A Century of Science Publishing
E.H. Fredriksson (Ed.)
IOS Press, 2001

Chapter 16
Medical Databases: Medline versus Excerpta Medica

Robert R. Blanken[a] *and Pierre J. Vinken*[b]
[a]Former Executive Chief Editor, Excerpta Medica,
Amsterdam, The Netherlands
[b]Former Chairman of the Board, Elsevier and Reed-Elsevier,
London, UK and Amsterdam, The Netherlands

1. Prehistory: the situation up to and immediately after World War II

During the 19th and the first half of the 20th century, efforts to make the world's medical literature more accessible to the end-user were dominated by the German *Zentralblätter* in Europe and the various publications of the Surgeon General's Office/Army Medical Library/National Library of Medicine and the American Medical Association in the United States. Later, these were complemented (and to a certain extent copied) by the *Bulletin Signaletique* in France and the *Referativnye Zhurnaly* in the Soviet Union.

Credit for the first major attempt to index the world's medical literature must be given to John Shaw Billings, who was responsible for the creation of the *Index-Catalogue of the Library of the Surgeon General's Office*, first published in 1880. This monumental work, the first series of which was finally completed with volume XVI (W to Zythus) in 1895, ultimately contained subject entries for 168,537 books and 511,112 journal articles, and author entries for 176,364 books; its major handicap, however, was the lack of currency that inevitably resulted from its alphabetical setup. In 1879, the cumulative alphabetical catalogue was therefore complemented by the first volume of the *Index Medicus*: *a Monthly Classified Record of the Current Medical Literature of the World*. Financial problems, however, caused this first attempt to fail (temporarily) after 21 volumes, in 1899. Following a three-year intermezzo during which the *Bibliographia Medica* was published in Paris, publication of the *Index Medicus* in the U.S. was resumed in 1903 and continued until 1927.

Meanwhile, in 1916, following caustic criticism of both the *Index-Catalogue* and the *Index Medicus*, the American Medical Association started publishing the *Quarterly Cumulative Index to Current Medical Literature*, which was to appear in

parallel with the *Index Medicus* for the next 43 years. The *Index-Catalogue* also continued into a second series (21 volumes, 1896–1915) and even a third (10 volumes, 1918-1932), thanks to the vigorous support of the Medical Library Association, but the setup and goals gradually changed and subject entries to the current literature were no longer included after 1925. In 1927, following prolonged negotiations, the *Index Medicus* and the *Quarterly Cumulative Index to Current Medical Literature* were merged to yield the *Quarterly Cumulative Index Medicus*; the cooperative effort ended, however, in 1931, after which the QCIM was continued by the American Medical Association.

Following a brief hiatus, the 'Friends of the Army Medical Library' together with the Medical Library Association started publishing the *Current List of Medical Literature*, a classified listing of the tables of contents of journals received in the library; monthly subject indexes were added in 1945, so that the *Current List* and the QCIM became more or less competing services, while the plans for future series of the *Index-Catalogue* were scrapped. Following its modification in 1950, the *Current List* contained a register section listing articles in serial numbered order under their journal titles, plus an author index and a subject index using standardized headings, clued to the descriptive bibliographic data by means of the serial numbers. These standardized subject headings became the basis for the later *Subject Heading Authority List* and still later *Medical Subject Headings* (MeSH). The American Medical Association finally discontinued publication of the QCIM in 1959, after which the Current List was transformed into the *Index Medicus* as we know it today.

An important characteristic of all of the above publications is that only bibliographic information was provided; although sometimes indexed in depth, the journal articles were not abstracted and no summaries were included. In Europe, the practice was quite different. The venerable, German-language *Zentralblätter* published abstracts of the current biomedical literature in a series of specialized, classified abstract bulletins intended for use by the medical specialist. In the prewar days, when most scientific articles did not even contain a summary, let alone an abstract, this meant an intellectual abstracting effort that inevitably delayed the appearance of the bibliographic reference. The *Zentralblätter* were therefore used more for retrospective searches and the compilation of bibliographies than for current awareness, a concept that had not yet acquired the importance we give to it today.

An essentially similar service, in Russian, was provided by the Soviet government in the *Referativnye Zhurnaly*, also a series of specialized, classified abstract bulletins, not only in medicine but in all areas of science. Originally intended for

domestic use in the USSR, and later disseminated (also in the form of tapes) to the 'satellite' countries of Eastern Europe, the *Referativnye Zhurnaly* attracted the serious attention of the West, and especially the U.S., after the launching of the first Sputnik. During the next decades, the U.S. government sponsored extensive translation programs of the abstracts in the *Referativnye Zhurnaly*. In Eastern Europe, however, preference was generally given to information from the West; in the medical area, this meant a preference for *Excerpta Medica* and (when available) *Index Medicus* above the comparable Soviet sources.

To a small group of German Jewish publishers who had been hidden in Amsterdam during World War II, it was clear that the key role of the *Zentralblätter* would not survive the defeat of Germany and the ascendance of English as the international language of science. The lack of abstracts in the *Current List* and QCIM meant that there was an urgent need and potential market for an English-language medical abstracting service. As a result, *Excerpta Medica* and its original series of 15 specialized abstract bulletins, with the slogan "By the medical specialist, for the medical specialist", was born.

The original basic concept of the series of semi-independent, specialized abstract bulletins or 'sections' bearing the joint name *Excerpta Medica* was very similar to that of the *Zentralblätter*. Each bulletin, with names such as 'Anatomy, Anthropology, Embryology and Histology', 'Physiology and Biochemistry', 'Endocrinology', 'Dermatology and Venereology' or 'Chest Diseases, Thoracic Surgery and Tuberculosis', had its own specialist editorial staff that was responsible for selection, classification and (ultimately) indexing, its own international editorial board of specialist advisors, and its own international staff of volunteer abstractors. The some 3000 journals then received regularly in Amsterdam were microfilmed with the collaboration of the Royal Netherlands Academy of Sciences and placed in a dated cabinet in the editorial offices of *Excerpta Medica* where they were screened by the specialist 'section editors' to select articles relevant to their abstract bulletins. Later on, historically speaking, this process was supplemented by a team of 'assignment editors' who wrote relevant section numbers in the table of contents, to be checked later by the section editor. After one week, the journals received on a particular day were taken out of the cabinet, torn apart into individual articles, and the articles were stapled to an indexing and abstracting form that also contained the section numbers, in presumed order of priority, to which the article had been assigned. The article then began on a sometimes-lengthy voyage through the hands of multiple section editors, each of whom added the necessary classification codes and index terms from his specialist point of view. The first editor who decided that the abstract should be published in his bulletin was responsible for creating

the abstract, either by modifying the English summary, by selecting portions of the text, by sending a foreign-language summary out to be translated, or by sending the entire article to one of several thousand volunteer abstractors, many of whom were also members of the international editorial boards. Ultimately, the product of this work was checked by a native speaker of English, after which the abstract was ready for publication in all sections that had selected it. In the case of multidisciplinary articles, this sequential processing sometimes meant that more than a year elapsed between journal receipt and publication of the abstract. Nevertheless, *Excerpta Medica* soon acquired a large number of individual subscribers, as well as medical libraries.

Actually, during the first few years of *Excerpta Medica*'s existence, the monthly abstract bulletins did not contain a subject index; retrieval was by means of decimal classification systems, specific to each section, with a cumulative subject index published at the end of each year. Initially, the terminology used in these annual indexes was also subject-specific and thus varied from abstract bulletin to abstract bulletin; all this was to change radically after the appointment of Pierre Vinken (a neurosurgeon who ultimately became the Chairman of Elsevier after the latter acquired *Excerpta Medica* in 1971) as President and Chief Editor in the 1960's, marking the beginning of the efforts at professionalization that would continue unabated for the next 30 years. The philosophy underlying *Excerpta Medica*'s ultimate system of semi-controlled subject indexing was that the specialist editor should be left free to write down all those terms he considered necessary to represent the content of the article, at the level of specificity he deemed necessary, using the preferred terminology of his subject specialty, and that these suggested terms would be controlled afterwards ('à posteriori') against a growing 'thesaurus' (initially a separate one for each specialty) to eliminate synonyms and spelling errors. This was felt to be preferable to a forced choice of indexing terms from a pre-existing list, such as was the case at the National Library of Medicine, since it not only spared the valuable time of the medical specialist but also enabled a much more rapid response to specific new concepts (such as drugs) appearing in the medical literature. The indexing was precoordinate, with a preferred length of two or three words per index term. In order to prevent an all too rapid growth of the thesaurus, there were some limiting rules or guidelines. Thus, all terms were in the singular noun form, in American spelling, and in the natural word order rather than rotated; furthermore, a philosophical distinction was made between 'primary terms', under which a reader could be expected to look in a printed index (names of diseases, anatomical terms, drugs, etc.), and 'secondary terms' such as child, diagnosis, treatment, the names of experimental animals, routes of drug administration,

etc. that are usually significant only when combined with a primary term; indexing terms in which a primary term and a secondary term are combined were generally forbidden. Only the primary terms were controlled, and in the printed subject index, only the primary terms created separate alphabetical entries, followed in each case by the other primary terms and then by the secondary terms. These secondary terms or 'secondary text' created a kind of 'mini-abstract' that frequently included quantitative information or more detailed findings or methods. An example of such a rotated index entry is shown in the box below.

brain infarction, carotid stenosis, hydrocephalus, radiodiagnosis, bilateral, 4-year-old child, 114
carotid stenosis, brain infarction, hydrocephalus, radiodiagnosis, bilateral, 4-year-old child, 114
hydrocephalus, brain infarction, carotid stenosis, radiodiagnosis, bilateral, 4-year-old child, 114

Although this approach to indexing had definite advantages over the use of a pre-existing thesaurus, there were practical problems that only really became apparent after automation and the integration of the separate thesauri into one 'Master List of Medical Terms' or 'Malimet'. First of all, as we shall see below, the thesaurus showed an irrepressible tendency to grow much too fast, with the addition of terms that were used extremely rarely, and secondly, the use of this increasing number of highly specific terms turned out to be relatively inconsistent, with similar articles often being indexed in different ways; although not disastrous in a printed subject index, this created difficult problems later in the training of users for on-line retrieval.

2. Automation, professionalization, and thesaurus development

Given the geometric increase in the volume of medical literature, together with the increasing cost of human labor and the need for more rapid access to biomedical and especially pharmacological information, it soon became clear to everyone that automation was the answer. It was expected that automation of the production systems for indexes and abstract journals would also create improved possibilities for information retrieval. In the United States, however, the initial experience with a punched-card system was disappointing in this respect.

In the 1940's and 1950's, the Welch Medical Indexing Project gave high priority to the development of machine methods for the production of the *Current List*, together with more theoretical work on coordinate indexing and the development of a 'Subject Heading Authority List' that was ultimately to result in the first MeSH. By 1960, a production system had been created consisting of Flexowriter composing machines for the index copy, IBM keypunches and sorters for alphabetizing the

copy, and a Listomatic step-and-repeat camera for composing column-width film. It quickly became evident, however, that this was inadequate as a retrieval system. The fastest card sorter then available could handle only 1000 cards a minute; to search a 5-year file containing 750,000 subject cards would take some 12 hours. A far-reaching decision was therefore taken to invert the objectives, i.e. to design a retrieval system from which a publication system could later be derived. This ultimately led to the computer-based system now known as MEDLARS (Medical Literature Analysis and Retrieval System).

Early in 1964, a system based on a Honeywell 800-200 computer and GRACE (Graphic Arts Composing Equipment) for phototypesetting was able not only to produce *Index Medicus* but also to perform experimental retrospective searches in batch mode. By 1968, there were 12 search formulation centers in the U.S. and several in foreign countries, and searches were being performed on four computers outside the National Library of Medicine. Meanwhile, in the latter half of the 1960's, plans were being made for the on-line input of indexed material, and the first experiments were carried out with on-line retrieval. In 1971, MEDLINE (MEDLARS On-line) was officially started, and a new system based on two IBM 370/158 computers, coupled together as a multiprocessor system to operate as one, was delivered in January 1975.

At *Excerpta Medica* in Amsterdam, the problems were similar, but the efforts were concentrated initially on automating the production of the abstract journals and standardizing the indexing terminology, rather than on retrieval as such. Between the late 1940's and the middle 1960's, the volume of the *Excerpta Medica* operation had increased significantly. Thanks partly to subsidies and stimulated by pressure from both specialist subscribers and specialist editors, several new abstract journals came into existence and a number of larger 'sections' were split to yield more specific daughters. The volume of literature processed had also increased significantly, to about 250,000 articles (100,000 of them with abstracts) per year, derived from some 3500 journals yielding ca. 20,000 individual issues annually; all of these articles were indexed on the basis of a thesaurus that had in the meantime grown to about 400,000 terms (preferred terms and synonyms), and classified in a polyhierarchic system containing more than 3500 'pigeonholes' at four levels in 39 independent 'sections'. The need for cost reduction coupled to the desire to provide better and more rapid access to medical information led to the decision to transform the loosely connected series of manually published abstract bulletins into an integrated, uniformly indexed and classified, electronic database from which the abstract bulletins could be obtained as a by-product.

In 1965, Pierre Vinken contacted Frans van der Walle, the director of a small

software development company named Rescona; this contact was to lead to the creation of Infonet, which was given the contract to create *Excerpta Medica*'s 'Mark I' system for the automated storage and retrieval of biomedical information.

At that time, computer technology was still in its infancy. Standard word-processing software and disk memories did not yet exist and the 'mainframe' processors had a power that is dwarfed by any present-day PC. Electronic phototypesetters had just started to become available. The choice of hardware eventually fell on two NCR 315-501 RMCs (Rod Memory Computers), 10 CRAM-V magnetic card storage systems, and an NCR 321 communications controller. The internal memory capacity of the configuration was 40k. The CRAM-files were equipped with large information cartridges, each containing 384 3.65x14-inch magnetic cards, each containing 144 recording tracks with a recording density of 936 bits/inch and a resultant capacity of 1500 six-bit characters; these cartridges had to be exchanged manually, since two were required for each year of *Excerpta Medica*, but this took less than a minute. Any card from a cartridge could be dropped into read/write position within 125 milliseconds. The developed software was revolutionary, being completely randomly organized and using a CRAM-card storage facility with direct addressing possibilities similar to those of present-day disk systems, together with NCR's 'FAMOUS' index-sequential software package that made the on-line update and recall of the thesaurus for each separate index term possible with response times measured in seconds, at a time when comparable information systems were still completely magnetic tape oriented with thesaurus update runs of some 24 hours (F. van der Walle, personal communication).

The *Excerpta Medica* production system also comprised 16 magnetic tape units and a paper-tape input unit that read 600 characters/second. Although all information for input was normally punched on paper tape, input was also possible via a Micro-Image Card Reader. The software included a systems supervisor that controlled the legitimacy of all input, as well as checking on the presence or absence of certain types of information; its most important component, for controlling indexing input, was based on Malimet and included a program that controlled the logical consistency of the relationships between Malimet preferred terms and synonyms. Finally, there was a publishing subsystem that provided for the compilation of the bibliographic information and abstracts in the database, assignment of abstract and page numbers, make-up of the final pages and compilation of the author and subject indexes for each abstract bulletin, resulting in a magnetic tape that was used to drive a Digiset photosetter.

Another and more intellectual aspect of the automation of *Excerpta Medica*'s production system was the creation of an integrated thesaurus to control the index

terms to be input into the integrated database. As early as 1963, *Excerpta Medica*'s Board of Chief Editors had decided to try to alleviate the existing chaos in medical terminology and the resultant inconsistencies in the subject indexes of the individual abstract bulletins. On the other hand, they had no desire to create an à priori thesaurus from which indexers would be forced to select terms. Taking the 1962 cumulative annual indexes as the starting point, the Chief Editors discussed the how and why of each entry and each cross-reference with the responsible indexers, which soon resulted in a number of small thesauri, one for each medical discipline. However, when an attempt was made to integrate these thesauri, it was found that many of the terms had several different meanings and that the cross-references were often mutually incompatible. Furthermore, the number of terms was so large that even punched cards and conventional IBM equipment did not suffice to control the thesaurus input. A computer program was therefore requested and obtained (see above).

At the end of this initial project, Malimet represented a file of about 25,000 preferred terms and 50,000 synonyms. With this as a starting point, the indexing entries suggested daily by the specialist indexers were checked against this growing authority file. Any term not recognized by the computer was printed out on a weekly 'error list', which was referred to a team of medical specialists who were experienced in the terminologies of all medical disciplines. Each term on these error lists had to be either 'translated' into an existing term or accepted as a new term in the thesaurus. Unfortunately, however, the error lists were so large as to be practically unmanageable in the time available, and to make matters worse, Malimet was not yet available on-line, so that the editing work of this team was based on periodic printouts or (later) microfiche versions that were quickly out of date. As a result, Malimet grew very rapidly and in a somewhat uncontrolled fashion, so that the number of preferred terms soon reached 125,000 and the number of synonyms perhaps twice that. Despite the guidelines referred to above, the transfer of the processing of the error lists to internal staff and the later availability of Malimet on-line, the growth of this à posteriori authority file continued at an alarming rate and a decreasing percentage of the preferred terms were frequently used.

Automation also made possible (or necessary) a number of other changes in *Excerpta Medica*'s production system and retrieval facilities. Thus, the tables of contents of the individual abstract bulletins were cast into a consistent decimal form, if necessary, and integrated into EMclass, a polyhierarchic classification system with a maximum of four levels, the first of which was the section number. The subclassifications within each section remained independent and pragmatic, being

designed to divide the literature into a large number of more or less equal piles rather than to provide a logical breakdown of the field. New subclassifications could be created at any time, although changes in the hierarchic structure were discouraged. In order to make a selected list of secondary concepts retrievable by projected future users of the database, these concepts were given numbers and compiled into the Item Index (later to be known as EMtags); these were terms representing, for example, the type of article, routes of drug administration, age groups, geographic concepts or the names of experimental animals, and were similar to the 'checktags' of Medline.

Since the bibliographic information ('reference' or 'citation') for all selected articles was now input first, before the abstract or any indexing, a new type of product also became possible: the literature index. Among the ca. 250,000 articles selected annually for the database, about 150,000 never received an abstract but would nevertheless be indexed and classified, often by multiple sections. It therefore became tempting to use some of these for saleable products (even before database tapes became a product), and the first such 'literature indexes' to be produced were the *Drug Literature Index* and *Adverse Reactions Titles*. Especially the *Drug Literature Index* (section 37 in the database) was an impressive product, including upwards of 50,000 articles per year, derived not only from the 3500 'normal' journals but also from some 200 specially selected chemical and pharmaceutical journals, and indexed in depth from both a medical, a pharmacological and a chemical point of view (with separate fields, for example, for trade names, manufacturer's names and the Wiswesser Line Notation). As DrugDoc, these two sections would come to represent an unusually valuable portion of the total database.

In the interest of getting the information into at least the literature indexes and the database more quickly, the routing of the articles and indexing forms was also streamlined. Now, instead of a single index form to be used by all the assigned sections, the system responded to the input of the bibliographic information by printing out separate forms for each section, which were attached to the article and sent along to the first editor (this always being the DrugDoc editors if relevant). When the article with its forms was returned (within a strictly controlled time period), the indexing for the first section was immediately keyboarded and sent for input while the article and the remaining forms went on to the second section. As a result of the input of indexing and classifications, the reference became available for printed publications such as the Drug Literature Index and for output onto database tapes. This process was repeated for each assigned section, until ultimately only the abstract form, on which the editors had indicated whether or not they wished to publish the abstract, was left. This abstract could very often be prepared

by the internal abstracting department that had in the meantime been organized. This sequential input of indexing from the point of view of several assigned sections of course meant that the database user (tape subscriber) would receive the same item repeatedly, sometimes with only minor additions. As we will see later, this was a major objection, leading to various attempts to prevent or alleviate it.

3. 'Mark II', on-line access and the battle for currency

Early in the 1970's, the Directors and Chief Editors of *Excerpta Medica* became convinced that the long-term future lay in the sale of electronic information via database tapes, and that the existing production system and the hardware used for it were no longer the most suitable for the purpose. They therefore again turned to Infonet with the request to make an inventory of the problems and ideas in the minds of the *Excerpta Medica* staff and to come up with a concept for a new system. Meanwhile, in line with earlier attempts to professionalize and streamline the processing of biomedical information for the database, two full-time Executive Chief Editors had been appointed to help the Chief Editors (themselves part-time with responsibilities elsewhere) run the system. These two would play an active role in the next two decades in the attempts to accelerate the input of information and at the same time make it more readily retrievable, beginning with a key role in the consultations with Infonet on what would become the Mark II system. Following a detailed analysis of the bottlenecks and the possible solutions in terms of hardware and software, it was decided to abandon the NCR equipment and to replace it with a network of Digital minicomputers, linked together to provide the necessary speed and capacity; it was felt that this would provide greater flexibility, at lower cost, than the choice of a mainframe. The CRAM-cards were therefore replaced by disk drives and magnetic tapes, and the thesaurus control group was given improved access to Malimet. Very soon, the first experiments could also be organized on search formulation for the retrieval of information, as a result of which the quality control over the medical indexers was tightened up.

With a view toward accelerating the input of the abstracts, the role of the volunteer abstractors was gradually phased out; this was made possible by the fact that the overwhelming majority of the articles from important journals now had English-language summaries, combined with an increased contribution from in-house personnel. At the same time, various experiments were made with the input of bibliographic information and indexing at different stages, separately and combined. For many articles from important journals, a bibliographic reference and an abstract were input first, together with the assigned section numbers, providing rapid (albeit unindexed) information for on-line retrieval by means of free-text

searches. In other cases, highly specialized articles from important journals were indexed and classified before input of the bibliographic information, so that everything could be input together. Tape subscribers, however, continued to complain about the multiple receipt of essentially the same information. This was aggravated by the increasingly multidisciplinary nature of the medical literature and by the subdivision of the abstract bulletins into increasingly specific daughters, so that the average number of sections to which an article was assigned tended to increase. Although these more specific abstract bulletins were attractive to the specialist individual subscriber, and the more specific classifications were useful for retrieval, the multiple receipt of the same information was an aggravation to librarians and database managers alike. To try and alleviate this, some arbitrary limitations were placed on the depth of assignment, particularly for articles from less important journals, and attempts were made to group the indexing input for several secondary sections. All of this made for a continuing process of change in the editorial procedures, guidelines and forms.

By the middle 1970's, *Excerpta Medica* was sending computer tapes weekly to a number of pharmaceutical companies and governmental agencies in foreign countries, and EMbase was accessible on-line via providers such as Dialog, DIMDI, DataStar, BRS, STN and JICST. Although the printed abstract bulletins were still the major source of revenue, on-line access was starting to represent an attractive alternative, particularly for the individual end-user. The attention of *Excerpta Medica*'s user training programs was therefore increasingly directed at retrieval from the database, and this in turn had an inevitable effect on editorial procedures and production streams. For example, the primary indexing terms on an article were further subdivided into A-terms and B-terms, depending on their relevance in that article, and only the A-terms were rotated in the printed indexes. The number of EMtags and classification subcategories was increased, and a continuing effort was made to reduce the time between journal receipt and input of indexed references.

4. EMbase versus Medline and the role of user aids

This period also witnessed the appearance of several articles in which retrieval from EMbase was compared with that from Medline. The general conclusion from these comparative studies was that the speed of input into the two databases was comparable, that EMbase often yielded more references, particularly in drug-related areas, but that the proper formulation of searches designed to yield comprehensive retrieval with a high degree of relevance was relatively difficult for EMbase. Users often complained about the inconsistent use of specific indexing terms, classifications and EMtags, and about the need to use several alternative formulations

simultaneously in order to retrieve all relevant articles. This underlined the need for user training and for the development of user aids. After all, even though most individual users would probably start with a free-text search of the terms appearing in the titles and abstracts, and although the higher percentage of abstracts in EMbase, combined with the deep indexing using both primary and secondary terms, gave EMbase a certain advantage in this respect, there are many theoretical reasons for not relying entirely on free-text searching if either comprehensive or highly relevant retrieval is desired. Retrieval using the controlled vocabulary is to be recommended, but then one must know how to use it.

The first of the many new user aids produced in the 1980's was the Index to EMclass. The classification categories represented an effective tool for the retrieval of broader concepts (more effective than the very broad Malimet terms), but the polyhierarchic nature of the classification system made it difficult for the on-line user to find the relevant classifications. Similar subcategories could be found in several sections, but their use of course depended on the assignment of the article to those sections, and the point of view of the medical discipline involved was implied in the definition of the classification subcategory. Following many serious discussions with the Chief Editors and section editors, an index was finally produced in which the users were referred from all concepts present in the entire classification system to all relevant subcategories in all sections. This guide was received with enthusiasm by the users.

A list of EMtags and the List of Journals Abstracted (with CODEN-codes and classified according to both subject specialty and country of origin) were of course relatively easy to produce and distribute. In contrast to MeSH, however, which was available in printed form in every medical library, Malimet was only available on-line or on (rapidly outdated) microfiches. Moreover, although control over its growth had been improved considerably, there continued to be a problem with precoordinated terms that were inconsistently used. Analysis showed that of the approximately 150,000 preferred terms then in existence, only about 20,000 were used with any frequency. It was therefore decided to produce a user aid (MiniMalimet) containing the most frequently used terms, and to encourage the more consistent use of these terms by distributing the list to the indexers as well. This list, which was eventually incorporated into a comprehensive *Excerpta Medica Guide to the Classification and Indexing System*, was also received with enthusiasm by users, although less so by the indexers, who felt that their traditional freedom was being curtailed. Simultaneously, an effort was made to limit the addition of new terms to the names of specific concepts such as drugs, syndromes, plant or animal species, etc. and to prohibit the addition of new precoordinated concepts.

This, however, would prove not to be the end of the story. During the 1980's, despite trials with the publication of various new types of printed products and even new abstract bulletins such as 'Toxicology' or 'Forensic Sciences', it became increasingly clear that the printed products were being supplanted by on-line access and electronic spin-offs such as sections of EMbase on CD-ROM or magnetic tape, EMSCOPES, EMbase Alert, etc. Although EMbase continued to be appreciated and used, particularly for the retrieval of drug-related and other highly specific information, users regularly complained of the difficulty of search formulation and the lack of hierarchic structure in Malimet, which made the retrieval of broad concepts particularly difficult. In 1988, therefore, the decision was finally taken to introduce a limited amount of hierarchic structure into (Mini)Malimet and create what was to become EMTREE or EMTHES.

As might be imagined, this was not an easy operation. First of all, a decision had to be taken as to the kind of structure to be introduced. The easiest solution might possibly have been to simply switch to MeSH, especially since this would have satisfied user demands that it should be possible to run searches formulated for Medline against EMbase as well with a minimum of modification. However, this would have meant a radical break with the past, making the information in the existing years of the database more or less irretrievable. It was strongly felt that the existing Malimet terminology had to be preserved, also for the benefit of the indexers. Furthermore, there was a feeling that MeSH, burdened by a history of more than 30 years with little change, could be improved upon in the light of new insights. The compromise reached was to take over the basic superstructure of MeSH, i.e. the 15 categories or "facets" at the highest level, together with some of the first-level subdivisions, and to attach the 20,000 most frequently used Malimet terms, plus 10,000 additional drug names and some terms newly created for 'umbrella' concepts at higher levels in the tree structure, to it. Moreover, MeSH headings would be added as synonyms of the EMTREE preferred terms wherever possible. This gargantuan task was largely accomplished within one year, so that EMTREE could be announced in 1989.

In its present form, EMTREE consists of about 40,000 drug and biomedical preferred terms or 'descriptors', plus over 170,000 synonyms (about 12,000 of which are included in the printed thesaurus); the descriptors are organized into a cascading tree-like structure with a maximum of seven levels of subdivision (15 facets at the top, divided into 127 subfacets, etc.), represented by about 10,000 numerical codes. Of the 40,000 preferred terms, about 5000 are 'explosion terms' with directly equivalent codes that have more specific terms under them, while about 30,000 are specific terms that are posted under broader concepts. Indexers

are expected to use the existing terms consistently, but may of course suggest 'candidate terms' (which are also appended to the reference in the database) that are reviewed regularly for inclusion in EMTREE. The inevitable changes in the hierarchic structure and terminology are carefully documented and announced annually. EMtags and EMclass are no longer used as such by the indexers, but can of course be used to search older EMbase files.

5. The present

At latest reports, EMbase continues to do well in its competition with Medline. The need for speed and economy, combined with the new editorial procedures associated with the use of EMTREE, have resulted in a further streamlining of the production process. All input is now on-line, section assignment and even the generation of many of the tables of contents for the printed products are now automated or combined with input of the bibliographic reference. In principle, bibliographic references ('citations') are input for all articles (about 400,000 per year) in all biomedical journals processed, without selection. Many of these references and abstracts are also obtained in machine-readable form, from Elsevier and associated publishers, and the role of the individual medical editors has been reduced to a minimum. The original concept of "By the medical specialist, for the medical specialist" with which *Excerpta Medica* began more than 50 years ago, has thus been sacrificed to a considerable extent in the interests of speed, economy, consistency and user friendliness.

These days, everyone's attention is on the Internet and the possibilities that this medium offers for the retrieval of information. EMbase is of course also available via the Internet, although not for free. Perhaps more importantly, an increasing number of original journals are also available via the Internet. An example in this direction is the ScienceDirect project of Elsevier Science, which offers Internet access on a subscription basis to a file that currently consists of more than 1.2 million articles. More on medical publishing via the Internet can be found in Ch. 19.

References

[1] Adams, S. & McCarn, D.B. (1976) Chapter II: From Fasciculus to On-Line Terminal: One hundred years of medical indexing. In: *Communication in the Service of American Health — A Bicentennial Report from the National Library of Medicine.* Bethesda, MD: National Institutes of Health.

[2] Anonymous. (1969) *NCR: a biomedical information storage, retrieval and dissemination system.* Amsterdam: Excerpta Medica & The National Cash Register Co.

[3] Anonymous. (1974) *The Excerpta Medica biomedical information system: computer tapes and abstract journals.* Amsterdam: Excerpta Medica.

[4] Blanken, R.R. (1989) Thesaurus. In: *Archiefbeheer in de praktijk.* Alphen aan den Rijn: Samsom Uitgeverij, Band 1 (3075), 1–17.

[5] Garfield, E. (1980) Excerpta Medica — Abstracting the biomedical litreature for the medical specialist. *Current Contents,* **28**, 5–10.

[6] Mehnert, R. National Library of Medicine. In: *Current Practice in Health Sciences Librarianship.* Vol. 7. Bunting, A. & McClure, L.W. (Eds.) Health Sciences Environment and Librarianship in Health Sciences Libraries. New York: Forbes, pp. 91–115.

[7] Vinken, P.J. (1969) Het Excerpta Medica Foundation systeem van informatieverwerking met de computer. *Nederlands Tijdschrift voor Medische Studenten* **15**(6), 276–281.

[8] Vinken, P.J. & Blanken, R.R. (1969) *Illustrated lecture on the Excerpta Medica automated storage and retrieval system.* Amsterdam: Excerpta Medica.

[9] Vinken, P.J. & Blanken, R.R. (1972) *Excerpta Medica. Encyclopedia of Library and Information Science.* Vol. 8. New York, pp. 262–282.

[10] Vinken, P.J. & Van der Walle, F. (1968) *Excerpta Medica automated storage and retrieval program of biomedical information.* Amsterdam: Excerpta Medica Foundation.

[11] Vinken, P.J., Van der Walle, F. & Warren, P.A. (1970) Design and operation of an advanced computer system for the storage, retrieval and dissemination of the world's biomedical information. In: *Proceedings of the Third International Congress of Medical Librarianship, Amsterdam, 1970.* pp. 149–154.