

WORLD HEALTH
ORGANIZATION



ORGANISATION MONDIALE
DE LA SANTÉ

HPP/WARC/1.65

ORIGINAL: FRENCH

THE INTERNATIONAL CANCER RESEARCH AGENCY
PROPOSAL FOR THE INFORMATION DEPARTMENT

Working paper prepared by Mrs M. Wolff-Terroine
Chief, Documentation Unit
Institut Gustave Roussy, Villejuif (Seine)

The information problem is a serious one encountered in all fields of knowledge. Everywhere the volume of literature is increasing exponentially, everywhere research workers are submerged under a flood of publications, and, as a result, lose considerable time before finding the relevant information, even then often incomplete.

It is for this reason that, for some time, numerous specialized services have been developing, responsible for supplying information to those interested.

Information retrieval is now a difficult problem, even in fields with quite well-defined limits, and becomes extremely complex in the case of a multi-disciplinary field, such as cancer research. Cancer, regarded as a whole, covers a very large number of extremely varied disciplines, ranging from pathology to the basic sciences such as biochemistry, virology, immunology, radiobiology, etc. . . . Research workers in this field are obliged to consult innumerable journals and many indexes without being sure, however, that nothing has escaped their notice, since the fields of investigation are varied and only partially overlap.

Thus it is particularly important, when the establishment of an international cancer research agency is envisaged, to make provision at the outset for an information department in such an agency. This study will endeavour to make clear what the activities of such a department might be and how they might be carried out, bearing in mind what has already been done in this field.

ACTIVITIES OF THE INFORMATION DEPARTMENT

Briefly, the role of an information department is to bring together original information (no matter of what nature) and to transform such information into working tools of various kinds with the aim of making it easier for research workers to collect data in a given field or on a given scientific question.

A. The information collected would be:

- (1) Published information: articles in journals, books and reports of congresses dealing with clinical and experimental aspects of cancer.
- (2) Unpublished information: the information department would collect all data concerning research work under way.
- (3) Information on research workers and institutions investigating cancer problems.
- (4) Information concerning all congresses, symposia and seminars during which cancer problems are discussed.
- (5) Precise information might also be collected with the aim of replying to definite questions on a given subject, such replies being direct ones and not lists of references.

It must be decided what "finished products" are required on the basis of the information collected; there are numerous possibilities here:

Indexing periodical (based on titles or articles)

Abstracts journal

Regular supply of the latest information to subscribers in their particular fields

Replies to specific queries

Publication of current bibliographies

Periodic publication of lists of congresses, research workers, research centres . . .

The choice should be made in accordance with needs and financial possibilities (it should be noted that, on the whole, such choices are not mutually exclusive). However this may be, whenever information in a scientific field is concerned, two "finished products" seem essential: rapid publication of an indexing periodical, and retrospective searches concerning a given subject on demand.

ACTIVITIES AND METHODS

The conversion of such information into "finished products" involves a series of different activities.

- A. Collection of documents.
- B. Analysis of the documents to extract the information and present it in the most suitable form for subsequent use.
- C. Processing of the information: i.e. the preparation of various "reference tools" from the results of the preceding analysis.
- D. Dissemination of the information: i.e. communication of the data obtained after the processing operations to the research workers concerned.

Each of these activities has its own special problems: in certain cases there are various possibilities between which a choice must be made from the outset.

- A. Collection: this has two essential aims.

- (a) Exhaustive literature coverage:

Some scientists are opposed to the principle of comprehensiveness in the collection of documents: they believe that a specialized documentation centre should deal only with the most important documents and restrict itself to collecting information coming from certain countries only or certain journals only. This attitude, although defensible in principle, is nevertheless a very dangerous one, for it is impossible to define which countries produce and which journals publish significant work, such an assessment being inevitably somewhat subjective in nature. Moreover, it would seem absurd to place national limitations on an international agency. Consequently, we feel it preferable to aim at the greatest possible comprehensiveness in collecting the documents, perhaps introducing an element of selectivity in their analysis.

(b) Rapidity

Information which is late has often lost its value for the research worker: rapidity is therefore absolutely essential. The very long delays usually intervening between the publication of an original document and any kind of reference to it (title, summary, review) are only too well known. Every effort should be made in the collection stage and in subsequent stages to reduce this delay. In particular, the International Cancer Research Agency might very well be informed by journals or by authors of all papers on cancer in course of publication. Similarly, the Agency could be informed as rapidly as possible of all research work in progress (see Part 3).

B. Analysis

Document analysis is an operation whose aim is to represent a given document in a form different from the original (translation, abstract, indexing, . . .) so as to make it easier for the specialists concerned to consult or find it.

Here we are confronted by several possibilities.

(a) Should a summary of every document be prepared?

(b) Should every document be indexed by means of key words or key phrases? Should such indexing be based on the full text, the summary or the title? Should the indexing be done by man or by machine?

(c) Should the titles of the documents be translated or only their index formulation?

(a) We shall not deal here with the theoretical problem of the usefulness of abstracts and of their cost. The field of cancer is, in fact, already very well covered by various abstracting journals, chiefly in English. These journals, which may be national, such as the Referativnyj Zurnal, the Bulletin Signalétique du C.N.R.S., or international, such as Excerpta Medica, or may deal with a well-defined subject, such as

Carcinogenesis Abstracts, Cancer Chemotherapy Abstracts, Leukaemia Abstracts, Berichte über die Gesamte Biologie, Biological Abstracts . . . etc., are, on the whole, all of high quality. There would therefore seem to be obvious overlapping here and it appears necessary to rule out straight away the publication of an abstracting journal as an activity of the information department.

(b) Indexing of all the documents: despite technical advances it is still out of the question to store in a system the full text of a document, and even if this were feasible, it would still be necessary to describe the text using standardized concepts. Expensive though it is, indexing is therefore indispensable for the storage of the information in a document.

(a) Mechanical indexing: this involves picking out automatically from the titles of documents certain terms regarded as important; tables are then compiled in which the documents are listed by main words, by authors, by collections . . . (e.g. KWIC index). This indexing is fairly easy to do because the relevant "programme" can be purchased.

A number of valid criticisms can be made of this method: in particular it is based solely on a reading of the title and is consequently often superficial (see statistical test carried out at I.G.R. on 1000 documents on cancer of the breast and lung); moreover these indexes are not very convenient to use.

(b) Manual indexing is the representation of the content of a text by a series of key words or key phrases taken from a lexicon or from an artificial language specially designed for data processing in a given scientific field. The compilation of such a lexicon adapted to cancer research is a difficult operation. The various existing lexicons are much too general (see communication of the M.D. Anderson Hospital at the Twenty-Sixth Annual Meeting of American Documentation Inst., Chicago, 1963). So far as we know, only one lexicon exists, that of the Gustave Roussy

Institute; it includes a thesaurus and has been in use for two years. It could be supplemented by the thesaurus still in process of preparation at the M.D. Anderson Hospital on radiobiology. The indexing methods must accordingly be specified, from the point of view of both lexical and syntactical problems (see "Plan of implementation", below).

(c) Translation can be envisaged at different levels: translation of the title or translation of key words. A pragmatic approach should be adopted to the solution of translation problems.

Theoretically it is perfectly possible to translate all the titles and all the terms indexed. In practice, this is difficult to do. Apart even from financial considerations (specialized translators are scarce and expensive), if it is accepted that speed is essential for the efficient running of an information department of this kind, the idea of translation into several languages must be abandoned and a single working language must be adopted. As matters now stand, it seems that this working language should be English. It would be necessary, therefore, to provide for systematic translation of all titles into English and direct indexing of articles in English. On the other hand, the publications of the information department not dealing with scientific information in the strict sense (lists of research workers and of scientific institutes, programmes of congresses . . .) could be published in the official languages of the International Agency.

C. Processing

After analysis, all the data must be coded for storage, and afterwards they have to be scanned when they are to be used (scanning will usually consist of comparing the terms of a "question" with the terms found in the indexed representation of documents in the system). At the present stage of data-recording techniques it seems necessary to provide for automation of data storage and retrieval from the outset (see under "Equipment and staff").

It should be noted that for bibliographical references, provision must be made for storing not only the indexed and coded representation of each document, but also its bibliographical reference and its full title: this will avoid a list of document numbers being produced by the tabulator - a method frequently employed which involves laboriously looking up references and then identifying the relevant document numbers extracted and their exact bibliographical reference.

D. Dissemination

The various methods of dissemination have already been mentioned: indexing periodical, document bibliographies, information supplied on request or subscriptions for recurrent bibliographies on special subjects . . . But two other types of service, independent of these, must be provided for:

- reproduction of the original documents, and
- their possible translation.

EQUIPMENT AND STAFF

Equipment

The problem of equipment will only arise after the indexing and coding phase.

Since these machines are developing very fast, it would be wise to provide for a storage system permitting an easy change-over from initial pilot equipment for study purposes to more developed facilities.

As things are at present, truly automatic indexing being still in the study stage (indexes of the KWIC types being screenings and tabulations rather than real "indexes"), manual indexing appears to be indicated, automation only beginning after a hand-written indexing card has been typed out. The system would be fed by means of cards, or rather punched tape. An automatic processing unit would be provided, with "conventional" data-processing machines, a punched-tape machine with auxiliary reader and auxiliary perforator, and a medium-power computer would be hired during a preliminary study phase.

All things considered, in regard to printing, the size of the Agency does not seem to warrant the use of machinery for type-setting direct from the material produced by the tabulators. Instead, a contractor would be commissioned to make offset reproductions of the informational material produced by the printer.

It would be advisable, however, to have an office offset machine for publications of small impressions.

Duplicating machines are necessary, and a very wide range of types (photocopy, xerograph, etc.) should be considered before making a selection.

Technicians

Staff varying considerably in qualifications will be required to carry out these different tasks.

For collection, a medical and scientific librarian is necessary.

Data reading can only be entrusted to competent medico-scientific staff. To be really competent, they must be conversant with the problems involved. It is hard to imagine doctors working full time for an information department, for they would very soon become too much out of touch with current problems.

In addition, a large number of specialties must be catered for: there must be biochemists to evaluate an article on biochemistry and virologists to evaluate an article on virology, not merely staff with a good general medical education. This specialist staff, itself engaged in research, will be required to make not only a descriptive analysis of the documents, but also an evaluation of them based on the following criteria: contribution of new findings, contribution of new theories, general reviews, popular articles.

For data processing on the other hand, staff will be required with a sound basic medical and scientific education and a training in data processing methods, as well as staff for operating the machines.

For scanning the stored information specialized analysts and technical staff are required.

If however the acquisition of a computer by the Agency from the outset is envisaged, it must be realized that all computers are not equally well suited to automatic documentation operations. Without wishing to enter into technical matters here, it must be mentioned that the computer selected from the electronic machines then available should have the following features:

- memory able to organize by individually addressable groups of units;
- very short access time;
- high speed of transfer inside the machine, and
- printer with a large variety of types.

Thus it would clearly be necessary to engage specialists capable of running a computing centre and of preparing and testing programmes.

ORGANIZATION AND STRUCTURE

Collection of information

The information department will have a large library, subscribing to the most important journals.

It will establish a system of relations with the research institutes. The research institutes will become its correspondents and will send it such documents as it does not receive on the spot.

Through its contacts with the various institutes it will be kept up to date with work in progress, the names and qualifications of research workers, etc.

Indexing will be as centralized as possible. It may be roughly estimated that the Agency should have among its scientific staff specialists able to analyse about two-thirds of the data. The remaining one-third, not suitable for analysis at the Agency on account of its specialized nature or for linguistic reasons, would be entrusted to correspondents: the latter would send the Agency a sort of telegraphic summary which would be indexed at the Agency (after a long "running in" period, correspondents might possibly be able to index by themselves: it was five years before Viniti ventured to make a similar attempt!).

Thus the automatic indexing, processing and scanning of documents must be centralized, for indexing is very specialized work which requires a working knowledge of a documentary language which has its own rules.

Since coding, processing and scanning are the essentially mechanical phases of the operation, they should be carried out where the equipment is located. It should, moreover, be pointed out that whatever lexical or syntactical refinements may have been introduced beforehand, scanning can only be performed satisfactorily - particularly in the case of replies to occasional questions - in close liaison with specialists in the field in question. (The fact is that to whatever extent indexing may have been standardized, there is rarely a one-one correspondence between the wording of a question and the way in which the most relevant documents have been recorded. The use of a thesaurus with its system of cross-references to cognate words and extended meanings facilitates collation of "questions" and "documents"; but in practice it is often only by consulting a specialist in the appropriate field that a limit to these extended meanings can be set, since, at the "growing points" of research, terminology is always essentially fluid.) This confirms the need for a specialized information centre to work as closely as possible in conjunction with the various specialists.

Dissemination

Since the main purpose of publishing a descriptive index, i.e. a list of classified references, is to provide rapid information, publication at weekly or fortnightly intervals is indicated. This, in conjunction with the fact that the majority of the documents will be indexed for recording and storage at the Agency itself (it is assumed that the Agency will be capable of indexing two-thirds of the documents and that only one-third will be analysed by correspondents), will make it possible for information on new publications to be supplied with considerable rapidity.

If this solution is rejected, a double system of indexing must be provided: a preliminary rapid indexing by the Agency staff for the purpose of rapid recording; and a second thorough indexing by specialists in the appropriate fields. This solution might perhaps be somewhat more rapid but would be less specific, more cumbersome and more costly.

Requests for information require two different types of service: replies to occasional specific questions; and the supply to subscribers of regular information on the latest developments relating to a specific subject.

Other publications should be considered: (lists of institutions, congresses, etc., see above) and a decision would have to be taken regarding their frequency.

PLAN OF IMPLEMENTATION

The creation of such an information service requires careful preliminary study, as regards both methods and organization. Time, money and energy could doubtless be saved by drawing - within limits that would have to be laid down - on what has already been done in this field. At present there is only one such centre practising the systematic storage and mechanical retrieval of information (it also publishes an index and retrieves information on request), namely the Gustave Roussy Institute at Villejuif, France, which has produced a lexicon and a thesaurus: however, the biochemical section is inadequate and is still in preparation. The M.D. Anderson Hospital at Houston is considering the preparation of a thesaurus on radiobiological terms used in cancer research, on the basis of a limited number of documents. The Chester Beatty Institute, London, publishes a classified monthly list of new publications. The National Cancer Institute is preparing an extremely detailed classification of all work carried out with the funds it has provided. There are certainly many other institutions which look for index documents relating to cancer, but cancer is not their prime concern and the vocabulary they use is never sufficiently specific.

What form then, in the light of what has already been done, should the preliminary studies take, and what should be the first practical steps?

Collection

The staff of the future information department would begin by drawing up:

- (1) a list of periodicals necessary for the Agency's library;
- (2) a list of the libraries of the main cancer research centres;

- (3) a permanent register of centres, laboratories and research workers (advantage could be taken of the relevant WHO lists);
- (4) a permanent register of research in progress (see the work carried out by WHO).

These stores of general information should be built up and processed automatically.

Analysis

The fundamental and very long term task of preparing a lexicon and thesaurus could largely be avoided by using the thesaurus of the Gustave Roussy Institute. Some points of detail might be improved by further specificity. The preparation of a biochemical thesaurus should be speeded up; advantage could be taken of the work of the M.D. Anderson Hospital in radiobiology.

It would be valuable to establish for this purpose a simple system of classification to be used in the indexing periodical: that of the Gustave Roussy Institute could be supplemented by that of the Chester Beatty Institute.

Methodological studies should also be undertaken to test the value of indexing methods; the use of a simple co-ordinate indexing system or the introduction of simple or complex syntactical factors; in this connexion use might be made of the studies now in progress at the Gustave Roussy Institute, where the value of the thesaurus and the validity of different systems of indexing in the fields in question are being studied by statistical methods.

It would further be necessary to make a list of the special subjects covered so as to decide which fields should be covered by the information service of the International Centre itself, and those for which use would have to be made of correspondents.

Processing

For the initial phase, a machine of the Flexowriter type with auxiliary punching unit and a data-processing shop of the conventional type, should be regarded as a minimum. At this stage there is no need to install a computer;

use could be made of the services provided by the big manufacturers of such equipment on a job basis. To save money and gain practice, it might be useful, in a preliminary study phase, to use the equipment of existing centres with experience of data processing.

At this stage, consideration should be given to the equipment that will be necessary at later stages. If a computer with the capacity specified above cannot be regarded as a practical possibility for the International Agency, the question of having the data processed by some institution (WHO, I.G.R., Chester Beatty) possessing the desired type of computer should be considered and a suitable system of programming worked out.

Dissemination

The types of publication produced by the information department will have to be studied in this connexion, and it must be decided whether to start an entirely new indexing periodical or simply to extend existing services.

The necessary duplicating and printing machines must be selected and purchased.

According to the range of questions considered (and the funds available), this preliminary study phase will last from one to two years.

It is practically impossible to give even approximate figures for the budget of an information department of this type. As has been said, a wide variety of choices, both theoretical and practical, are open, which makes it difficult to give any estimate, so that we cannot guarantee the accuracy of the note attached to this report.¹

Thus the setting up of an information department of this kind is a delicate problem, both in conception and execution. Whatever the decisions taken and the choices made, two points seem to be of prime importance for the quality of the future work: a thorough study of lexicological problems; and the recruitment of genuinely competent specialists, themselves engaged in research, to carry out the analysis.

¹ To be sent under separate cover.

If the essential methodological and intellectual prerequisites are thus fulfilled, and if the mechanical equipment is adequate, such a department would be of the greatest value to all those engaged in the struggle against cancer.