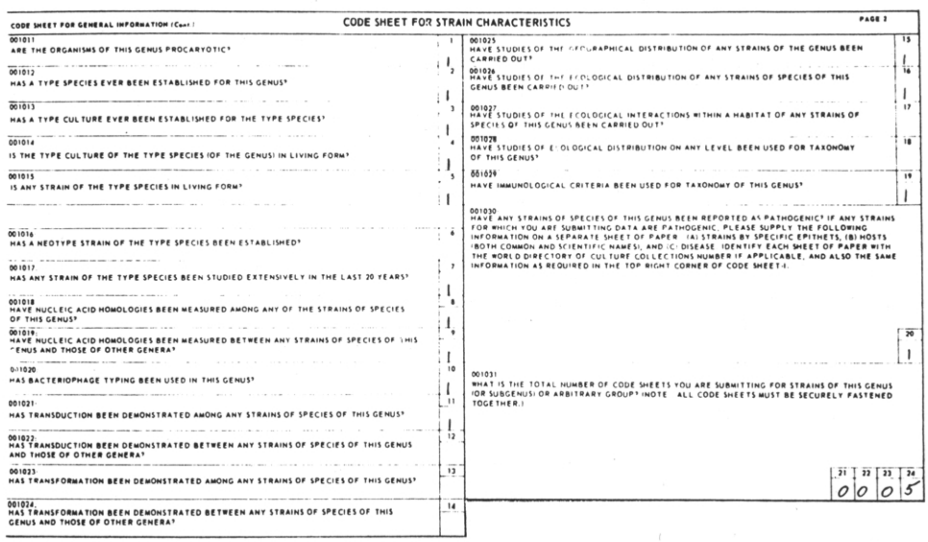

Fig 2 Detail of code sheet for strain characteristics (Rogosa et al., 1986)

to one node on the network for each country). The data in each set were to have the same format and were to be accessible by the common BASIS utility (Information Dimensions Inc., Dublin, Ohio, U.S.A.) . The database included large collections in the Netherlands, Belgium and the United Kingdom. The format for data entry has been published (Gams et al., 1988) and the system named Microbial Information Network Europe is planned to be on-line in 1992, but at present the International Mycological Institute's data and that of the National (U.K.) Yeast collection are available through the BT Tymnet company. The format is defined in a PASCAL sense of 3 record types within a main data record - each record then being subdivided into fields, 114 of which are described in the publication (Gams et al. 1988). These fields may or may not be duplicated in the record types. The fields are grouped into blocks as shown in Table I, presumably to aid in the entry of data. The data is also schismatic because 84 fields are not universally available on the network. Eight of these, grouped in a block called 'internal administration' contain such essential data as the accession number of the organism and the dates of acquisition and update. Obviously some of this information (though not all; Table I) has to be duplicated in the fields of the records that are available to the public. Seven fields dealing with references to metabolites, ecological, pathological, taxonomic (duplicated) genetic and patent literature will also be, surprisingly, unavailable to network users at the onset. Table I gives a cursory indication of the information stored in the various records and it will be seen that although it is less comprehensive than that envisaged by Colwell and her coworkers (Rogosa et al., 1986), it is, nevertheless, voluminous.

It seems likely that these two databases will provide microbiologists with very useful and important services in the future. One hopes that the relatively large funds needed for their maintenance and development will be available, but even if they are, these

Table I Data blocks used in the Microbial Information Network, Europe data bases

Block description	Number of Fields	Fields available on network	Notes
Administration	8	0	
Species name	13	7	Serotype & taxonomic data not on network
Strain administration	11	5	Accession number given; date of receipt not available
Authenticity of specimen	1	1	
History of isolate	14	7	Herbarium specimen numbers not on the network
Symbiosis, parasitism toxicology etc.	8	1	Only information on pathogenicity on network
Sexual state & behaviour	2	2	
Histology & cytology	10	0	
Genetics	12	2	Mutants & 'killer' yeasts information only available
Growth requirements	23	2	Only conditions for growth (incl. T factors) on solid media available on network
Enzymology, metabolites etc.	6	0	
Practical applications	6	3	Industrial applications included in 2 records - assay & patented organisms.

databases have several deficiencies from the point of view of the working scientist. Among these shortcomings are the impossibility of foreseeing the needs of future microbiologists, the inevitable expense of consulting the services, and perhaps most important, the difficulty of providing a mechanism of enabling microbiologists to interact and therefore improve the system. These problems are evident in the Chemical Abstracts database which contains many errors and omissions that are apparently immune from correction by working chemists.

About 12 years ago we installed a search and retrieval system that enabled scientists to locate fungal metabolic products reported in the literature (Brewer et al., 1978). The references were the product of more that 30 years reading by two working mycologists. The data were updated regularly, on the basis of our routine perusal of the literature and was supported by the National Research Council. The number of accessions of the system on a yearly basis were as follows: 1981, 18; 1982, 26; 1983, 31; 1984, 109; 1985, 52; 1986, 90 (the figure for 1986 is an estimate subtracting the increased usage after the culture collection database was added to the system, see below) and the average cost (i.e. discounting overheads such as storage charges) of a single search was about 90 cents. The use that was made of these data showed that it was possible to allow public access to a utility on a large main-frame computer at modest cost and without accidental or deliberate corruption of the system. It therefore seemed a worthwhile experiment to attempt to provide a similar system which would give access to information about live cultures held in Canadian culture collections. At the outset the data were restricted to fungi, because it became apparent to a number of mycologists attending a meeting of culture collection curators in London, Ontario in 1986, that a general agreement to establish a national culture collection database was likely to take a long time. Accordingly 10 curators agreed to make their data available and these data form the basis of the system described in this paper.

Methods

Computational equipment Most of the data was originally written using a Texas Instruments terminal equipped with dual casette drives. The data lines were corrected using the limited editing facility of this machine. Later an Apple Macintosh computer with its greater editing capability was used and the data were stored on 3.5 inch discs. All data were transmitted to the mainframe computer via a telephone line using the 'kermit' utility. Thus all crude data have been retained on one or other of these media. The database and its attendant programmes were originally implemented on a Control Data Cyber 170 machine with the NOS operating system. After about 4 years of use the system was transferred to a Digital VAX 8800 unit with VMS operating system. The VMS procedures giving access to the files are described in this paper; earlier NOS procedure files are available to interested readers on application.

Sources of Collection Data Eleven culture collections took part in this experiment. Ten of these were Canadian and of the other (International Mycological Institute, Kew, U.K.) only about one third of its holdings were incorporated into the database. In Table II the addresses of the collections, the name of the scientist donating the data, an acronym to distinguish the collection in accession numbers, the medium on which the data was submitted, and the numbers of cultures about which data were given, can be found. The data records for the collections at the University of Alberta and at Kew were more extensive than required for this database (see below). The donators of the data of these collections therefore submitted tapes containing the data we required, thus this

Table II Description of sources of culture collection information used in the 1991 database

Title of Collection	Postal Code	Curator or Donator	Acronym	Number of Cultures submitted	on database	Data received on
Natl. Research Council						
Halifax	B3H 3Z1	D. Brewer	HLX	1037	751	Index cards
Ottawa	K1A 0R6	R. Latta	NRC	478	478	Typed list
Saskatoon	*	R. Haskins	PRL	1266	1084	Index cards
Can. Forestry Service						
Edmonton	T6H 3S5	Y. Hiratsuki	NOF	652	652	Catalogue
Fredericton	E3B 5P7	L. Magasi	FSC	642	642	Catalogue
Forintek	K1G 3Z5	K. Seifert	FTK	2060	2245	Catalogue
Laval Univ.	G1K 7P4	R. Fortier	CRBF	380	380	Typed list
Labatt's Brewery Co.	N6A 4M3	I. Russell	LBT	354	353	Typed list
Agriculture Canada	K1A 0C6	S. Needham	DAOM	8329	7923	Notebooks, index cards
Univ. of Alberta	T6G 2E1	L. Sigler	UAMH	5041	5075	Magnetic tape
Commonwealth Mycological Institute (U.K)	TW9 3AF	D. Hawksworth	CMI	~8050	2624	Magnetic tape

*Now incorporated into the Agriculture Canada collection

editing of the information may well have introduced some of the errors we ultimately detected. All the information submitted was checked for mycological accuracy, for compatibility with published data on appropriate isolates and for internal discrepancies. In most cases these problems were resolved by correspondence with the curators and/or the mycologists who originally isolated the fungi. In certain cases accession numbers had to be changed. A list of these numbers, the new numbers assigned and the reasons for the changes are available upon enquiry by curators.

Data Format In selecting a format for the data to be recorded for each isolate, we were greatly influenced by the arrangement of the data records devised by Martin and his colleagues (Martin, 1964) because it appeared to be inexcusable to promote a format into which all previous work could not be included. We therefore adopted the concept of one record for each isolate, and subdivided this into 4 fields, 3 of which (the organism binomial name, its accession number and the substrate from which it was isolated) were identical to those used by Martin (1964). The fourth field contained information on the source, maintenance (i.e. culture and/or storage conditions) and the date of accession of the organism to the collection as in Martin's scheme. However, because the number of methods of storage of fungi are small and the sources of fungi, i.e. own isolation or a culture collection, relatively small, this information was stored numerically, and decoded by incorporating the translation in the search programme or in a suitably indexed file. Thus this fourth field was a multidigit number, the justification of which permitted retrieval of the information in a comprehensible way. For example, 20331984 indicated on translation that the culture was acquired in 1984 (4 right digits), was stored under oil and regularly subcultivated (33=100001 binary; each digit in the binary number corresponding to one method of maintenance) and that it was originally obtained from a culture collection in the file of sources whose index is 20 (the two left justified digits). As Colwell and her colleagues (Rogosa et al., 1986) have pointed out, this technique is to some extent open ended, since the number 20331984 can be extended (left justified) to the limit imposed by the computer's arithmetical processing capability. We have used this possibility by adding to the data information on the availability of the organism i.e. whether the collection charges a fee for supplying a culture or whether it is freely available, the pathogenicity of the organism, and a cross reference to the file of fungal metabolic products described in the introduction to this paper. This information could be added as two further left justified digits e.g. 1220331984, the most significant digit (1) indicating a reference on the 'mycotox' file and the next (2) indicating that the culture could be obtained for a fee.

Computer programme development

Assembly and Updating of Database Files The database was written using the beginners all-purpose symbolic instruction code (BASIC). The main reason for choosing this language was that each organism in a collection could be assigned one data statement on one line of text - each line being numbered. It was found that this format simplified checking for typing errors, spelling errors in the original documents, omission of data, computational errors and editing procedures e.g. replacement of common names with binomial names in the substrate field. The database was assembled in blocks of about 500 organisms and was written in two ways. The first was stored in binomial format and was indexed with respect to organism name and accession number. The second was written in readable text, each culture collection being a separate subfile. The format of both was the same and was as follows. Each record was limited to 105 characters (bytes) divided into the following fields: 0-13, accession number; 14-26, the numerical code described above; 27-57, the substrate description and 58-104 the binomial name.

The indexed file was used for public access i.e. for search and retrieval purposes and was recreated from time to time as changes in the collections occurred. These changes were made in the second file and could be cumulative i.e. changes could be made as received, and then at regular intervals a new indexed binomial file written for public use. Much of this editing was done on the Macintosh computer which was especially helpful for updating the literature cross references and compiling statistics of the database such as those reported in this paper.

Programmes to Search and Retrieve Information from the Database Information is retrieved from the database by a user employing a computer programme written in the common business-orientated language (COBOL). This programme was written in such a way that versions of it could be called by a user in any language that uses Greco-Roman script, with the restriction that input and output statements had to fit the defined picture formats. The source text of the programme was broken down into a number of files, one being the code that retrieved and formatted the information required. This information was obtained from three sources: 1, the main database; 2, a file of names and addresses of culture collections and 3, files containing on-line help (see below). In addition (in this particular implementation) there were two files that contained definitions of all programme prompts and messages (one in English and one in French). This programming architecture is shown in Fig 3. The 'logical name' facility of the VAX(Digital)/VMS operating system was used to build the executable programmes. The COBOL 'copy' statement shown in Fig 3 refers to the language file by the name

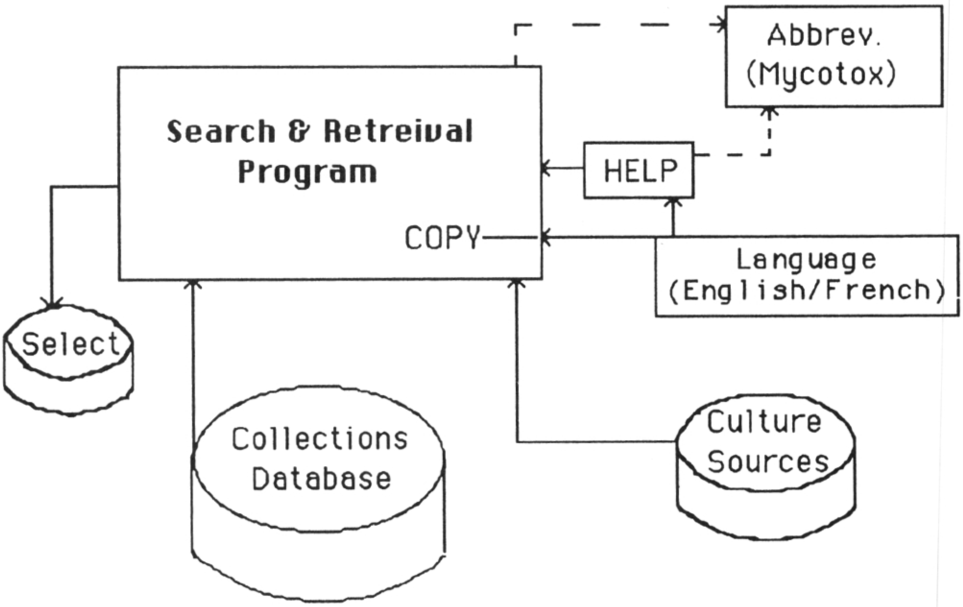

Fig 3 Files used by the programme for retrieval of information from the culture collection database. These files are called by the procedures 1, 2, & 3 in Fig 4.

'query–language'. Before passing the main source file through the COBOL compiler a VMS logical name is defined, thus allowing an English or French language file to be selected without changing the text of the main programme. The same technique is used to select a help file in the language of the user's choice. Hence there are separate executable query programs for each language of the system, each being assembled by defining the query language, compiling the COBOL main programme and linking the resulting object file to make the (English/French) executable program.

Two other features of the programming need to be mentioned. First it was desired to impose a limit on the output data because many cultures of the same species were maintained in some, or even most collections. Thus an enquiry for such a common species resulted in the output of several hundred lines of data. Output was controlled by writing this data to a scratch file before displaying anything to the user. The user was placed in control of how much of the data from the scratch file would be printed and/or displayed on the terminal screen. Thus the full data for one culture could be displayed or abbreviated details: binomial name, accession number, and the substrate on which it had been found for each of 10 organisms – could be printed. A subroutine was written to create a pause at this point giving the user the following choices - 1, a list of a further 10 organisms in the abbreviated format; 2, display of the full data for one organism by typing its accession number, and 3, to terminate the enquiry and/or relist the data. Help was offered to the user at this point by displaying the number of cultures remaining on the scratch file.

The second feature of the programming that requires comment relates to searching the database for organisms isolated from particular substrates. The term 'substrate' was used loosely and this part of the record contained a variety of information e.g. the binomial name of the plant on which the organism was found, the approximate geographical location of the isolation, characteristic enzymology etc. Code was therefore written to 'inspect' the substrate field for the occurrence of a string input by the user. Only when this string was found in the substrate field of the organism selected, was the data recorded on the scratch file.

Security The culture collection database and the list of references to mycotoxins are intended to be available to everybody. Though they contain no confidential information there remains the danger that the data and/or the computer system may be destroyed or corrupted deliberately or unintentionally by a user. It is therefore necessary to reduce such possibilities to a minimum.

Computer System Security In the VAX(Digital)/VMS system each valid username has a record in the system's 'authorization file' that contains e.g. the user's password, the number of central processor units that may be used etc. It is possible to invoke an unusual configuration where no password is required. This simplifies the login procedure for the user and discourages password exploration. The user's record in the authorization file also contains the specification of a command file (login.com) which is normally executed after the computer has checked the validity of the username and password but before access to the system is granted. In the present instance, this login.com file has been carefully written to try to ensure that a user cannot induce it to completion and thus gain access to all the computer facilities. The login.com file presents the user with a limited number of choices. One of the choices logs off the user, thus terminating the login.com procedure file (instead of gaining access to the computer system as a whole). The remaining choices in the login.com file give the user access to the culture collection database or the list of references to mycotoxins. An incorrect choice merely results in the menu of choices being repeated. This procedure is written as a loop so that the session is not terminated after a search but again presents the user with the menu of legal choices. In addition to the precautions taken in the login.com procedure, the username used for searching the database is flagged in the authorization file as "capture" which means that access to the system other than under the control of the login.com procedure is not allowed, thus protecting against errors in the login.com procedure. The procedures described are illustrated in the flow-chart of Fig 4.

Database Security The databases represent many hours of work in their assembly and the checks that have been made for accuracy. In addition, the curators who have supplied us with the data presumably own copyright to their segments of the data, which therefore have to be protected. We have sought to achieve this protection by the following methods. The datafiles are regularly copied to magnetic tapes, by ourselves and by the backup system of the computer. One tape is always kept outside the computer center. However, this does not exclude unauthorized or inadvertent changes that do not change the data format but which are cumulative and difficult to detect (gross changes would result in input/output errors during execution of updating programmes). To prevent user tampering with the files the username used to log onto the search programmes was given no library of files i.e. the databases and the programmes that search them were not located in the directory of that user number. Instead, permission was given for the user to read these, and only these, files that resided in the library (directory) of another user. This second user was conveniently ourselves, since this permitted us to update the files, improve the programming and exclude our use of the files from the outside use of the system. It also permitted two copies of the databases to be kept, one of which was not used by the public and hence comparison of the two files provided a further check on the integrity of the data.

Instructions for searching the databases, when on line are given in Appendix A.

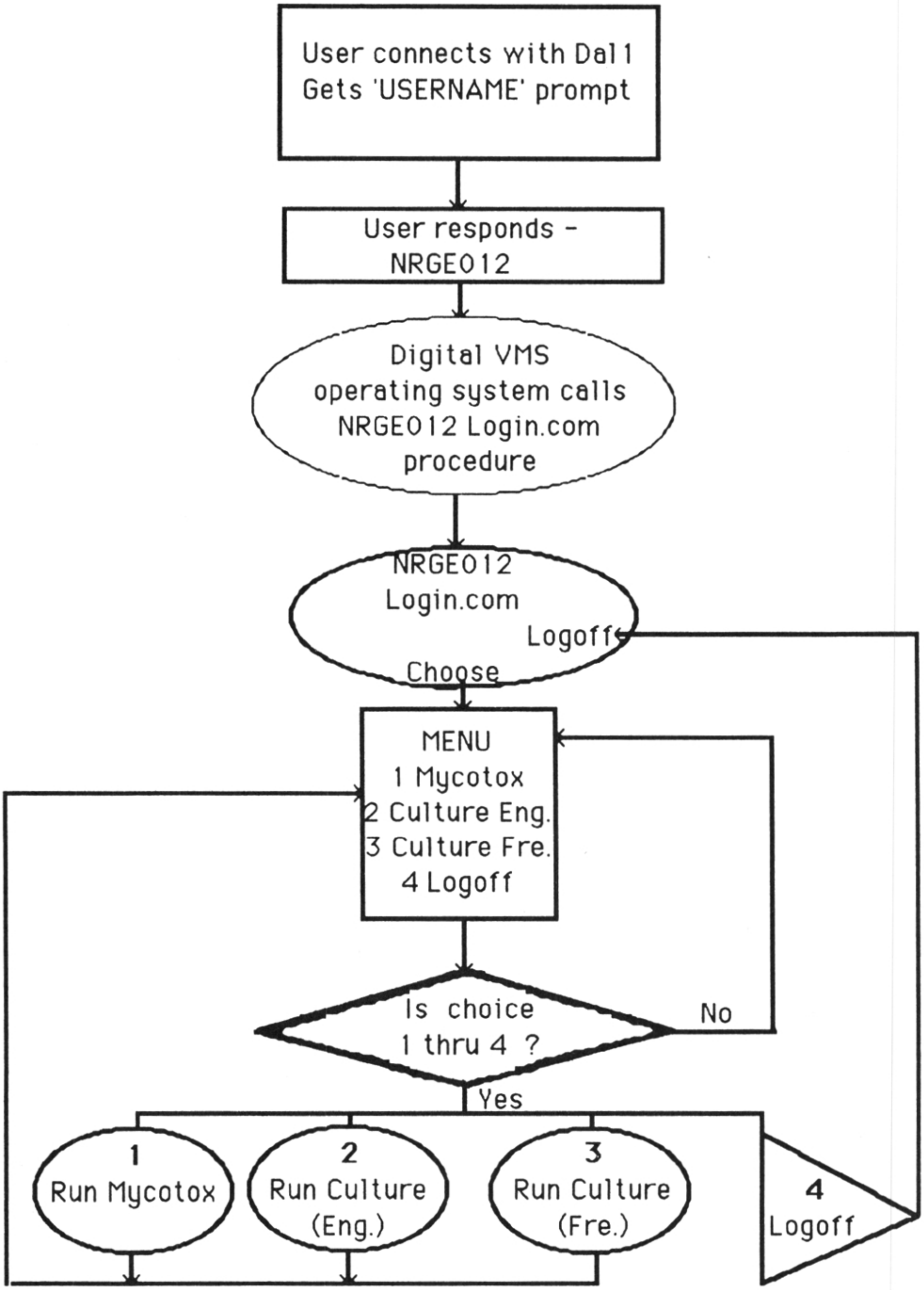

Fig 4 Execution flow-sheet of the Login.com file associated with the user number for public access of the database. Note that this Login.com file never proceeds to completion.

Results

Expenses incurred in setting up the databases

Literature reference file ('mycotox') This file and the programmes that searched it were set up in 1978 and the cost for doing this was $7426 broken down as follows: secretarial assistance $5567, computer processing $1667 and publication $192. The cost of writing the search and retrieval programmes was paid by Dalhousie University Computer Center because they were interested in the project from the point of view of open access to the computer services.

Culture collection database This file and the programmes that retrieved data from it were set up in 1986 at a total cost of $33598, broken down as follows: secretarial assistance $25662, contracts awarded to the University of Alberta ($500) to provide us with data and Dalhousie University ($2000) to write search & retrieval programmes, computer processing charges ($5111) and the expenses of demonstrations, correspondence etc.,$325. To this must be added part of the capital cost of the computer terminals used; this is deferred as they were also used in the maintenance of the databases.

Costs of maintenance of the databases

The disc storage charges for the literature reference file and its ancillary programmes in the period 1979-1985 inclusive were $505±40 y^{-1}. These storage charges increased after the culture collection database was added; the mean cost for the years 1987, 1988 and 1989 was $2505 y^{-1}. Another major cost of maintenance was that of updating the system and of correction of the data on the files. The system was updated when it was transferred to the VAX 8800 computer system and this was done by contract to Dalhousie University ($6000). The data on the literature reference file were corrected - usually by adding new references as they were published - on a weekly basis and the backup tapes rewritten monthly, the mean cost of these activities in the period 1979-1987 was about $90 y^{-1}. After the culture collection data had been added this expense increased to $2112 in 1988 and $1549 in 1989. A major item in this increased expenditure was the update of the cross references in the culture collection database to cover new work that had appeared in the preceding year.

Finally there is the capital costs of terminals $2900(=50%), tapes $75, floppy discs $150 etc. thus the total for setting up and maintaining the system was $65670.

Users and cost of usage of the databases

The number of users of the mycotoxin literature references file in the period 1981-1986 are given in the introduction of this paper. These numbers increased after the culture collection data was added as follows (year, number of accessions, total connect time in hours): 1986, 106, 5.63; 1987, 137, 9.699; 1988, 124, 5.111; 1989, ?, 6.677. The large number of connect hours in 1987 is probably due to curators of the collections learning how to use the system. The number of accessions to the system in 1989 is not known because the VAX accounting algorithm does not report this statistic. However, it is clear from the increases in central processing units used and connect time, that there was greater use of the system in 1989. The total cost of this activity was $1688 and on a yearly basis (rounded to the nearest dollar) 1981, $21; 1982, $30; 1983, $36; 1984, $120; 1985, $50; 1986, $295; 1987, $397; 1988, $232; and 1989, $507. Thus the total cost of the system over the period 1979-1989 was about $70,000. More significantly, perhaps, the running costs (allowing for amortization of capital costs and major system updates over 10 years) per annum are approximately $6000. None of these costs include the time and knowledge of two mycologists who were primarily responsible for setting up the system and maintaining it. The value of this work is difficult to compute because it was in part voluntary because of their scientific interests.

Statistics of the databases

At the end of 1989 there were 19585 records of fungal cultures from 10 Canadian collections and 2624 cultures from the collection of the Commonwealth Mycological Institute, on the database making 22209 altogether. There were 1482 genera of fungi represented in this data; the 30 most common, together with the numbers of isolates of each, are given in Table III. Eleven more genera were represented by more than 100 isolates; there were 53 genera with 50-99 isolates and 242 genera with 10-49 isolates. However the number of single species i.e. genera represented by only one culture was 525 or 35.4% of the total and 200 genera were represented by only 2 isolates. The maintenance of these 19585 organisms is summarized in Table IV. The oldest culture *Phellinus robinae*, was placed in the Agriculture Canada collection in 1912 and the numbers accumulated in the following decades were: 1912-1920, 32; 1930, 261; 1940, 555; 1950, 1428; 1960, 3320; 1970, 4882; 1980, 4492; and 1981-1986 inclusive 2945. The date of acquisition of the remaining 1670 cultures is not known.

Table III Number of isolates of the thirty most common genera on the database

Order	Genus	No. of isolates	Order	Genus	No. of isolates	Order	Genus	No. of isolates
1	*Trichoderma*	755	11	*Polyporus*	296	21	*Verticillium*	148
2	*Penicillium*	726	12	*Candida*	266	22	*Pleurotus*	144
3	*Fusarium*	638	13	*Ceratocystis*	257	23	*Acremonium*	144
4	*Aspergillus*	526	14	*Tyromyces*	245	24	*Paecilomyces*	141
5	*Saccharomyces*	475	15	*Phialophora*	244	25	*Coriolus*	139
6	*Chrysosporium*	455	16	*Arthroderma*	241	26	*Scytalidium*	138
7	*Poria*	389	17	*Suillus*	225	27	*Scopulariopsis*	127
8	*Trichophyton*	367	18	*Phoma*	198	28	*Malbranchia*	124
9	*Chaetomium*	354	19	*Fomes*	191	29	*Pythium*	123
10	*Peniophora*	348	20	*Phellinus*	184	30	*Alternaria*	123

Table IV Numbers of isolates in Canadian collections maintained by different methods and kept as herbarium specimens

Method of maintenance	Number of isolates
By regular subcultivation	8849
refrigeration at 4°	5733
lyophilization	9915
refrigeration at -196° (i.e. under liquid nitrogen)	479
under mineral oil	8497
Herbarium specimens	7339

Table V Canadian Provincial and Territorial distribution of isolates recorded in the database

Province/Territory	Number of Isolates
Alberta	2541
British Columbia	1148
Manitoba	243
New Brunswick	322
Newfoundland (including 5 from Labrador)	60
Nova Scotia	488
North West Territories (including 19 from Baffin Island)	50
Ontario	2817
Prince Edward Island	65
Quebec	1346
Saskatchewan	239
Yukon	18

Almost half (47.7%) of the cultures (9338) were isolated from Canadian sources, distributed among the Provinces and Territories as shown in Table V. The relatively large number from Alberta reflects the fact that two of the collections were from that Province, in one of which about 20% of its holdings were species of Arthroderma, Trichophyton and Chrysosporium (? human pathogens). Forty percent of the cultures (7874) in the database were isolated by the curators of the collections or by their colleagues, but 8553 (43.7%) were of unknown origin. A relatively small proportion (2442, 12.5%) were isolated from soil samples. Only 1243 (6.4%) were bought from the American Type Culture Collection (375), the International Mycological Institute (196) or the Centraalbureau voor Schimmelcultures (672).

About 20% of the isolates (4221) on the culture collection database were species that have been reported to produce toxic metabolites. The file of literature references contained 3382 entries at the end of 1989 in which 1155 species were described. It must be emphasized that the identity of a species in the culture collection database with one reported in the literature to produce toxic metabolites does not imply that the former produces the antibiotic(s) in culture.

Downloading and storage

The programme and procedure files, some 18 in all, used for setting up the system described in this paper have been copied to 3.5" (8.9 cm) magnetic discs in both the Macintosh Apple format and the International Business Machines (IBM) format i.e. they can be read on either computer with the aid of a rudimentary word processor. These discs have been deposited and catalogued in the library of the Nova Scotian Institute of Science. Copies of them may be bought from the Librarian of the Institute. The files can be loaded via a utility such as 'Kermit' onto any Digital VAX computer using the VMS operating system and having a COBOL compiler. Help will be required from the operators of such a computer system to set up the appropriate authorization code. For other main frame computers, a COBOL compiler for the source code will be required but given the files (copies of these are on the discs) called by the main programme there should be little difficulty in linking them and setting up an executable program.

The 'mycotox' literature reference file has also been deposited in the library of the Nova Scotian Institute of Science, both on 3.5" magnetic discs in either Macintosh or IBM format and also on 9-track magnetic tape. This file is available from the Librarian of the Institute for a fee that will be decided by the Council of the Institute when the expenses of doing this work are known.

The culture collection database file copied on 9-track magnetic tape, has also been deposited in the library of the Institute. It can be consulted freely by members of the Institute at the library provided that they give adequate notice to the Librarian. Copies of the data are only available after written permission has been obtained from the appropriate curators whose names and addresses are given in Table I, and signed copies of this permission are sent to the Librarian of the Institute. It is unnecessary to obtain permission from curators whose data is of no interest to the user, since any or all of the 11 culture collections data can be copied. As in the case of the 'mycotox' file it is the Institute's policy to apply a fee for these services which will be decided when their cost is fully known.

Discussion

Canadian microbiologists were pioneers in the use of computers to store and process information on bacteria, fungi, protozoa and other micro-organisms. In their publications (Martin, 1964; Simpson et al., 1971) they present strong arguments for the mechanical documentation of such information and these reasons remain valid (Russell & Sanderson, 1988). How is it, then, that this considerable body of work is presently unavailable to the public ? There are, of course, several reasons, some of which are worthy of comment. At the time (1964-1970) access to large bodies of data stored on mainframe computers could only be through the intermediary of computer scientists, who were uninterested in the microbiological use of the data, or by microbiologists prepared to learn computer programming. Inevitably the information became otiose and increasingly redundant. In addition, one of the main reasons for creating the database was to explore the feasibility and utility of such systems - their maintenance and longevity were secondary considerations. Thus when the interests of the scientists who created the database changed it was easy to terminate the support required for maintenance and development simply by default. Traditionally, storage of scientific knowledge has been achieved on a long-term basis by scientific societies in their published journals, which were then (and are) accumulated in libraries. These societies have been slow to react to the changes that have occurred in the last 10-15 years in the technology of information dissemination. This prudence was and is based on sound reasons; the tendency of purveyors of information media - the paper tape to compact disc progression - to change the format and hence the machines to read and write information, the high cost of computational machinery and so on. However, in recent years the cost of personal computers has declined to the point where many scientists can afford their own machines and the 3.5 inch magnetic disc medium, capable of recording about 10^6 alphanumeric characters is now available for less than 50¢.

Thus a scientific society has the option, when moving into the new technology, of presenting scientific data to the public in the form of magnetic discs and/or in the form of an online utility accessible on a main-frame computer. These considerations have led us to give a cost analysis for the implementation and maintenance of a small mycological database on line. It is clear from the analysis that the maintenance costs per year, of this database are not trivial, but the fact that they are about the same as one issue of these Proceedings makes their undertaking by a scientific society feasible. As in all other publication activity by scientific societies, the style or format of this new medium will have to be as concise as possible. Thus the format for the culture collection database was chosen not only to conform with that used by preceding workers (Martin, 1964; Simpson et al. 1971), but also, uniquely, to embrace our conception of the minimal essential information required. We see this choice as an advantage for other workers because they will be able to use these data as a basis for their own databases and they will have the minimum inconvenience in updating their segment of the whole.

The financial analysis also provides information on the merits of an online utility as compared to one based on magnetic discs. Thus when 2624 isolates from the collection at Kew (U.K.) were added to the database - an increase of about 12% - the cost of a search of the data (~90¢) remained the same but the storage cost increased by about \$500 y^{-1}. About four times this information could be stored on a 800 Kbyte disc for <50¢. Thus it is conceivable that the distribution of information on discs will compare favourably with on-line facilities. Details of a proposal to implement such a cooperative database are given in Appendix B.

If this happens the role of the scientific society will remain traditional. It will serve to collect, review and publish scientific information in an electronic format as well as by the familiar printed form. Its success or failure in this endeavor will depend, as always, on the support it receives from scientists. The statistics given in this paper show that there is a demand in Canada for mycological information in an electronic format, but the concept of cooperation for a national culture collection database is less secure.

References

Ainsworth, G.C., 1961. Storage and retrieval of biological information. *Nature, (Lond.)*, 191: 12-14.

Brewer, D., Taylor, A. and **Evans, A.**, 1978. A mechanical search and retrieval system for mycotoxins. *Proc. N.S. Inst. Sci.*, 28: 163-170.

Gams, W., Hennebert, G.L., Stalpers, J.A., Janssens, D., Schipper, M.A.A., Smith, J., Yarrow, D. and **Hawksworth, D.L.** 1988. Structuring strain data for storage and retrieval of information on fungi and yeasts in MINE, the microbial information network Europe. *J. gen. Microbiol.*, 134: 1667-1689.

Martin, S.M., 1964. Conservation of micro-organisms. *Ann. Rev. Microbiol.*, 18: 1-16.

Philpot, C.M., Krichevsky, M.I. and **Rogosa, M.**, 1982. Coding of phenotypic data descriptive of selected group of fungi for entry into computers. *Int. J. Syst. Bacteriol.*, 32: 175-190.

Quadling, C. and **Martin, S.M.** 1968. Organisation of information about micro-organisms. In Progress in Industrial Microbiology, 7: (Ed. D.J.D. Hockenhull). CRC Press, Cleveland, Ohio, pp. 125-148.

Rogosa, M., Krichevsky, M.I. and **Colwell, R.R.**, 1971. Method for coding data on microbial strains for computers. *Int. J. Syst. Bacteriol.*, 21: 1A-184A.

Rogosa, M., Krichevsky, M.I. and **Colwell, R.R.**, 1986. Coding microbiological data for computers. Springer-Verlag, New York.

Russell, I. and **Sanderson, K.E.**, 1988. Report of the task force on the status of culture collections in Canada. Ministry of State for Science and Technology, pp. 23-25.

Simpson, F.J., Okeren, I. and **Peardon, G.**, 1970. MICSYS, A system for preparing lists and catalogues of culture collections and selective retrieval of information employing the digital computer. National Research Council Canada, publication No. 11388, Ottawa, pp. 1-139.

Simpson, F.J., Okeren, I. and **Peardon, G.**, 1971. MICSYS, an information system using the digital computer for culture collections. *Can. J. Microbiol.*, 17: 403-414.

Weldon, J., Ferguson, J. and Schindler, D., 1986. Directory of Canadian culture collections; a report prepared for the national biotechnology advisory committee, Ministry of State for Science and Technology (Edited copies will be available from the Librarian).

Appendix A

Instructions for using the Databases

In the following instructions certain conventions are used. The sign ® is an instruction to press the 'enter' or 'return' key; CAPITAL letters are used to indicate the response of the computer and *italic* letters are used to indicate information typed by the user (regardless of the actual case input or output). Some of the commands can be abbreviated e.g. the list command can be either list or simply l. In the following discussion abbreviated commands are printed thus *l)ist*.

Use of the system is similar to that described in the earlier publication (Brewer et al., 1978). One logs onto the system either using a Datapac line (No. = 76100256) or by a telephone line (No. = 01-902-494-2500). When a connection is established - a buzz is heard and the user responds with ®. The following procedure is then followed:

WHAT SYSTEM? *dal1* ® ® (i.e. return or enter 2 or 3 times)
DALHOUSIE UNIVERSITY DECSERVER 200 TERMINAL SERVER V3.0(BL33)-LATV5.1
DS0006
LOCAL-010-SESSION1 TO DAL1 ESTABLISHED
USERNAME: *nrge012* ®
CHOISSISSEZ (PAR NUMERO) L'APPLICATION SVP

1. MYCOTOX - MYCOLOGICAL TOXINS
2. FUNGAL CULTURE COLLECTION DATABASE
3. BANQUE DE DONNES DES SOUCHES FONGIQUES
4. LOGOFF (Q ALSO LOGSOFF)

SELECT BY NUMBER OR TYPE 'Q' TO QUIT

The user now selects the utility they wish to use by typing *1* or *2* or *3*, or *4* and ®. By typing *1* access is obtained to the file of literature references to fungi that produce a toxic metabolite in culture. The organization of the file and the method of searching it is essentially the same as in 1978. The list of abbreviations that is part of the 'mycotox' utility (Brewer et al., 1978) has been expanded to include abbreviations used in the culture collection database. Thus to interpret abbreviations used in the latter file it is necessary to exit the culture collection search utility (see below) and enter the 'mycotox' utility through the menu selection process.

Functionally, choices 2 & 3 in the above menu differ only in the language used - according to the preference of the user (such language choice in the case of the 'mycotox' utility seems to be unnecessary since the names of the organisms, metabolites and journals are universally understood). For example when *3* is typed in response to the menu followed by ® the computer responds e.g.

BANQUE DE DONNES POUR LA COLLECTION CANADIENNE
DES SOUCHES FONGIQUES
(COMPILE PAR LE CONSEIL NATIONAL DE RECHERCHES)
POUR OBTENIR DE L'AIDE OU DES RENSEIGNEMENTS VEUILLEZ TAPEZ 'AIDE'
COMMANDE?

There are 3 ways to respond to the prompt COMMAND? when searching for information from the database:

1. If the accession number of an isolate is known (by for example from a list of isolates of a selected organism (see below)), typing (e.g.) a)ccess HLX 136 will generate on the terminal screen all the information on the file of the gliotoxin producing organism *Penicillium terlikowskii*, held in the collection at the National Research Council's laboratory in Halifax.

2. The second way to search the database is to use the command l)ist, thus *l)ist Penicillium terlikowskii* results in the output of the total number of cultures of this organism recorded on the database, all of the accession numbers of these organisms and finally, the substrates from which these organisms were isolated.

This information is printed on one line of <80 characters, hence details of many isolates can be revealed in a small amount of space.

As discussed above the information is output in blocks of 10 organisms when the programme pauses and asks the user:

X MORE RECORDS TO COME; TYPE 'END' TO STOP HERE, A BLANK LINE TO CONTINUE, OR AN ACCESSION NUMBER: (X being the remaining isolates waiting to be printed)

In this particular case the user merely types the accession number without a command word e.g *HLX 136*, if *end* is typed the user is returned to the menu of choices.

3. The third way of searching the database enables the enquiry to restrict a species or genus chosen to those found on, or restricted to, a particular substrate. For example if the substrate is recorded as 'roots of Abies rugulosa N.S.' the search can be restricted to those organisms isolated from 'root', or from 'roots of Abies', or from Nova Scotia, and so on. To do this the required substrate is entered *before* the name of the species sought as follows:

COMMAND? *s)ubstrate root* ®

COMMAND? *l)ist Penicillium terlikowskii* ®

and in this example only those isolates of *P. terlikowskii* found on the roots of plants will be printed. One reverts to the default mode, where in the searches the substrate is ignored by the following:

COMMAND? *s)ubstrate* ®

The information revealed about each organism is not restricted to its name, accession number and the substrate on which it was found. The following series of commands prints out all the information on the file for each organism:

COMMAND? *long* ®

COMMAND? *l)ist Penicillium terlikowskii* ®

This additional information is as follows:

1. Whether the organism is freely available from the collection where it resides, is available for a fee, or is available only after consultation with the curator.

2. Whether the organism is pathogenic to man.

3. From where the organism was obtained by the collection - where this is not known, UNKNOWN is printed. If the organism was obtained from another culture collection, the name and address of this collection is displayed.

4. Methods of maintenance of the organism - by one or more of the following methods: subcultivation, refrigeration at 4° or -196° (i.e. under liquid nitrogen), under paraffin oil and/or lyophilized.

5. Whether an herbarium specimen is kept.

6. The year the collection obtained the isolate.

7. Whether or not there is a literature reference to an isolate of this species producing a toxic metabolite.

Thus this degree of detail can be obtained either via an accession number or by a general (l)ist) enquiry for genus or species in the 'long' format. It is important to appreciate that internally the computer writes the full data to a scratch file. Thus segments or all of this scratch file can be printed in either or both formats by use of the 'relist','short' or 'long' (see below) commands; no further searches are performed - the choices are entirely in the hands of the user.

In the cases of organisms that are laboratory isolates, genetically selected and/or the products of molecular genetic manipulation, the substrate field is used for information on genetic characterization - usually the presence (+) or absence (-) of enzymic markers. These are given in accepted biochemical abbreviations using the 3-letter code for amino acids (e.g. ala=alanine) and sugars (e.g. Glc=glucose). Thus the database can be searched for organisms with or without such characteristics. The substrate field is also used to indicate (when known) the status of the organism as an ecto or endo symbiont.

When it is necessary to decipher one or more abbreviations, or to obtain literature references to metabolites of fungi found in the database the following sequence gains access to this information.

COMMAND? *end* ®
when the menu of choices is displayed by the computer
1 ®

*** NRC ATLANTIC REGIONAL LABORATORY ***
FUNGAL TOXIN REFERENCES

FOR HELP OR INFORMATION PLEASE TYPE "HELP"
COMMAND? *l)ist Penicillium terlikowskii* ®
PENICILLIUM TERLIKOWSKII NRRL A-537-A
GLIOTOXIN ACETYLGLIOTOXIN JACS 75 2110
PENICILLIUM TERLIKOWSKII HLX 136
GLIOTOXIN DEHYDROGLIOTOXIN J 1966 1799
PENICILLIUM TERLIKOWSKII O
BIS-DETHIODIMETHYLTHIOGLIOTOXIN CPB 34 340
PENICILLIUM TERLIKOWSKII O
GLIOTOXIN CPB 34 340
COMMAND? *a)bbrev NRRL* ®
NRRL
U.S. DEPARTMENT OF AGRICULTURE, NORTHERN RESEARCH SERVICE, 1815 NORTH UNIVERSITY STREET, PEORIA, ILLINOIS, 61604, U.S.A.
COMMAND? *a CPB* ®
CPB
CHEMICAL AND PHARMACEUTICAL BULLETIN
COMMAND? *end* ®
when the menu of choices is displayed by the computer.

It will be noticed that a few minor changes have been implemented in the 'mycotox' file since the preceding publication (Brewer et al. 1978). These changes result from rewriting the utility in COBOL to improve the response time and to use the powerful editing and search facilities in the Digital VMS software.

Users of the system can obtain help for their searches of the culture collection and the mycotox files on line. We have tried to make this help self explanatory and to use the same commands to give the same results on each file. The following illustrates the use of this on line help:

COMMAND? *help* ®
FOR HELP ABOUT A PARTICULAR TOPIC TYPE 'HELP XXX' WHERE 'XXX' IS THE NAME OF THE TOPIC
FOR A LIST OF TOPICS FOR WHICH HELP IS AVAILABLE TYPE 'HELP INDEX'
FOR A SUMMARY OF COMMANDS TYPE 'HELP COMMANDS'
TYPE "END" IF THIS IS ENOUGH HELP? *help commands* ®
COMMANDS
COMMANDS ARE ENTERED IN RESPONSE TO THE COMMAND? PROMPT.
COMMAND ABBREVIATIONS ARE SHOWN IN PARENTHESIS.

ACCESSION	(A)	- LIST THE WHOLE RECORD FOR A PARTICULAR ACCESSION NUMBER.
END		- LOGOFF THE SYSTEM.
HELP	(?)	- REQUEST HELP.
LIMIT		- SET THE NUMBER OF RECORDS PRINTED BETWEEN PAUSES.
LIST	(L)	- LIST RECORDS FOR A SINGLE GENUS OR SPECIES.
LONG		- LIST RECORDS IN THE "LONG" FORMAT.
RELIST	(R)	- RELIST THE OUTPUT IN LONG OR SHORT FORMAT.
SHORT		- LIST ONLY NAME, ACCESSION NUMBER AND SUBSTRATE.
SUBSTRATE	(S)	- SELECT ONLY RECORDS WITH THE DESIGNATED SUBSTRATE.

TYPE "END" IF THIS IS ENOUGH HELP? *end* ®
COMMAND? *end* ®
when the menu of choices is displayed by the computer.

To logoff from the system completely the user responds to the menu by typing *4* ® when the computer responds:

ACCOUNTING INFORMATION

BUFFERED I/O COUNT	XXXXX	PEAK WORKING SET SIZE	XXXX
DIRECT I/O COUNT	XXXXX	PEAK PAGE FILE SIZE	XXXX
PAGE FAULTS	XXXXX	MOUNTED VOLUMES	XXXX
CHARGE CPU TIME	XX secs	ELAPSED TIME	h. m. s

LOCAL-011-SESSION1 DISCONNECTED FROM DAL1

This information is worth noting so that telephone or Datapac charges can be checked.

Appendix B

A Proposal to Establish a Cooperative Canadian Culture Collection Database

The format used to describe a culture in a collection in the database reported in this paper requires a record of length 105 bytes. In the Directory of Canadian culture collections (Weldon et al., 1986) the curators of 140 collections reported their holdings. A summation of these holdings indicates that about 120,000 organisms are maintained in the collections. Thus this information could be recorded in 1.26×10^7 bytes. In practice the recording medium required is less than this because 47 bytes are allocated for the binomial name of the organism when usually less than 30 are used. For example the data describing the 7923 cultures in the DAOM collection occupy 6.06×10^5 bytes, or about 77 bytes each. Thus all the cultures in the directory (Weldon et al., 1986) could be recorded on fifteen, 800 Kbytes, 3.5" magnetic discs, or 3 of the newer 4 Mbytes discs. The cost of such media is about $4.00 and since it can be quickly and easily reproduced in a number of disc formats, this technology offers an attractive way of assembling a national culture collection database as an alternative to the conventional on-line facility described in Appendix A.

The following is an outline of how such a system might operate.

Input of information into a national culture collection database

It is now possible to buy a reconditioned computer for less than $300 so we assume that in the future all scientists will have one. This being the case we assume that the scientist will keep records of his/her cultures on this sort of equipment and because of the biological characteristics of *Homo sapiens,* that the data and the data format used by each scientist will be different. Our hypothetical microbiologist will, from time to time update their data and we propose that when this is done the machine automatically write 2 files of the data. The first will be in the format used by the scientist and the second will involve editing the information to format it as described in this paper. This second file can then be copied to a 3.5" disc. The programming for this operation is easy and to some extent can be standardized. The disc can then be mailed to a central repository. Discussion of the latter is deferred for the moment.

Output of information from a national database

We envisage three kinds of output normally, but not restricted to, the central repository.

1. *News of and about the information available* This will be provided by publication in the Proceedings of the Nova Scotian Institute of Science, and updates will also be published in this journal. Information will gradually accumulate and be revised from time to time and all of it will be available from the Librarian of the Institute in the 3.5" disc format at a modest fee.
2. *Specific queries* By specific queries we mean straightforward questions designed to elicit information from the database e.g. Who has a culture of *Fusarium graminearum* that produces deoxynivalenol ? Such queries can be made to the Librarian of the Institute by mail using a 3.5" disc and the reply sent, again at reasonable cost on the same disc. In this way scientists will be able to build their own culture sources file in whatever format they choose. Requests for statistics can be treated in the same way. We have some evidence that users of the system in the last four years have done this sort of work.

3. *Downloading* Here we are talking about fairly large blocks of data desired for use by e.g. curators who by misadventure have lost their files, or by biologists interested in taxonomic studies of a group of organisms. Such copies of blocks of data may infringe the intellectual property rights of the owners of the data. In these cases the procedure described in the paper is proposed. The minimum nature of the information on each organism will ease its elaboration by the recipient scientist in any desired format.

Central repository

It is proposed that the repository and clearing house for the national culture collection database be a scientific society such as the Nova Scotian Institute of Science. We believe that a dynamic national catalogue of culture collections held in Canada can be developed on these lines; its future will depend, as always, on the cooperation of the scientists who will use it in the course of their studies. They normally document their results by publication in a suitable journal, hence the proposal can be regarded as an extension of this familiar process. The discs they deposit with the Institute of Science will not therefore, differ materially from any other scientific publication.