



THE WHO BIOMEDICAL RESEARCH INFORMATION SERVICE

Research in medical and related sciences is growing at an enormous rate. It is said that quantitatively speaking, its growth is exponential, doubling itself every 10 years. This appears true, if one is to judge by the number of publications describing the results of such research. But the growth is not only quantitative. More important perhaps than the increase in the number of investigators and research projects, is the increase in the number of sciences and disciplines applied to the solution of biomedical problems. Thus, the traditional biologist is no more the central figure in the so-called biological research. He now shares his place with physicists, chemists, mathematicians and others. New disciplines such as biophysics, molecular biology, biomathematics, radiobiology, etc. have sprung up like mushrooms in an attempt to combine biological approaches with the theoretical sciences.

A scientist today has to keep up not only with progress in his own field, but also with progress in many other related fields. Furthermore, he requires rapid access to developments which may affect the progress of his own work. In other words, the problem facing the scientist is one of communication. Traditional methods of communication do not meet the new needs of scientists and of science administrators. New methods are under investigation.

The advent of the electronic computer with its potential for handling large masses of material at great speeds, has created new possibilities in the field of communication. A good example of these possibilities is the MEDLARS system of the United States National Library of Medicine which deals with medical documentation.

One aspect of biomedical research that awaits solution is information about ongoing research. True enough, most of this research will find its way into the medical literature, in due time. However, the lag period is likely to be fairly long, often 2-3 years after the research is completed. In an attempt to collect information on research in progress and make it available to scientists and research administrators, the World Health Organization has established, on a pilot basis, a Biomedical Research Information Service. The Service is designed to provide comprehensive and up-to-date information about current activities in the biomedical field. The information collected includes data about research institutions, departments, scientists and their current research projects. The data is coded, analysed and stored on large capacity random access disk files and magnetic tapes connected to an IBM system 360 computer. An up-dating system allows for the whole material to be revised and up-dated at annual intervals.

The Method

1. Collection of Information

The information is collected through a series of questionnaires as follows:

Questionnaire A: Data on research organizations and administrations at the national levels (ministries of science, research councils, academies, etc.). To be completed by secretary or administrative officer.

RC/70.2

The issue of this document does not constitute formal publication. It should not be reviewed, abstracted or quoted without the agreement of the World Health Organization. Authors alone are responsible for views expressed in signed articles.

Ce document ne constitue pas une publication. Il ne doit faire l'objet d'aucun compte rendu ou résumé ni d'aucune citation sans l'autorisation de l'Organisation Mondiale de la Santé. Les opinions exprimées dans les articles signés n'engagent que leurs auteurs.

Questionnaire B: Data on research institutions (institute, school, university department, etc.). To be completed by the director of the institute.

Questionnaire C: Data on research projects by units (departments, divisions, laboratories, etc.). Project titles and names of investigators are requested. To be completed by the unit chief or director.

Questionnaire D: Biographical data on scientists mentioned in questionnaire C as engaged in a research project. To be completed by the individual scientists.

It will be seen from the above system that no one person is asked to complete more than one questionnaire, with the possible exception of the director of an institute or unit chief, and then only if he or she is engaged at the same time in a research project. Those people would have to complete questionnaire B or C and questionnaire D. The questionnaire method for the collection of information has its drawbacks. However, preliminary testing of the method has shown that 60-70 per cent. replies can be expected within a reasonable period. By further contacts such as reminders, personal appeals, etc. one has been able to raise the percentage of response to about 80 per cent. It is doubtful whether this can be improved further.

2. Coding and Analysis

The material is coded and analysed prior to storage. Numerical code lists were prepared for the various questions.

The coding of data about institutions and scientists presents no special difficulty. The indexing of the projects, however, presents innumerable difficulties. An indexing system was developed after two years of study in order to solve, if not all, at least most of the difficulties encountered.

The main problems presented fall into two categories:

1. The material for indexing consists of project titles. In many instances these titles are complete and allow for full indexing. In other instances, however, the titles are couched in very general terms. Thus while one scientist may report a study on the "role of aflatoxin in the development of hepatic carcinoma in human beings", another may report simply "studies on the aetiology of liver carcinoma". The two scientists may actually be following the same lines of research and be interested in the same factors.

The retrieval system must be so designed as to retrieve both projects under a general heading of "Aetiology of liver carcinoma", and the first project under the specific heading of "Aflatoxin".

2. The study of almost 2000 queries received by WHO reveal that the scientists tend to frame the questions either in very general terms such as "cancer chemotherapy" or in extremely specific terms such as "chemical reactions of the nuclei of lymphoma cells compared to similar reactions in normal lymphoid cells".

It is evident that under these conditions no indexing system can satisfy all retrieval requirements. We have, however, endeavoured to devise a system that would allow for the maximum retrieval possibilities of the information in our possession. Furthermore, the retrieval possibilities are expected to improve as scientists are educated into employing fuller replies to our questionnaires.

Description of the Indexing System

The principle used in building up the indexing system is the categorization of each project title under four different headings, namely:

1. The Field of Activity (FA)
2. The Disease Orientation (if any) (DO)
3. The Test Object (TO)
4. The Technique Used (TU)

Lists were prepared for each of the above 4 headings. In preparing the lists, a digital decimal hierarchical system was used, namely, a main heading subdivided with up to 9 sub-headings which themselves can be further subdivided. Each subdivision allows for a higher degree of specificity.

These 4 principal lists were built with the idea that each project title should theoretically supply answers to the following questions:

1. What is the general field(s) of activity or discipline(s) within which the particular research project falls? (List FA)
Example: Biology, Chemistry, Genetics, Immunology, etc.
2. Is the research oriented toward the solution of a particular disease or group of diseases? (List DO)
Example: Cancer, Parasitic Diseases, Diseases of the Circulatory System, etc.
3. In the performance of research, what object is being tested or what phenomenon is being examined? (List TO)
Example: Tissues, Cells, Laboratory Animals, Drugs, Chemicals, etc.
4. What are the techniques used in the study? (List TU)
Example: Histological Techniques, Physical Techniques, Chemical Techniques, etc.

In addition to the above 4 categories, a list of "Keywords" or "Descriptors" as prepared. The system therefore combines the principles of content analysis with those of keyword analysis, the latter being a mere supplement to allow for a still higher degree of specificity. However, the two systems are fully integrated. The Descriptor List (DE) follows a simple serial number given to each descriptor as it is met during the actual indexing process. However, in order to facilitate the indexing and retrieval processes, the descriptors are also listed under a general category within one of the main lists. Thus glucuronic acid carries the descriptor number DE 1235, and also is listed under the general heading of "carboxylic acids" in the Test Object List (TO 521240). In the Descriptor list, glucuronic acid will appear as follows:

TO 521240 - DE 1235 Glucuronic Acid

The indexer will therefore code glucuronic acid both in the TO and DE lists. Similarly, Trypanosoma gambiense would be coded as follows:

TO 272004 (Trypanosoma) - DE 0222 (Trypanosoma gambiense)

This system allows therefore for a highly specific retrieval when the descriptor is specifically mentioned. On the other hand, the same projects will be retrieved under a more general heading whether the particular descriptor is specifically mentioned or not (see Annex I - Indexing Instructions).

Advantages and Disadvantages of the System

To begin with the disadvantages, the system, at first sight, appears to be complicated and cumbersome to use. The indexer must consult 5 different lists in order to code each project. Furthermore, it demands that the indexers be highly qualified in the biomedical sciences. Even a qualified person requires a period of at least 6 months in order to master the indexing techniques. And finally, regardless of the training, indexers will differ in their choice of codes to a certain extent. In other words, there will be some individual variation. However, in our two years' experience, we found that this variation is not likely to affect the system seriously, as the variations are quantitative rather than qualitative. Thus, one indexer may have the tendency to use a larger number of codes for the same project in order to increase the retrieval possibilities, while another may be content with a smaller number of codes. On the whole, however, both use the same major code numbers. In some cases the correct categorization of a project may require expert advice, though the indexers are supplied with standard dictionaries and text books.

Turning now to the advantages, one may say that the main advantage of the system is its flexibility. It can be expanded and improved at will without affecting its usefulness. It does not require the preparation of fully comprehensive lists. Subdivisions of the various headings in the main lists can be easily made if and when required, while the descriptor list is built up serially as the specific keywords are met with.

The system is adaptable to the needs of the users in as far as the degree of specificity required is concerned. Thus, the digital system may be simplified by the use of smaller numbers of digits, if a high degree of specificity is not required. On the other hand, the number of digits may be increased to allow for a higher degree of specificity. Thus, the Disease Orientation list (DO), which is based on the International Classification of Diseases, is presented here as a six digit decimal system. As an example, Diseases of the Cervix Uteri is coded DO 122243, using up all the six digits. More specific terms like "Cervicitis", or "Cervical Stricture" are given descriptor numbers. However, if it is so desired, a seventh digit may be added to accommodate these specific terms.

In order to avoid overloading the list, some terms of a very general nature are so placed that they may be combined with other terms. Thus the code for Ecology is FA 07010, while the code for Marine Biology is FA 07210. By noting both codes, one can retrieve all the projects which involve the Ecology of Marine Organisms. Similarly, the code for Cell Division is TO 013000. The code for Reticular Cells is TO 114211. The combination of these codes enables the retrieval of Division of Reticular Cells.

3. Storage

The material after coding and analysis is stored as follows:

(a) The names and addresses of institutions and departments, as well as other data found in questionnaire B, are stored on a random access storage unit (disk). This unit also stores the biographical data on the individual scientists.

(b) The research projects are stored on a magnetic tape. The project title in full narrative form together with the identifying code number are stored in sequential order.

(c) On a second random access storage unit (disk) we store a dictionary of descriptors with a listing of all associated project numbers.

(d) The numerical code of the content analysis is included in the magnetic tape mentioned above, with each project.

4. Retrieval

A dual retrieval system is used, namely retrieval by content and retrieval by descriptors (keywords).

This dual system enables us to retrieve projects under general headings, such as the field of activity, the technique, etc. or, on a more specific basis, by using the descriptors.

The retrieval programme prepared includes 36 computer programmes, which are designed to allow for retrieval under various categories of questions.

It is envisaged that questions put to the system may include such queries as: "Who is doing what and where?", "What institutions are engaged in this or that type of research?", "Who is using this or that technique?", "What is the research potential of this or that country?", "How many scientists are engaged in this or that field of biomedical research?" and so on and so forth.

ANNEX I

Indexing Instructions

1. Every project should be coded as far as possible under the 5 coding lists, namely, the Field of Activity (FA), the Disease Orientation (DO), the Test Object (TO), the Technique Used (TU) and the Descriptors (DE).

2. A project may be coded several times within the same coding list to a maximum of 8 times per list, thus making a maximum of 40 code numbers per project.

3. In selecting the code numbers, it is best if one selects the most specific number under a particular heading. Thus "Radiation Carcinogenesis" should be coded as FA 34132. It will, however, be retrieved under FA 34130 (Physical Carcinogenesis), FA 34100 (Carcinogenesis, General) and FA 34000 (Oncology). There is no need to use the more general code number if a more specific code number within that heading is available.

4. The 5 lists may be used in combination. Example: "Neurophysiology" would be coded as follows:

FA 43000 (Physiology), TO 160000 (Nervous System)

Similarly "Pathology of Oropharyngeal Tumours" would be coded:

FA 39000 (Pathology), DO 011100 (Oropharyngeal Tumours).

5. There are a number of general terms which have been placed separately for combined indexing. These terms fall into two groups.

(a) General terms that apply to all categories. These are grouped separately in the FA list under two headings, namely, FA 98000 (Types of Research) and FA 99000 (General Category). Concepts such as "Clinical Studies", "Experimental Studies", "Epidemiological Studies", etc. are grouped together under the heading "Types of Research", while the heading "General Category" includes concepts such as "Diagnosis", "Therapy", etc. Ideally, each project should be coded under one or both of these two headings in addition to the more specific coding. Thus, "Tumour Transplantation in Rats" would be coded as an "Experimental Study" (FA 98120); while "Atromid in the Prevention of Ischaemic Heart Disease" would be coded both under "Clinical Studies" (FA 98110) and "Chemoprophylaxis" (FA 99311).

(b) General terms that apply to a particular category. These are placed immediately below the main category heading. Example:

010000 CELLS
010100 Normal Cells
010200 Malignant Cells
011000 CELL ORGANELLES
011130 Chromosomes

Under this system, "The Study of Chromosomes of Malignant Cells" would be coded as follows:

TO 010200 (Malignant Cells), TO 011130 (Chromosomes)

The same terms may be used in combination with codes from other categories. Thus, "Malignant Cells in pericardial fluid" would be coded:

TO 010200 (Malignant Cells), TO 111310 (Pericardial fluid)

6. In some cases, a certain concept may fall under more than one category. In order to avoid confusion or having the same concept coded differently by different indexers, the list indicates the need for double coding by the sentence "Code also". For example, "Pharmacological Standardization" is placed under the general category of "Standardization and Reference Activities". However, it is important to retrieve it under "Pharmacology". Therefore, it appears in the list as:

FA 01420 Pharmacological Standardization (Code also Pharmacology, NEC FA 40900).

EXAMPLES

- I. ROLE OF AFLATOXIN IN THE DEVELOPMENT OF HEPATIC CARCINOMA IN HUMAN BEINGS
- FA 98110 Clinical Studies DO 010001 Malignant Tumours TO 210000 Man TU 611000 Surveys DE 0154 Aflatoxin
- FA 98200 Epidemiological Studies DO 010130 Carcinomas TO 333221 Aspergillus TO 332900 Plant Products, NEC
- FA 99010 Aetiology DO 011500 Liver Tumours TO 382900 Plant Products, NEC
- II. STUDIES ON THE AETIOLOGY OF LIVER CARCINOMA
- DO 010001 Malignant Tumours TO -- TU -- DE --
- DO 010130 Carcinomas
- DO 011500 Liver Tumours
- III. CANCER CHEMOTHERAPY
- FA 99220 Chemotherapy DO 010001 Malignant Tumours TO -- TU -- DE --
- IV. CHEMICAL REACTIONS OF THE NUCLEI OF LYMPHOMA CELLS COMPARED TO SIMILAR REACTIONS IN NORMAL LYMPHOID CELLS
- FA 98120 Experimental Study DO 010001 Malignant Tumours TO 010100 Normal Cells TU 200000 Biochemical DE 0148 Malignant
- FA 98410 Comparative Studies DO 017100 Lymphoma TO 010200 Malignant Cells and Chemical Lymphoma
- FA 07490 Cell Biology, NEC TO 011100 Cell Nucleus Techniques
- FA 06900 Biochemistry, NEC TO 114130 Lymphocyte
- V. STRUCTURE OF FRIEND VIRUS
- FA 98120 Experimental Studies DO 010001 Malignant Tumours TO 313030 Virus Structure TU -- DE 0598 Friend
- FA 07430 Molecular Biology DO 017400 Leukaemia TO 313820 Mouse Leukaemia Virus
- FA 28020 Microbial Structure
- FA 28400 Virology
- VI. COMBINATION SURGERY AND RADIOTHERAPY IN THE TREATMENT OF EARLY BRONCHOGENIC CARCINOMA
- FA 98100 Clinical Studies DO 010001 Malignant Tumours TO 219200 Hospital Patients TU 413400 Surgical DE --
- FA 99201 Combination Therapy DO 010130 Carcinomas Removal
- FA 99240 Radiotherapy DO 012210 Bronchial Tumours TU 340000 Radiation Techniques
- FA 99250 Surgery DO 991200 Early Disease States
- FA 47000 Radiology
- FA 50020 Cardiovascular and Thoracic Surgery
- VII. PRODUCTION OF SKIN TUMOURS IN THE HAIRLESS MOUSE BY APPLICATION OF TOBACCO SMOKE CONDENSATES
- FA 98120 Experimental Studies DO 010001 Malignant Tumours TO 221241 Mice TU 900000 Techniques, DE 1084 Hairless
- FA 34120 Chemical carcinogenesis DO 014100 Skin Tumours TO 363213 Nicotiana other Mouse
- DO 019200 Experimental Tumours TO 382111 Tobacco Smokes Constituents DE 0161 Topical Application