

Computerization of strain data in the Microbial Information Network Europe (MINE)*

W. GAMS¹, J.A. STALPERS¹, G.J. STEGEHUIS¹ & J. SMITH²

¹ Centraalbureau voor Schimmelcultures, P.O. Box 273, 3740 AG Baarn, The Netherlands

² University of Amsterdam, Department for Medical Informatics, Academic Medical Centre, Meibergdreef 15, 1105 AZ Amsterdam

GAMS, W., J.A. STALPERS, G.J. STEGEHUIS & J. SMITH (1990). Computerization of strain data in the Microbial Information Network Europe (MINE). – *Sydowia* 42: 218–230.

The Microbial Information Network Europe (MINE), sponsored by the CEC, has been established to coordinate, harmonize and integrate data on microbial cultures held in collections of so far nine EC countries. The aims of these databases differ from other systems designed for either identification or bibliographic documentation. The software-hardware configurations have been critically selected. Uniform data formats have been agreed for the major groups of micro-organisms to assure compatibility between all systems used. Data structures have been implemented according to the standardized formats, and the available information (full data set) for all strains is being computerized. The databases of some of the participating collections are now accessible on-line. User-friendly menu screens have been designed that do not require a detailed knowledge of the data structure from the user. To allow a simultaneous search of organisms that answer certain requirements, an integration of a selection of data (minimum data set) has been executed for the fungi and yeasts and for bacteria, and the combined data will be made accessible on-line on a Central Database Node. Data Integrating Nodes take care of the integration, and Responsible Committees have been organized for the various groups of organisms to supervise harmonization of the data. The amount of data to be computerized and integrated, the number of participating collections, and the types of organisms represented in the system will be expanded in the near future.

In the EC countries numerous microbial culture collections have developed independently for decades. They are serving research, education and industry, and particularly biotechnology is strongly interested in these resources. Altogether, these collections are estimated to store a total of about 150,000 strains. In recent years most collections felt a great need to computerize their holdings. The awareness of the need for a qualitatively better information service is growing both internally from the management of the collections and externally from the users who wish to

* Paper based on a talk given at the Fourth International Mycological Congress, Symposium G-2, Computers and Information Systems, held in Regensburg, FRG, 28th August – 3rd September 1990.

- find organisms that respond to a particular set of conditions
- obtain all relevant information on these cultures.

Many European culture collections are collaborating in the European Culture Collection Organization (ECCO). In view of the importance of microbial cultures in biotechnology, the Commission of the European Community (CEC) in its Biotechnology Action Programme (BAP) has also decided to promote the computerization of collection data in a centralized way. The network was named **Microbial Information Network Europe (MINE)**.

The harmonization of information on organisms held in European culture collections will make them more readily accessible to the users. European collaboration in microbiological research is stimulated in a much broader frame by the CEC in order to achieve "European Laboratories Without Walls" (ELWW, see CEC, 1989). The present project can profit from the experience of some predecessors such as the government-supported MiCIS (Microbial Culture Information Service) in the UK and the On-line Services system of NCYC (National Collection of Yeast Cultures, Norwich, UK). The MSDN (Microbial Strain Data Network) is a similar approach, but designed to act as a central directory of depositories for microbial strains and cell lines, and to disclose information on the sites where particular types of data are available. It uses the RKC Code (ROGOSA & al., 1986). This system supplements MINE and does not duplicate the specific strain databases.

The principal aim of MINE is to disclose strain data on taxonomy (nomenclature), origin, physiological properties, industrial applications, etc., but not to serve directly for identification. This entails a major difference with other approaches of computerization, where a flat table structure of large numbers of codes (e.g. the RKC code) is used to disclose a vast amount of data which mainly serve for identification, ecological analyses, etc. It is also not intended to compete with existing bibliographic database systems. Bibliographic data are needed in a culture collection database, as far as particular strains are concerned and for references to the most important publications on each species. In addition, the origin of all data on strains must be carefully documented.

In 1985 collections of the F.R.G., the Netherlands, the U.K., Belgium and Portugal, started the project. They were soon joined by those of four other EC countries.

Nodes participating in MINE

Belgium: Belgian Coordinated Collections of Micro-organisms (BCCM), Science Policy Office of

- Belgium, Rue de la Science 8, B-1040 Brussels (3 collections).
- F.R.G.: Deutsche Sammlung von Mikroorganismen und Zellkulturen (DSM), Mascheroder Weg 1B, D-3300 Braunschweig.
- France: Bureau de Ressources Génétiques (BRG), 3 Avenue Victoria, F-75100 Paris (4 collections).
- Greece: Dairy Laboratory, Agricultural University of Athens (ATHUM), Botanikos, GR-11855 Athens (4 collections).
- Italy: Collezione dei Lieviti Industriali, Dipartimento di Biologia vegetale, (DBVPG), Univ. di Perugia, 74, Borgo 20 Giugno, I-06100 Perugia.
- The Netherlands: Centraalbureau voor Schimmelcultures (CBS), P.O. Box 273, 3740 AG Baarn (8 collections).
- Portugal: Portuguese Yeast Culture Collection (PYCC), Gulbenkian Institute of Science, Apartado 14, P-1093 Oeiras Codex.
- Spain: Colección Española de Cultivos Tipo (CECT), Department of Microbiology, Univ. of Valencia, 50 Doctor Moliner, E-46100 Burjasot.
- United Kingdom: CAB International Mycological Institute (CMI), Ferry Lane, Kew, Surrey TW9 3AF (9 collections).

Long discussions have taken place about how to realize the goal of a uniform access to the data of many collections. The choice of a structure for co-ordination and combination is determined by technical and political considerations. Each participating collection is interested in safeguarding its rights of intellectual property and responsibility for the data. Moreover, updating of combined data sets is quite problematic.

Therefore it was decided that a **network of distributed databases** should be established. In this system each participating collection can retain its full set of data and update this regularly. Databases of the major collections of each country will be connected in such a way as to allow them to retain the property of and the responsibility for the data in their keeping, while at the same time allowing optimal access to the data of all participants (GAMS & al., 1988).

In MINE computers situated at the **national nodes** for each participating country are connected by means of DECNET software (Digital Equipment Corp.), which allows the on-line connection of an external user to the major nodes and through-connection from that to the others in the system. It was envisaged that, with the further development of software, it would become possible to simultaneously access different databases not only on one but on several computers.

During the first phase of the project, it was, however, recognized that it was easier to integrate data physically into one database than to realize the simultaneous access to several computers, especially as the number of different computer and database systems grew.

At the moment a so-called **minimum data set** (generally the same amount of data as that reproduced in printed catalogues, the contents of 33 fields) is contributed by each collection for a **physical integration** into one database that is made accessible on-line. These combined data will be located at a Central Database Node (CDN) but may also be loaded on other national computers as far as required. The combined minimum data sets should later be extended to cover as many data as possible for each strain, thus approaching the **full data set** (which comprises 115 fields for fungi and yeasts).

The **practical work** comprised the following activities:

- select software-hardware configurations
- define general and standardized formats for the data (independent of the software)
- implement the format in the software chosen and entering of data
- provide user-friendly on-line access
- integrate strain data

In the following text these phases will be illustrated largely with an outline of the development at the Dutch MINE node, the Centraalbureau voor Schimmelcultures, Baarn.

Software – hardware configurations

It is desirable to limit the number of software systems in view of the compatibility within MINE. After defining the requirements of this project, the collections from the F.R.G., the Netherlands and the U.K. chose for the combination BASIS – VAX. The package BASIS from Battelle (Columbus, Ohio, USA) is a very flexible, non-relational system, which combines the advantages of free field length, the capacity to cope with very large amounts of data, the possibility of various kinds of indexing single words, subfields, even of several different ranks, rapid search, also in combination with various logical operators, good security provisions preventing access to certain non-public data, establishing thesaurus files, etc. For the smaller collections, the combination ORACLE – IBM/AT (with 1.5 MB extended memory) was chosen and conversion programs were written to make the two systems compatible. The BASIS system will be converted to the relational BASIS-plus system in 1990, without losing the above mentioned advantages.

At a later stage, the Coordinated Belgian collections chose the relational database system Rdb (DEC), which also runs on Micro-VAX computers. They link this system to MacIntosh computers and 4th Dimension software.

Definition of general and standardized formats for the data

Adherence to standard formats is a prerequisite for harmonization and compatibility within the European network. Decisions on the format are largely independent from the software and must precede the implementation.

Most of the data are textual, and software coping with indexed text has been chosen as a suitable way to store and retrieve strain data, rather than codified tables as in the RKC procedure (ROGOSA & al., 1986). The available advanced software packages are highly flexible in using long fields with differentiated subfield structures. After careful planning, agreements on convenient numbers and definitions of fields were reached, the contents of which are sufficiently independent so that distinct search arguments are meaningful. For some fields rapid search facilities are provided, for others there is no such need.

Detailed accounts of the chosen formats for fungi and yeasts (GAMS & al., 1988) and bacteria (STALPERS & al., 1990) have been published. The two formats differ in some fields from each other, but similar fields have in principle the same contents, so that the databases can easily be combined.

It is inherent to culture collections that a number of strains belong to one species. Besides the name, some data relate to the species in general and need not be repeated with each strain. This situation led to the definition of two record types, **SP**(ecies) and **STR**(ain). The STR records can be linked to SP by including the record number of the respective SP record in a special field, APOS (alternate posting).

Frequently several synonymous names are available for one organism, amongst which only one can be correct according to the rules of nomenclature, but others are still in wide use and the computer must allow to retrieve the appropriate name and strains from any applicable name. This problem is best solved by establishing a third record type **SYN**(onym), which also is subordinated to SP. For fungi, a fourth record type is used to cope with the frequent case that teleomorphs and anamorphs (perfect and imperfect states) of one fungus have separate names, both of which are valid. The record type **ALT**(ernative morphonym) links one of them to the preferential alternative morphonym (GAMS & al., 1988). The teleomorph name has

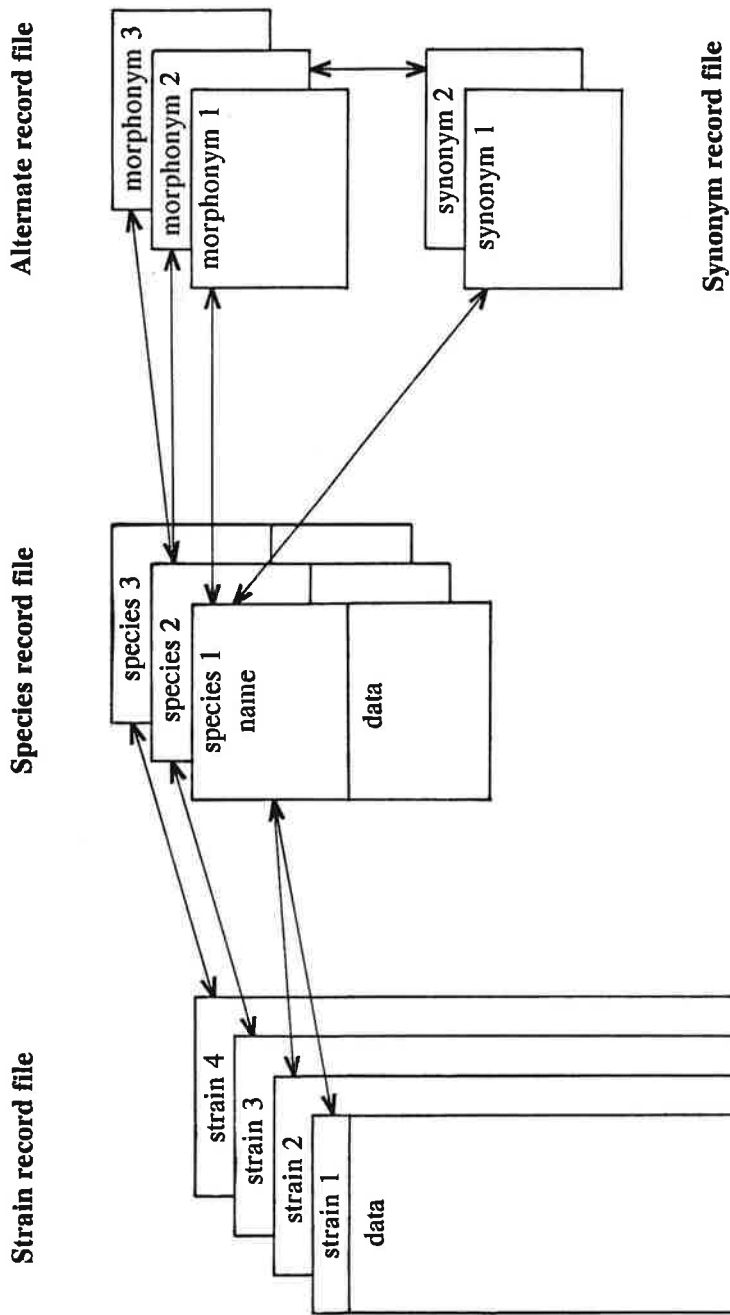


Fig. 1. - Connection between species and strain records, and between strain, synonym and alternative morphonym records. (drawn by G. L. HENNEBERT, from GAMS & al., 1988).

greater importance in mycology than the anamorph name, but in practice the anamorph is the form usually encountered *in vitro*, and sometimes the correlation between anamorph and teleomorph is equivocal. For these reasons, the preferential use of anamorph names is justified in some cases. The interlinkage of the four record types is illustrated in Fig. 1.

The data are arranged in 12 blocks of fields: internal administration – name – strain administration – status – environment and history – biological interactions – sexuality – properties (cytology, biomolecular data) – genotype and genetics – growth conditions – chemistry and enzymes – practical applications.

The origin of information (from the depositor, the institute of the culture collection, or literature) can principally be indicated in each field after a delimiter (“<”), which excludes the subsequent part from indexing. Entering of literature references is possible in general fields (LIT) and fields pertaining to the major blocks (ECOLIT, PATHLIT, PROPLIT, GENLIT, CHLIT, APPLIT).

Implementation of the format in the software chosen and entering of data

After definition of the formats, databases had to be created that are able to store the data accordingly. With BASIS, this is accomplished by compiling definitions in a “Data Definition Language” (DDL), and by creating the database files using the compiled definitions. The DDL can be used to define:

- field names of all fields in the database;
- field labels used during display of the data;
- fields which are indexed, and how they are indexed;
- fields for which thesaurus files are used;
- validations to be performed during data entry;
- preformatted screens to be used during data entry;
- levels of security for fields and records in the database;
- names of users and their privileges.

Thesaurus files in BASIS aid in maintaining consistency during data entry. Only terms that are accepted by the thesaurus can be entered in the appropriate fields. Incorrect terms can be refused, or if recognized, they are automatically changed to the preferred terms. This is both helpful during data entry and search; if a user asks for all strains from ‘Ceylon’, this search argument will automatically be replaced by ‘Sri Lanka’, the preferred term in the thesaurus. Thesaurus files for names of countries, culture media, collection acronyms, and enzymes with their abbreviations have been completed, others will be added soon.

To provide information on the classification of genera, it was found preferable to build a separate **genus database** rather than a thesaurus file in the main database. Data have been extracted for this purpose from the database file of AINSWORTH & BISBY's Dictionary of the Fungi (HAWKSWORTH & al., 1983), ERIKSSON & HAWKSWORTH's (1988) text file of Ascomycete genera, text from CARMICHAEL & al. (1980) on Hyphomycete genera together with supplementary data accumulated at CBS. This database will soon be available.

After the definitions of the databases had been created, data entry could start. At the CBS, the text of the List of Cultures was available in an ASCII file, and a program was written to convert this file so that it could be used as input for BASIS. Because it was difficult for the conversion program to recognize the fields into which the text had to be distributed, many corrections had to be carried out individually. After this, the process of adding data from the central card system started. On photocopies of the cards a mycologist indicated the fields to which the information pertains and a data typist entered the data. Reports were printed to enable the specialists to check the input. With increasing experience, less pre-processing of the cards was necessary. Adding all the data from the card system took one data typist about three years (half-time); the responsible mycologists then thoroughly checked the quality of the data. This work was ended early in 1990. Data on new strains are entered in the database immediately on accession. A card printer allows printing of cards from the database.

In a next phase, additional information is gathered from other sources. At the CBS, some specialists have personal card systems and also the chemical department has much valuable information that is still to be entered.

At present the databases contain already much more information than what can be published in catalogues. Moreover, the information that is accessible on-line reflects the current situation and is up-to-date, whilst a printed catalogue always lags behind the actual state.

In the Netherlands, the Netherlands Culture Collection of Microorganisms (NCC) is situated in the Centraalbureau voor Schimmelcultures (CBS) at Baarn. The CBS collection comprises about 32,000 fungal cultures. Other collections of yeasts and bacteria are computerized at other places in the Dutch network. The CBS Baarn directly updates its data in the BASIS system. At the other institutes, strain data are maintained on PC/AT computers with ORACLE software. As often as desirable, records that have been added, updated or deleted can be extracted from these databases and sent to the node at Baarn, where the central database is updated to

reflect these changes. The Dutch national node thus contains data on the strains of all participating Dutch culture collections.

Providing user-friendly on-line access

Full profit of the MINE system can be made by customers who use the database on-line. Access can be gained by using packet switching systems (PSS). Access to the PSS is possible either through a direct line or by a modem connected to a PAD (packet assembler-disassembler). Users can also use a modem and a telephone line to connect to a MINE node directly, but this works only over short distances. On-line access is at the moment possible to the databases of NCYC, CBS, DSM, and CMI. The on-line connection is also used for electronic mail between the members. The Mail Utility also enables users to order cultures, or to leave messages or questions for the microbiologists.

The databases of CBS, NCYC and of the Information Centre on European Culture Collections at Braunschweig (IC ECC, where the MiCIS database resides now after its conversion to the MINE format) can now be reached through a gateway from Telecom Gold in the UK and from Dialcom in the USA. Through this gateway, researchers who have a mailbox in the MSDN network can access the databases after logging in in their mailbox. Especially for users in the USA it is difficult otherwise to get connected to European national packet switching systems.

Because users cannot be expected to know all the field tags defined in the format, a **user-friendly menu system** has been developed at CBS that allows retrieval of information from the database, without much knowledge about the MINE format. This menu system is not only used at CBS, but also at the ICECC.

A first version had been written with the BASIS Menu Language, a kind of programming language which is part of BASIS; as this turned out to be too slow, the programs were rewritten and considerably extended in FORTRAN programs that call BASIS sub-routines.

Users can retrieve strains in many ways: not only collection numbers and species names can be used as search arguments, but also substratum, origin, depositor, substances produced or decomposed by the strain, names of organisms affected by the strain, symptoms caused, and industrial applications. Logical 'AND' and 'OR' combinations can be used. A special option allows to survey parts of the index files to see the available searchable terms. When a number of strains has been retrieved, their strain number and species name are displayed, and a user can ask for limited or full information on the strain (Minimum or Full Dataset). Users seem to be satisfied

with the menu system, although extensions and improvements are still necessary and will be implemented in the near future.

Integration of strain data

After much discussion it was decided to integrate data at strain and not at species level. This choice is laborious but is the most efficient in order to match all bits of information (naming, origin, properties . . .) for each strain held in several collections and supposed to be identical.

Data from all participating collections are loaded as sequential ASCII files in one database, and strains that have at least one collection number in common are assumed to originate from the same strain. This does not necessarily imply that the combined strains have the same properties; but a strain that has been kept in one collection for many years will probably also not have exactly the same properties as when it was accessioned. Data on the 'same' strain held in different collections can only be integrated, if the source of each piece of information concerning the strain is carefully recorded, and, moreover, not all fields can be meaningfully integrated.

In view of integration, fields can be divided in some groups:

- A. Fields concerning the history of the strain, like original substratum, location where it was found, collector, isolator and, partly, identifier. These fields can be integrated: at the end the database should contain only one value, which is accepted by all collections.
- B. Fields like method of preservation, date of accession, depositor, etc. The data of these fields depend entirely on the individual collection, and integration is meaningless.
- C. Fields in which properties of the strain are recorded, like production of enzymes, cell wall constituents, co-enzyme Q system, etc.. For a user it will be most valuable if for these fields the 'sum' of information of all collections can be shown, but for each value the source of the information should be indicated. It should not be suggested that all properties mentioned apply to each of the substrains in different collections.
- D. The name of a strain and all its derivatives should, ideally, be identical, but in practice this will often not be the case.

Mycologists may disagree about species concepts, nomenclature and preference of teleomorph or anamorph names. Separate identifications may have yielded different results.

It is still being discussed, how far the physical integration of individual databases should proceed. It has been shown that integration is feasible in both relational and non-relational database systems and, outside a database, using FORTRAN programs with sequential text files. At the moment, minimum data sets have been integrated at **Data Integrating Nodes (DIN)** for fungi and yeasts of ten collections at Baarn, and bacteria at Ghent. This first exercise of integration has shown that there are numerous inconsistencies in the contributing databases, which come to light and have to be straightened out by **Responsible committees** established for each group of micro-organisms. Besides a lot of editorial details, the required effort will help to provide the most complete data on strains in each collection and also serves to discover some errors in cross references. These advantages must be related to the amount of extra work involved in straightening out the data.

Further developments

The BAP programme of the CEC has ended in 1989. MINE has reached its goals defined until 1989, but it is not yet completed. Most of the participating culture collections have produced printed catalogues in a comparable lay-out and some have implemented their full data sets according to the standardized formats. The further integration and harmonization of the data must still take place and the viability of the system must be assured for the future.

The CEC is now starting a new programme, BRIDGE, in which continued support to MINE can be expected. The planned work will include a further integration of more comprehensive data with subsequent feed-back of the harmonized data to the individual collections; the collection of additional data on the strains from regularly published and 'grey' literature and new research, and addition of this information to the database; the expansion of the system by soliciting the participation of additional collections of EC- and non-EC countries and extension with databases of other groups of organisms. The menu structure for access will further be improved and software will be developed for simultaneous access of different databases.

In the first phase the national nodes were the primary structure within the network. In the second phase the Data Integrating Nodes (DIN) and Responsible Committees (RC) will have the most important function, whilst project leaders designated at national level will steer the development in a Board of MINE (Fig. 2).

CENTRAL EUROPEAN DATABASE SYSTEM

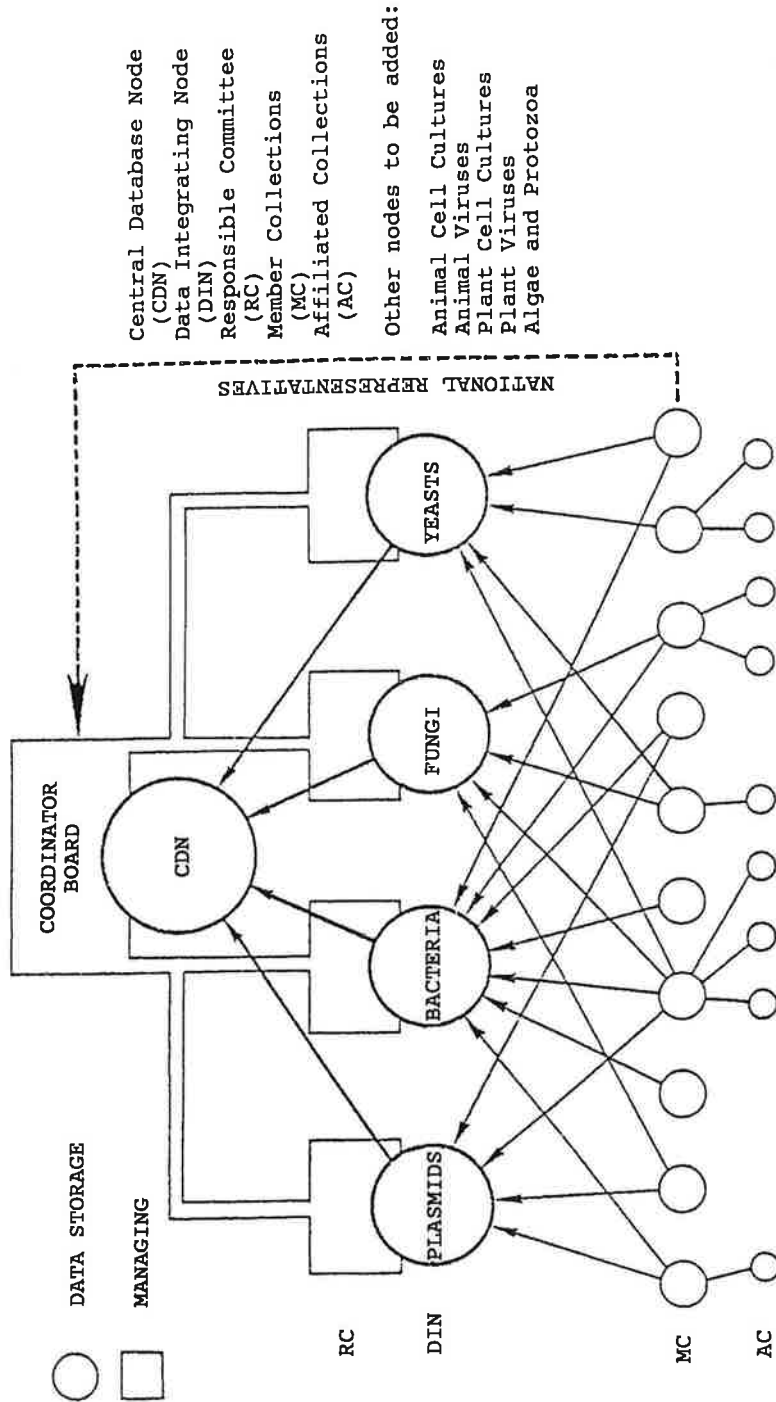


Fig. 2. - Proposed scheme of MINE (March 1989, drawn by R. Kokke).

References

- CARMICHAEL, J. W., W. B. KENDRICK, I. L. CONNERS & L. SIGLER (1980). Genera of Hyphomycetes. – Univ. of Alberta Press.
- COMMISSION OF THE EUROPEAN COMMUNITIES (1989). European Laboratory without walls, in the field of MINE, the Microbial Information Network Europe. – Bruxelles, 20 pp.
- ERIKSSON, O. E. & D. L. HAWKSWORTH (1988). Outline of the Ascomycetes – 1988. – *Systema Ascomycetum* 7: 119–315.
- GAMS, W., G. L. HENNEBERT, J. A. STALPERS, D. JANSSENS, M. A. A. SCHIPPER, J. SMITH, D. YARROW & D. L. HAWKSWORTH (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.
- HAWKSWORTH, D. L., B. C. SUTTON & G. C. AINSWORTH (1983). AINSWORTH & BISBY'S Dictionary of the Fungi (including the Lichens). 7th Edition. – Commonwealth Mycological Institute, Kew.
- ROGOSA, M., M. I. KRICHEVSKY, & R. R. COLWELL (1986). Coding microbiological data for computers. – Springer, New York.
- STALPERS, J.A., M. KRACHT, D. JANSSENS, J. DE LEY, J. VAN DER TOORN, J. SMITH, D. CLAUS & H. HIPPE (1990). Structuring strain data for storage and retrieval of information on Bacteria in MINE, the Microbial Information Network Europe. – *Syst. Appl. Microbiol.* 13: 92–103.