The microbial strain data network

M.I. Krichevsky[1], E. Ross[2], C. McManus[1] & B. Kirsop[2]

[1] Microbial Systematics Section, Epidemiology and Oral Disease Prevention Program, National Institute of Dental Research, National Institutes of Health, Bethesda, MD 20892 USA

[2] MSDN Secretariat, Institute of Biotechnology, Cambridge University, 307 Huntingdon Road, Cambridge CB3 0JX, UK

Krichevsky, M.I., E. Ross, C. McManus & B. Kirsop (1991). The microbial strain data network. – Sydowia 43: 123–134.

The Microbial Strain Data Network (MSDN) is a distributed network of users, databases and communication services endorsed initially by three component organizations within the International Council of Scientific Unions: the World Federation for Culture Collections (WFCC), the International Union of Microbiological Societies (IUMS), and the Committee on Data for Science and Technology (CODATA). Recently, the Committee on Biotechnology (COBIOTECH) joined the others. In spite of its name, the MSDN holds no strain data per se. Instead, a Central Directory points to collections of such data. A number of major culture collections provide their catalogues on line either directly or through automatic gateways. A link to the World Data Center on Microorganisms makes that on line directory available as well. Databases of interest in biotechnology are available and more are planned. The general services expected of a full service on line information system (literature searching, travel information, news services, computer conferencing, bulletin boards, etc.) are available as well. In addition to the on line services, MSDN helps develop and promotes standardization of computer data management of strain data. Training courses are given in electronic communication techniques and facilities and strain data acquisition, management, analysis and transmission. Computer programs for these purposes are distributed as shareware by the MSDN. The MSDN has an ongoing collaborative initiative on building conventions for computer coding of mycological data. The participation by the community of mycologists is invited and welcome.

Keywords: databases, fungus collection, computer application, microbiology.

Traditionally, we seldom communicate the detailed laboratory observations we make of microbes to others. The results of groups of observations are accumulated, summarized, and conclusions drawn. We form the summaries and conclusions into a structured format and publish these as a „paper". For example, we describe a „species" in terms of summarized lists of characteristics which we deem relatively consistent for the taxon. These descriptions may be statistical (as feature frequencies) or natural language phrases wherein we use words such as positive, usually, negative, seldom, variable.

The formalism of the published paper, structured into Abstract, Introduction, Methods and Materials, Results, Discussion, and References, evolved from informal narratives such as Leeuwenhoek's letters to the Royal Society of London. Now we have our formal formats to structure the published communications, peers to comment and evaluate, editors to accept or reject, learned societies and publishing houses to establish and disseminate the periodicals, and bookshelves in laboratories and libraries to archive these summaries.

This process is useful for disclosing generalities among groups of organisms. Classification of strains into groups (taxa) allows us to think in terms of relationships, trends, evolution, ecology, epidemiology and all the other phenomena involving classes of organisms. The disclosures and insights contained in the papers' conclusions are the raison d'être for their publication.

This open literature forms a poor source of information that requires knowledge of the specific traits of a single strain of microorganisms. Very few papers contain comprehensive descriptions of individual strains. At best, a paper may describe a particular trait or restricted complex of traits (e.g., a metabolic pathway or set of genes) for one or more strains.

Culture collection catalogues provide the best available printed source of the characteristics of individual strains. The major service collections of the world publish listings of their holdings. For economic and other reasons, the strain descriptions are limited at best to special or unusual features of the strains, brief histories, growth conditions, and reported utility. As with any printed database, comprehensive searching of culture collection catalogues is difficult. Printed indices are inherently inflexible. The catalogues are out of date as soon as they are published.

The vast majority of microbial strains reside in collections other than those of service collections. They may be institutional collections of very large size (such as large governmental research laboratories, industrial research and development facilities, academic departmental collections) or the personal collections of individual microbiologists. Almost none of these latter collections publish and distribute catalogues or even simple lists of their holdings.

The current emphasis on the potential of biotechnology and the widespread use of computers and digital communication provide the motivation as well as the solution for ready access to detailed descriptions of individual microbial strains. Persons requiring microbes with specific properties often have not the resources or time to isolate the desired strain from nature. Rather, the strain might be found by a search of the strain descriptions held by culture collections.

Rationale for a microbial strain data network

In 1979, such reasoning led Professor O. EL-TAYEB, then at UNEP, to search for a database describing microbial strains that would be useful in biotechnology. Since no such database was known to exist, a round table was held at the International Congress of Microbiology in 1982 to consider the utility of such a database. The participants concluded that such a database would be useful.

A workshop, under the auspices of UNEP, the Commission of the European Community, and the Belgian Science Policy Office was held in Brussels in November, 1983 and a second one in Bangkok, Thailand in November, 1984 (HILL & KRICHEVSKY, 1985). The network design and initiation of the MSDN resulted from these workshops.

The first workshop was attended by forty-three people, two thirds microbiologists and the rest computer specialists of various kinds. The group concluded that assembling any significant portion of the world's strain data in a single place was not practical. Constructing a pointer system to the location of existing sources of strains and strain data was preferable as being more attainable. Hence, a distributed Microbial Strain Data Network (MSDN) was conceived. The good offices of three component organizations within the International Council of Scientific Unions: the World Federation for Culture Collections (WFCC), the International Union of Microbiological Societies (IUMS), and the Committee on Data for Science and Technology (CODATA) were sought. All three organizations agreed to endorse the creation of the MSDN. To this end, CODATA established a Task Group on the MSDN to oversee the creation of the MSDN.

The MSDN was to establish a network with the strain data residing and remaining in the collections themselves and a Central Directory which described the data elements that each collection assesses in describing their strains. In effect, the collection need only to submit a blank record form or list of table headers to the MSDN for incorporation into the directory. In spite of the name, the Microbial Strain Data Network collects no strain data per se.

The central directory

The underlying logic rests on the concept that most searches will be for collections that have rare types of information. For example, consider finding an organism that effects the bioconversion of phenanthrene into some other product. The Central Directory would be interrogated for collections measuring the ability to utilize phenanthrene. If a collection does not make the measurement, it would be unproductive to ask if they had such organisms. Inquiry need only to

be directed to those collections assessing such an ability. Of course, a trivial question would be the ability to utilize glucose as this is likely to be measured by all collections holding heterotrophic organisms.

Standardization is compelled by the requirement that the Central Directory be searchable in a controlled, uniform manner to be useful at all. Thus, the Directory is composed of two parts.

The first part is a controlled vocabulary of legal search parameters. Each term consists of a statement with its identifying number. The vocabulary file defines strain characteristics in very precise terms and codes them numerically for storage in the directory. The coding system has been published (ROGOSA & al., 1986) and is known as the RKC Code. The vocabulary file is searched first by simple string matching on the words of interest. The statement numbers for the items of interest are noted. These numbers are used as search parameters in the second database, the actual descriptions of the collections. For example, the search on „alanine" yields a number of records as presented in Tab. 1.

The second part contains „administrative" information in a header for each collection record. Such elements as the name and address of the collection, the contact persons for the collection, and the general category of organisms, which may or may not be a taxonomic category, are contained in a record header which can be

Tab. 1. — Result of a search on „alanine" in the MSDN database.

MSDN VOCABULARY CODES
JULY 28, 1988

SECTION 29 AMINO ACID UTILIZATION

At least 1 amino acid	029001	L- Asparagine	029015
D- Alanine	029002	L- Aspartic acid	029016
L- Alanine	029003	Betaine	029017
DL- Alanine	029552	DL- Carnitine	029620
b- Alanine	029004	L- Citrulline	029570
m- Aminobenzoic acid	029005	DL- Citrulline	029018
p- Aminobenzoic acid	029006	Creatine	029019
DL-a- Aminobutyric acid	029007	Creatinine	029642
gamma- Aminobutyric acid	029008	L- Cysteine	029020
e- Aminocaproic acid	029009	L- Cystine	029021
delta- Aminolevulinic acid	029630	D- Glutamic acid	029022
DL-a- Aminovaleric acid	029010	L- Glutamic acid	029023
delta- Aminovaleric acid	029011	L- Glutamine	029522
Anthranilic acid	029012	Glycine	029024
L- Arginine	029013	Hippurate	029025
DL- Arginine	029014	L- Histidine	029026
		Homoserine	029651

listed separately. The numerical codes for the data elements recorded in strain descriptions by the collection comprise the body of the record. Thus, the searcher enters the numbers to identify records of interest but lists only the header information to identify the collections of interest. A search on the string „029003" from the record above, yields the list of collections presented in Tab. 2.

The collection identifying information can be listed separately from the rest of the record. The list of six digit numbers comprises a list of the characteristics assessed for the yeasts by the specific collection. Another collection will likely be described by a different set of characteristics.

Tab. 2. — Results of a search on the string „029003" in the MSDN database.

MSDN

4 DECEMBER, 1989

CL YEASTS
DI CECT
AD Collección Española de Cultivos Tipo
AD Departamento de Microbiología
AD Facultad de Ciencias Biológicas
AD Universidad de Valencia, Burjasot
AD Valencia, Spain
CO Professor F. Uruburu
E-MAIL 42:CDT0428 (Dialcom) TEL 346 3864300
TELEX FAX 346 3637649

001995	001996	001997	002001	002004	002009	002012
003001	003004	003007	003023	003027	003028	003146
003186	004008	004009	004029	004030	004031	004037
004038	004039	008020	008342	008344	008398	008456
008634	008635	008636	008638	008639	008640	008641
008662	008663	008684	008705	008706	008713	008715
008716	008719	012021	012022	014001	014003	014004
014008	014011	014012	014018	015004	016028	016030
016031	016041	016043	016044	016046	016047	016137
016138	017015	018032	018034	024138	024148	024248
024252	024463	025007	025012	025020	025022	025029
025030	025036	025037	025038	025039	025040	025041
025042	025043	025044	025195	025213	025252	025253
025269	025270	025271	025272	025273	025274	025275
025276	025284	025296	025297	025300	025302	025307
025310	025311	025313	025326	025327	025328	025329
025330	025331	025332	025333	025334	025341	025342
026352	026355	025365	025367	026515	026556	026559
026567	026575	026578	026589	028002	028022	028027
028054	028057	028072	028134	028164	028173	028717
028723	029003	029008	029013	029016	029023	029024
029026	029030	029037	029041	029044	029045	029049
029570	030025	030141	030427	034103	034109	034143

The evolution of the MSDN

The original design of the MSDN concentrated primarily on the creation and use of the Central Directory. To this end, the CODATA Task Group established various Committees. A Secretariat was envisioned to support the workings of these Committees. The Technical Committee was charged with the task of finding a suitable electronic mail system to allow the Committees and the Secretariat to communicate efficiently. The Training and Education Committee evolved into a panel of faculty members to conduct training courses related to the aims of the MSDN. The current aims have evolved from these beginnings into a complex of services to the microbiological and biotechnical communities.

The aims of the MSDN are five-fold. The first is to build the Central Directory to be used to locate strains with specific properties; the second is to provide storage and access to other databases of importance to biotechnology and microbiology; the third is to provide an electronic communications service. The combination of these three provides a comprehensive information system. Using and promoting an integrated set of descriptors for properties of microorganisms facilitates the fourth aim of promoting data communication standards in microbiology. Conduct of training courses, consultation on communication, support to establishment of cooperating regional networks, and dissemination of strain data-related computer programs all serve the fifth aim of raising the level of use of computer and digital communication technology by microbiologists and scientists and technologists in related disciplines.

These aims are reflected in the following services available through the MSDN Network:

- Database of information sources about microorganisms and cultured cells with specific properties (the Central Directory)
- Access to other scientific databases
- Facilities for providing international access to databases
- Electronic communications (electronic mail, telex, fax)
- Electronic bulletin board and conferencing
- Online culture ordering
- Training courses in the use of computers in culture collections and microbiology
- Micro-IS software distribution
- User support
- Access to Telecom Gold and BT Tymnet services
- Access to scientists on other Telecom Gold or BT Tymnet systems

Today, the MSDN continues to have the endorsement of CODATA, WFCC, and IUMS. To ensure flexibility and sustainability,

CODATA suggested, upon establishment of a permanent Secretariat, that the MDSN become an organization in its own right. Thus, the MSDN is an independent non-profit organization that is managed by an international Committee of Management. Its constitution ensures that all proceeds shall be retained for the development of the MSDN. The database and other „products" of the MSDN are recognized as the property of the international scientific community.

MSDN Nodes

Nodes may be individual research workers, laboratories, institutes, culture collections or data centres. The only requirement is that they are willing to answer queries about the data they record. There is no obligation to supply cultures. There is no obligation to be computerized.

To contribute to the information network, laboratories need only supply copies of any uncompleted data sheets or laboratory notebook pages or record sheets used to record data. These provide the MSDN with information on the kinds of data recorded by individual nodes. The collaborating nodes benefit by making the kinds of information they record available to the scientific community. Additionally, they themselves gain access to the kinds of data held by other nodes. Furthermore, they may use the telecommunications network described below for rapid communication with fellow scientists worldwide.

Other databases

The MSDN provides international access to other scientific databases. These are either stored on the MSDN computer or are accessed through an electronic gateway to another computer. Other databases available at the end of March, 1990 include (* accessed via electronic gateway):

- Hybridoma Data Bank – contains data on publicly available immunoclones and their products (USA, Europe, Japan nodes)
- Information Centre for European Culture Collections (MiCIS and DSM databases)
- American Type Culture Collection (ATCC) Recombinant Clones and Libraries
- * National Collection of Yeast Cultures (NCYC) Computer Services
- * Netherlands Culture Collections Databases (CBS/NCC) (filamentous fungi, yeast, bacteria)
- * World Data Center on Microorganisms

- Culture collection catalogues:
  ATCC Animal Cells
  ATCC Algae and Protozoa
  ATCC Bacteria
  CAB International Mycological Institute
  European Collection of Animal Cell Cultures
  * UK National Collection of Food Bacteria
  * UK National Collection of Yeasts (NCYC)
- Telecom Gold and BT Tymnet Databases

The MSDN is currently working with other collections and organizations to provide international access to their databases through the MSDN network. These include other service culture collections as well as more general scientific database services.

MSDN computer system

The MSDN information network uses the commercially available telecommunications network systems, BT TYMNET and TELECOM GOLD. The advantage of using such systems is that they are readily available to scientists in industry, universities and research institutes throughout the world. Moreover, they provide data storage and backup systems combined with an established electronic mail system that is available to all parts of the world having access to the International Packet Switching System (IPSS). In addition, users have access to the services provided by BT Tymnet and Telecom Gold and are able to communicate with others using the same system. All MSDN services are accessible either through the BT Tymnet or Telecom Gold networks.

Electronic communications

Electronic communications facilities provided by the MSDN include:

- Electronic mail
- Telex and fax
- Bulletin board
- Computer conferencing

Currently over 400 scientists and administrators from 28 countries are linked through the network. Users are from academic institutions, administrative organizations (CEC, ICSU, IUBS, WFCC, CODATA) and industry. The electronic communications system is used for sending messages to colleagues, for committee work, for

joint publications, data transfer, administration and arranging meetings. The service is particularly valuable since it provides an efficient and economic means of locating information and participating in international scientific activities.

The electronic mail system may be used for transfer of bulk data as well as messages. Thus, if collaborating laboratories wish to share data, this is easily accomplished. If a joint publication is being prepared, joint editing is easy. The system has been found to be invaluable for the arrangement of international meetings and conferences, where organizers are in widely separated countries, operating in different time zones.

The electronic mail service is also used to order cultures from service culture collections through online sales mail boxes. Currently, 10 major culture collections provide this service. It is expected that others will follow as the use of the MSDN increases and the advantage of online ordering becomes apparent.

Telex and Fax services are available in the same way as electronic mail. Users on other Dialcom systems may be contacted from the MSDN network.

Bulletin board services are available for publicizing meetings, publications, and other announcements. Computer conferences may be set up by groups for discussions on specialized subjects. These services are managed by the MSDN Information Officer.

Software and technical support

The Secretariat of the MSDN provides support for accessing and using its services and also distributes the Microbial Information System (MICRO-IS) software. The MICRO-IS is a comprehensive data management system especially designed for strain data. It is available as shareware (i.e., at cost plus handling and postal charges).

Training

The MSDN conducts training courses at locations throughout the world. The courses provide instruction in the use of computers in microbiology and for database construction and management, culture collection administration, catalogue production, online communications and database searching. Courses have been held in Czechoslovakia, Brazil, USA, Guatemala, USSR and Egypt. Additional courses are planned. Individual training is also available if appropriate support can be obtained.

Regional collaborative networks

Plans are underway to set up MSDN regional offices to provide first language support. These regional offices will provide more personal contact with users and will facilitate the location of laboratories which have data that could be included in the Central Directory. Also, the services available through the MSDN will become better known through regional meetings and publications. In collaboration with the International Affairs Committee of the American Society for Microbiology, a regional network covering the Americas is being established. The Center for International Projects of the USSR State Committee for Environment Protection has agreed to a collaborative initiative for a regional network for Eastern Europe.

Using the MSDN

Computer facilities are not required in order to collaborate with the MSDN. Laboratories wishing to contribute to the Central Directory need only send their blank record sheets to the Secretariat of the MSDN, together with basic information on the categories of cell types studied and contact information.

Communication with the MSDN is available to all through conventional mail and telephone. Those wishing to be linked electronically to the system for access to the databases and other computer services, or for communications purposes, will need to contact the MSDN Secretariat to be registered and supplied with documentation, an ID number and password.

Computer requirements for linking to the MSDN Network are a computer (either mainframe, mini or micro), a modem, communications software and a Network User Identification number (NUI, obtainable from the local telecommunications office). The Secretariat of the MSDN or computer departments will be able to advise on the necessary requirements to be linked to the system. Over 400 people are already linked, including people from developing countries, proving that it is not difficult to become part of the network. For those without access to the necessary computer facilities, the MSDN Secretariat will carry out database searches upon request.

Standards for coding mycology strain data

The MSDN promotes the use of communication standards in microbiology. The RKC Code for coding of microbial strain data into computers is one of the main elements in facilitating standardization. A second is the distribution of the MICRO-IS computer pro-

grams which use the RKC Code in managing strain data. The RKC Code has grown to encompass, with varying degrees of coverage, most cellular microorganisms: bacteria, fungi, protozoa, and algae. The main thrusts of the current activities are to expand the descriptors for fungi, include biotechnological (application) descriptors, allow for incorporation of genetic information into a strain record, formalize the coding of plasmid data, and begin to interface the system with virus descriptions.

The RKC Code is a statement oriented vocabulary to meet the needs of the microbiologist for a defined context for each feature. Further, as discussed previously, synonyms are in frequent use since no accepted standard is available. We use numbers as the base on which to build the Code as well as in the MICRO-IS. Numbers are difficult to recall while sitting in front of a computer. We cannot realistically define a universally understandable (intuitive) set of mnemonics for more than 12,000 features.

The RKC Code is an open-ended list. While it is reasonably comprehensive with respect to the common yeasts, black yeasts, and the genus *Phythophthora,* much work remains to be done in mycological coding. For example, the richness of metabolic end products largely awaits codification. The editors of the Code are actively soliciting help from the community of mycologists to aid in this effort*.

Acceptance of the RKC Code and the MICRO-IS as de facto standards depends on wide spread use. The RKC Code is distributed through publication as a book (Rogosa & al., 1986) and supplementary articles in journals (Daggett & al., 1980; Philpot & al., 1982; Jong & al., 1988; Jong & al., 1989). We are establishing a mechanism for distribution of the updated Code in ASCII text on floppy disks as well. Selected portions are accessible online to the public. Both the disks and online access are through the facilities of the MSDN**.

---

* The Editor of the RKC Code is Dr. Candace McManus, Microbial Systematics Section, Epidemiology and Oral Disease Prevention Program, National Institute of Dental Research, Park 5 Building Room 451, National Institutes of Health, Bethesda, MD 20892 USA

** For further information on the MSDN and its services contact MSDN Secretariat, Institute of Biotechnology, Cambridge University, 307 Huntingdon Road, Cambridge CB3 0JX, UK. Telephone: 0223 276622. Fax: 0223 277605. Telex: 812440 CAMSPL G. Telex via Telecom Gold: 265451 MONREF G ATTN 10075:DBI0001 or DBI0005. Electronic mail: TELECOM GOLD/DIALCOM 10075:DBI0001 or GOLD/DIALCOM 10075:DBI0005. JANET: MSDN@UK.AC.CAM.PHX. EARN: MSDN%CAM. PHX@EARN. RL. AC. UK

References

DAGGETT, P., M.I. KRICHEVSKY, M. ROGOSA, J.D. CORLISS, & J.P. GIROLAMI (1980). Method for coding data on protozoan strains for computers. – J. Protozool. 27: 353–361.

HILL, L.R. & M.I. KRICHEVSKY (1985). Needs and Specifications for an International Microbial Strain Data Network. – Proceedings of a Workshop held in Brussels, Belgium 15–17 November 1983 and Executive Summary of the Working Group Meeting, Bangkok, Thailand, 23–25 November 1984. – UNEP, Nairobi, 1985.

JONG, S-C., H.H. HO, C. MCMANUS & M.I. KRICHEVSKY (1989). Computer Coding of Strain Features of the Genus *Phythophthora*. – Binary 1: 187–193.

—, L. HOLLOWAY, C. MCMANUS, M.I. KRICHEVSKY, & M. ROGOSA (1988). Coding of strain features for computer-aided identification of yeasts. – Mycotaxon 31: 207–219.

PHILPOT, C.M., M. ROGOSA & M.I. KRICHEVSKY (1982). Coding of phenotypic data descriptive of selected groups of fungi for entry into computers. – Int. J. Syst. Bacteriol. 32: 175–190.

ROGOSA, M., M.I. KRICHEVSKY & R.R. COLWELL (1986). Coding Microbiological Data for Computers. – Springer-Verlag, New York, 299 pp.