Excellence in Evaluation: Early Landmarks at the National Library of Medicine

Barbara A. Rapp

Abstract

F. Wilfrid Lancaster has earned a reputation for greatness in the evaluation of information storage and retrieval systems. Many of his extensive contributions stem from his early experience with the National Library of Medicine (NLM) MEDLARS system. His evaluation of the MEDLARS Demand Search Service in 1966 and 1967 was an important landmark as one of the earliest evaluations of a computer-based retrieval system and as the first application of recall and precision measures in a large, operational database setting. In 1971, his evaluation of the MEDLARS AIM-TWX system was an important study of early online systems and their direct use by end users. This paper summarizes Lancaster's two major evaluations of the MEDLARS system, including the information environment at the time and their impact in the field of information science. Examples of Lancaster's other evaluation work with information retrieval systems are provided, followed by discussion of the textbooks that grew out of his evaluation experience and expertise. The article closes with comments from current and former NLM staff regarding Lancaster's time at NLM or his influence on their own career.

Introduction

F. Wilfrid Lancaster established himself as a giant in the evaluation of information storage and retrieval systems early in his career, and his reputation for greatness in this arena stands today.

Many of Lancaster's extensive contributions stem from his experience with the National Library of Medicine (NLM) MEDLARS system. As one of the earliest evaluations of a computer-based retrieval system, his evalu-

LIBRARY TRENDS, Vol. 56, No. 4, Spring 2008 ("The Evaluation and Transformation of Information Systems: Essays Honoring the Legacy of F. W. Lancaster," edited by Lorraine J. Haricombe and Keith Russell), pp. 859–887

ation of the MEDLARS Demand Search Service in 1966 and 1967 was widely regarded as an important landmark, earning praise as the "beau ideal" in the *Annual Review of Information Science and Technology (ARIST)* (Brandhorst & Eckert, 1972). A few years later, his evaluation of the MEDLARS AIM-TWX system in 1971 was an important study of early online systems and their direct use by end users.

Lancaster undertook these evaluations in an environment of innovation and rapid change—in computing, in information retrieval applications, in information science research, and in information system evaluation.

In this paper, I first summarize Lancaster's two major evaluations of the MEDLARS system and discuss their impact at NLM and more generally in the field of information science. Next, I provide examples of his other evaluation work with information retrieval systems and discuss the books that grew out of his evaluation experience and expertise, the books that instructed so many of us about information systems—their design, analysis, and evaluation. The article closes with comments from current and former NLM staff regarding Lancaster's time at NLM or his influence on their own career.

Evaluation of the MEDLARS Demand Search Service

Lancaster is widely associated with the MEDLARS evaluation—but what was it and why was it so important?

MEDLARS stands for MEDical Literature Analysis and Retrieval System and was developed to computerize the production of *Index Medicus*, a major printed index to the biomedical literature produced by NLM. The computer-based searching component was called the Demand Search Service. When launched in March 1964, there was no other publicly available, fully operational electronic storage and retrieval system of its magnitude in existence (Miles, 1982).

At the time of Lancaster's evaluation, the MEDLARS database contained about 800,000 bibliographic records from January 1964 forward, growing at the rate of about 200,000 records annually. Articles were indexed from a set of 2,400 journals, using the hierarchically organized MeSH controlled vocabulary that consisted of about 7,000 "fairly conventional pre-coordinate type subject headings" (Lancaster, 1968a). This was an offline, batch search system. Search requests were submitted in writing to NLM staff, who created and entered the search strategies. The searches were then run sequentially against the database tapes.

The Information and Evaluation Environment at the Time

Appreciation of the importance of the evaluation, and why it was influential for NLM and the information retrieval field, may be helped by providing a sense of the information environment at the time and the visibility

of NLM's initiative to provide computer access to bibliographic data.

The use of computers for bibliographic retrieval systems was in its infancy, and many of the extant systems were experimental or small in nature. In their comprehensive history of online information services, Bourne and Hahn (2003) credit MEDLARS as "one of the earliest large-scale online retrieval operations," and describe an environment of tremendous increase in medical research publications and need for more efficient methods of information retrieval. In a recent historical paper on the development of the MEDLARS system, Dee (2007) characterized the environment by saying, "NLM's accomplishments regarding MEDLARS were cutting edge, placing the library at the forefront of incorporating mechanization and technologies into medical information systems" (p. 416). Dee also noted "enthusiastic public interest" in MEDLARS, citing coverage in the *Wall Street Journal* and other newspapers. In his comprehensive history of the NLM, Miles (1982) summed it up by saying "On the whole the system was one of the largest and most successful library automation projects. Its success marked a milestone in the evolution of modern libraries."

The first year of MEDLARS operation was characterized by NLM's Deputy Director Scott Adams (1965) as

> one of intensive trial, test, experiment, evaluation, and change. Internal and external pressures alike have been brought to bear on the system . . . MEDLARS has been highly conspicuous nationally and internationally, and the variety of challenge and the Library's necessarily experimental response have made for an extremely busy year. (p. 139)

There was also high interest among the scientific community in computerized access to biomedical information, as evidenced by the publication in *Science* of a paper on MEDLARS. Coauthored by NLM Director Martin M. Cummings, the paper reported on the first year's experience with automated access (Karel, Austin, & Cummings, 1965). The *Science* paper also foreshadowed Lancaster's formal evaluation and characterized somewhat the environment into which he was recruited. Describing the evaluation approach, the authors wrote:

> Appreciating that there is as yet no wholly satisfactory method of objectively evaluating the effectiveness of information storage and retrieval systems, the library has relied heavily on consumer reaction and appraisal. Evaluation of critical reports indicates that the percentage of missed entries is minimal; furthermore the relevance of retrieved citations as determined by the individual requester's evaluation of demand bibliographies, appears to be satisfactory. New and more precise measurements of relevance are under study. (p. 769)

Why Lancaster?

It was in this environment that the NLM director received a visit from Cyril Cleverdon, librarian of the College of Aeronautics in Cranfield, England.

Cleverdon was well known for his research on evaluating the efficiency and effectiveness of information systems by determining their recall and precision ratios. He explained his ideas to Cummings and recommended Lancaster for the job (Miles, 1982). Saul Herner, another information science pioneer, concurred in the recommendation.

Cleverdon's experience with Lancaster came from their work together in England on projects using the Cranfield collection and evaluation techniques. Lancaster served as senior research assistant on the Cranfield Project from 1962–63 and published a summary of the Cranfield research in *American Documentation* (Lancaster & Mills, 1964). He also drew on significant prior practical experience in librarianship, classification, and indexing in conducting his evaluation research.

At the time of his recommendation to NLM, Lancaster was head of the Systems Evaluation Group at Herner & Company in Washington, DC, working on a project for the Technical Library at the U.S. Navy Bureau of Ships (Lancaster, 1964) and utilizing procedures similar to those used in the Cranfield studies and later used in the MEDLARS evaluation. The approach was described as follows: The purpose was

> to evaluate and maximize the effectiveness of a computerized information retrieval system based on a specialized thesaurus used in conjunction with the Engineers Joint Council (EJC) system of role indicators and links. . . . The evaluation method used was that developed by Cleverdon in the ASLIB Cranfield Project. . . . Retrieval effectiveness was expressed in terms of relevance and recall ratios. . . . Reasons for search failures were analyzed in terms of indexing faults, searching faults, and system faults. (Herner, Lancaster & Johanningsmeier, 1965, p.92)

The detailed failure analysis was important as "a basis for remedy and correction" (p. 95), also a key characteristic and important contribution of the MEDLARS evaluation. Lancaster's approach and attitude toward evaluation was also conveyed in the Bureau of Ships paper:

> Relevance and recall ratios cannot be construed as figures of merit; they do not tell us whether we have a good or bad system in any absolute sense. What they do tell us is what kind of system we have, and it is for us to decide whether what we have meets our needs. . . . No evaluation technique can tell us what we want or need. These we have to decide for ourselves. (p. 95)

This early articulation of Lancaster's evaluation viewpoint is revealing of the perspective he brought to bear not only on the MEDLARS evaluation, but throughout his career in other evaluation projects and in his influential books on the subject. Evaluations provide information for making decisions within a particular context and for measuring the effects of system or operational changes.

So upon the recommendation of Cleverdon and Herner, NLM Director Cummings engaged Lancaster in December 1965 to evaluate MED-

LARS. He appointed a committee of knowledgeable computer specialists, including Cyril Cleverdon and Calvin Mooers, to review the test procedures and results.

Evaluation Description

Planning of the evaluation began in December 1965, when Lancaster joined the NLM staff as Information Systems Evaluator. As a newcomer previously uninvolved in the design or operation of MEDLARS, he was able to approach the job with a spirit of impartial analysis that was maintained throughout (Lancaster, 1968a).

The one-year evaluation was launched in August 1966 and ran through July 1967. The Demand Search component of MEDLARS had been in place for nearly two years. The evaluation results were published in a 1968 report to the National Library of Medicine (Lancaster, 1968a), followed by two journal articles, one in *American Documentation* for the library and information science audience (Lancaster, 1969a), and the other in *JAMA* for the scientific and health professional user community (Lancaster, 1969b). The following description of the evaluation and its results is based primarily on these three published accounts authored by Lancaster.

The main objectives were to study the requirements of MEDLARS users, determine the effectiveness and efficiency of MEDLARS in meeting their requirements, identify factors adversely affecting performance, and suggest ways to make improvements. The evaluation was designed to provide information on MEDLARS performance relative to user requirements around several key factors of a retrieval system: coverage of the literature, recall power, precision power, response time, format of the results, and the user effort needed to achieve a satisfactory search result. The team "wanted to identify the principal causes of search failures, thus allowing corrective action to be taken to upgrade system performance" (Lancaster, 1969a, p. 120).

Lancaster summarized the evaluation as follows in the October 1966 issue of NLM's newsletter:

> In an effort to refine and improve MEDLARS services to the biomedical community, the Library has initiated a new project designed to provide data on the usefulness of demand bibliographies. This project is believed to represent the first extensive study of a large-scale operating information system. The evaluation is based on two measurements: "recall," or the proportion of useful citations in MEDLARS actually retrieved; and "precision," the ability to withhold citations to non-relevant documents. To measure "recall," it is necessary to compile a list of relevant documents by some means other than MEDLARS. This is done, first, by having the recipient of a demand search provide a list of citations already known to him; and, second, by conducting a manual search of the literature, using reference tools, such as Science Citation Index, not generated by NLM. The recipient assesses the citations identified by the manual search. Those which he finds relevant, plus

> the ones he originally suggested, constitute the recall base. "Precision" is measured by having the recipient of a demand search examine photocopies of journals articles selected at random from the search output to determine their relevancy. (MEDLARS Evaluation, 1966, p.3)

Although these concepts are familiar to us now, they were not at the time. In Lancaster's study report and the papers that followed, each element of the evaluation was carefully and clearly defined and explained—in and of itself a contribution to the knowledge of the field.

Methodology Challenges From a methodology perspective, two critical challenges involved sample selection and the method for determining recall and precision performance.

To achieve a representative sample of users and requests, Lancaster decided on a stratified sample of twenty organizations that had used the MEDLARS service in 1965. The organizations were selected to achieve balance among the following factors: volume of requests, likely subject area of requests, type of organization, and mode of user-system interaction in terms of the level of contact the requestors had with the librarians or search analysts who served as the interface to the system at that time. The users were the individual physicians and scientists affiliated with the organizations in the sample, plus some private practitioners not aligned with the organizational users.

To establish the recall and precision performance figures in the MEDLARS evaluation, Lancaster relied on a formal search request representing the user's actual information needs, followed by the user's assessment as to the relevance of documents to that need. Relevance judgments were made on a random sample of the retrieved citations, for which the full text of the articles was provided. Users rated the value of an article as major, minor, or of no value in addressing the information need that prompted the search, with reasons to support the rating. If not directly relevant, users were asked if it was relevant to some other project to gather information on the serendipity factor. This approach was important for obtaining valid precision figures as well as for obtaining data important to other analyses in the study.

Lancaster's views about relevance judgments were clear and strong:

> We believe categorically that, within the environment of an operating retrieval system, where the performance of the entire system is being evaluated, a "relevant" document is nothing more nor less than a document of some value to the user in relation to the information need that prompted his request. (Lancaster, 1969a, p.121)

The difficulties of estimating a recall ratio in a large-scale operational system were well known at the time. While feasible in certain experimental settings to identify the complete set of documents relevant to given requests, it was not so for a database containing hundreds of thousands of documents. Lancaster's approach to calculating recall by comparing

search results to a set of relevant documents identified completely outside the MEDLARS system was an innovative method for studying recall in an operational setting.

Lancaster is of course well known for his emphasis on the use of recall and precision as key performance measures in an information system, but his attitude toward the meaning and use of this information is equally important and instructive. He did not consider them as absolute indicators of the quality or success of an information retrieval system, writing that "recall and precision figures are merely yardsticks by which we measure the effect of making certain changes in our system or in ways of operating the system" (Lancaster, 1969a, p.122). Lancaster was also careful to point out that citing average recall and precision percentages for a system can be misleading, and that the detailed analysis of each failure is more important in providing information on specific changes to improve system effectiveness and efficiency.

Analysis of Results A detailed and thorough analysis of approximately three hundred specific search failures revealed by the precision and recall calculations was carried out, resulting in many recommendations for change. In describing this aspect of the evaluation, Lancaster wrote, "The 'hindsight' analysis of a search failure is the most challenging aspect of the evaluation process" (Lancaster, 1969a, p.123).

He went on to explain what the analysis entailed, which I quote in full to emphasize the detail with which the process is described as well as the enormous effort that went into this crucial aspect of the study.

> It involves, for each "failure," an examination of the full text of the document; the indexing record for this document (i.e., the indexing terms assigned, which are obtained by printout from the magnetic tape record); the request statement; the search formulation upon which the search was conducted; the requestor's completed assessment forms, particularly the reasons for articles being judged "of no value"; and any other information supplied by the requestor. On the basis of all these records, a decision is made as to the prime cause or causes of the particular failure under review. (Lancaster, 1969a, p. 123)

For each failure identified, a specific system recommendation for the affected area of service was put forward.

This thoroughness of failure analysis particularly distinguishes the corpus of Lancaster's work in the evaluation arena. Throughout the MEDLARS analysis, there was a careful distinguishing among the different types of errors, and careful explanations of the distinctions. At every turn, detailed analyses and mini-studies were being done to further investigate the findings, to consider possible changes, and illustrate the effect on retrieval of implementing the recommended change.

Evaluation Findings The failure analysis identified the principle system components responsible for recall and precision failures—indexing,

searching, index language, or user-system interface. The results have had a lasting impact on how we view and approach the evaluation of information retrieval system functions. Key findings for each subsystem are summarized below. Other factors considered in the evaluation of system performance included journal coverage, foreign language literature, journal usage factors, system response time, serendipity value of searches, and quality of output screening.

Within the Indexing Subsystem, Lancaster identified two types of failure—indexer error and policy related to exhaustivity of indexing.

Indexer error was of two types, omission of needed terms and use of an inappropriate term. Errors of omission were more common and partly attributed to an inadequate entry vocabulary in MeSH, leading to a recommendation to augment the entry vocabulary.

Exhaustivity of indexing, or depth of indexing, refers to the number of index terms assigned. Failure related to exhaustivity of indexing had to do with two policies. The first policy specified that certain journals be indexed at a nondepth level, receiving about three terms per article instead of ten terms per article. Lancaster found that exhaustive indexing is better in general, and he recommended that decisions be made on the basis of individual articles rather than an entire journal. The second policy specified that nondepth journals also be indexed at a general level, rather than the most specific level allowed by the MeSH vocabulary, and led to both recall and precision problems. Lancaster recommended against indexing any article at the general level. Commenting on their irretrievability and demonstrating his flair for words, he wrote that such a policy was "indefensible" in an environment of machine retrieval and that articles "indexed in such general terms are merely occupying space on the citation file" (Lancaster, 1969a, p.131).

The Index Language Subsystem failures were of two types: those due to lack of specificity in the available terms, and those due to ambiguous or spurious relationships between terms. The analysis revealed overall index language deficiencies and identified specific subject areas in which the vocabulary was weak. Recommendations included augmentation of the entry vocabulary and addition of specific terms or term combinations to explicitly cover needed topics.

Lancaster's strong interest in vocabulary control was clear in his characterization of the Index Language role in an information retrieval system:

> The quality of the index language is probably the most important single factor governing the performance of a retrieval system. Poor searching strategies and inadequate or inconsistent indexing can mar the performance of a system, but indexing and searching, however good, cannot compensate for an inadequate index language. (Lancaster, 1968a, p. 80)

Considering the depth of analysis surrounding vocabulary issues in the MEDLARS study, and in the prior studies with the Bureau of Ships and the Cranfield Project, it is not surprising that Lancaster soon became an expert in vocabulary control in information retrieval systems. His book *Vocabulary Control for Information Retrieval* (1972c, 1986) still stands as an important classic text.

The Searching Subsystem was found to be the greatest contributor to MEDLARS failures. Four types of searching errors were identified: omission of topics from the search strategy, use of inappropriate terms, defective search logic, and inappropriate levels of specificity or exhaustivity in the search strategy. With respect to this last type of error, Lancaster (1969a) commented, "In fact, the central problem of searching is the decision as to the most appropriate level of specificity and exhaustivity to adopt for a particular request" (p.132). Additional training was one of the recommendations for addressing these problems.

Failures due to User-System Interaction were of two types: inadequate capturing of the information requirement in the search request, and inadequate ability to interact with the system. The mode of interaction with the system affected search success in a way that was not expected at the outset of the evaluation. Search request forms that were filled out directly in the requestor's own natural language resulted in better search results than those in which information specialists interpreted the information need and completed the search request form as a third party. Recommendations regarding improvements in the search request form and the system interface resulted from this aspect of the evaluation.

Evaluation Conclusions The conclusions of the study were based on a careful examination of searches at each end of the recall-precision distribution. Lancaster recommended actions in several areas to improve overall operating efficiency. Some of the more important are: improve user request statements; record recall and precision tolerances of the user; establish standard strategies for recurring search elements; abandon the distinction between "depth" and "non-depth" indexing; obtain input from the indexing and searching operations to further develop the MEDLARS vocabulary; expand the entry vocabulary and make it readily available to every indexer and searcher; extend the use of subheadings; and develop greater integration between the activities of indexing, searching, and vocabulary control.

Lancaster also recommended that NLM begin continuous quality control of MEDLARS searches to ensure good performance and collect necessary data for continued improvement. Emphasizing the ongoing nature of evaluation, he wrote, "Only by continuous self-appraisal can a large information system make itself responsive to the needs of the scientific community" (Lancaster, 1969a, p. 142).

Impact of the Evaluation at NLM

At the final meeting of the Evaluation Advisory Committee in January 1968, NLM Director Cummings expressed his praise for Lancaster and his hope that the committee fully endorse the evaluation findings. In his speaking notes, he wrote: "I have carefully read [the] report of [the] study and its findings. I share your view that it is a job well done! Thanks to Wilf Lancaster" and "You have my pledge that NLM management will carefully receive and review your recommendations with a view towards improving the system and distribute our findings for the benefit of others" (Cummings, 1968).

The evaluation was useful both inside and outside the library. Within the library, the results were taken quite seriously and resulted in numerous changes in policies, procedures, and content. Cummings ordered that Lancaster's recommendations be adopted, and he set up a quality control unit to check the effectiveness of every search requested by a patron (Miles, 1982). He also refined MEDLARS indexing, expanded MeSH terminology and hierarchical tree structures, and established additional MEDLARS training programs. Later the designers of MEDLARS II for online searching built on the results of this evaluation.

Quality Control Program at NLM The quality control program established in response to the MEDLARS evaluation was described by Jenkins (1972). In this paper, Jenkins (later McCarn, former chief of Bibliographic Services Division at NLM) directly links the program to the evaluation, writing,

> As a result of Lancaster's recommendation, the National Library of Medicine, in March 1968, established a Quality Control Unit as part of the Bibliographic Services Division (BSD). A small staff began to plan and implement the program, and the first steps have been taken toward setting up a program to evaluate MEDLARS on a continuous basis. (p. 423)

The first project was to revise the Search Request Form and Search Appraisal Form along the lines that Lancaster had recommended. In addition, detailed failure analysis of selected searches, taking approximately eight hours per analysis, was conducted. Jenkins' discussion of the time-consuming process of failure analysis sheds additional insight into the painstaking and thorough nature of Lancaster's work on the MEDLARS evaluation. Jenkins also outlined future plans for implementing a comprehensive quality control program, including continuous interaction with users to obtain feedback on system limitations, per Lancaster's recommendation. A program of integrated quality assurance for vocabulary, indexing, literature coverage, and user support services continues at NLM today.

MeSH Enhancement Internal NLM correspondence illustrates the impact on MeSH even before the evaluation was completed. In a January

1967 memo to the NLM director, the associate director for Intramural Programs wrote the following regarding early case reports from the MEDLARS evaluation:

> Mr. Lancaster is beginning to provide feedback analyses of search evaluations. This is a very useful procedure and we anticipate that these will provide useful information to Search, Indexing, and MeSH. I am requesting Mr. Lancaster to provide broader circulation of these analyses to our staff. (Leiter, 1967)

This was followed by a memo from Lancaster (1967a) to the head of MeSH, saying:

> We have begun preliminary analysis of test searches from the MEDLARS evaluation program. I hope it may be of some value if I report to you on any searches in which terminological problems have been primarily responsible for poor system performance.

Two topics that were then brought to his attention were "separation anxiety" and the subheading "complications." Terminology problems regarding drug information were directed to Winifred Sewell in the Drug Information Program; the first example dealt with a search on "toxicity of organic selenium compounds" (Lancaster, 1967b).

Lancaster continued to work with NLM on vocabulary control, looking ahead to plans for the online environment of MEDLARS II. Harley and Lancaster (1969) analyzed the dynamics of a large controlled vocabulary for online implementation, and coined the term "lexicodynamics" to express the concept of construction, maintenance, use, and change of controlled vocabularies for information retrieval purposes.

Training Programs Following completion of the MEDLARS evaluation, Lancaster was named deputy chief of the Bibliographic Services Division in February 1968, then special assistant to the associate director for Library Operations in September 1968. In those roles, he developed training programs and training materials at NLM, consistent with his own recommendations. The *NLM News* announced a series of five seminars on information retrieval systems in May 1968, the first of which was presented by Lancaster (New MEDLARS Training, 1968). Later that year, Lancaster announced user training programs for MEDLARS Centers and Regional Libraries (MEDLARS Search Analysts, 1968) and issued a community call for samples of teaching materials (Share, 1968). In December, Lancaster's involvement in an all-day orientation program on the use of MEDLARS was credited in this way: "This curriculum is being further developed by Mr. F. Wilfrid Lancaster, Special Assistant to the Associate Director for Library Operations, as a base for a much broader user orientation program to be established on a national scale" (MEDLARS Orientation, 1968). He also wrote a seventy-seven-page, illustrated booklet called *Principles of MEDLARS* (Lancaster, 1970). Designed for MEDLARS users, it covered

indexing procedures, vocabulary, and search strategies, and sold for seventy-five cents (to be included with the order).

Since those early years, NLM has continued a strong training initiative, developing a national program to support training of librarians, health professionals, biomedical researchers, and the general public.

Impact Outside the Library

Outside the library, Lancaster's evaluation was of interest as the first large-scale evaluation of a major operating information system and was met with favorable reviews. The paper published in *American Documentation* (Lancaster, 1969a) received its Best Paper award for 1969.

In a memo to the Director of the National Institutes of Health, Dr. Cummings (1969) wrote:

> The critical evaluation of MEDLARS searches conducted by Mr. F.W. Lancaster was a landmark report on the evaluation of large operating information systems. Mr. Lancaster has forwarded copies of reviews of his work, which I should like to share with you. I hope you will find them as interesting as I did!

In a review for *Library Association Record*, Brian Armitage of Charing Cross Hospital Medical School (1969) praised it as "an impressive piece of work." In a lengthy and thoughtful review for the *Journal of Documentation*, Glyn Evans (1968) of the Royal Society of Medicine (later of Washington University School of Medicine) addressed not only the evaluation per se, but also the underlying measurement and methodology issues. He wrote:

> We are in debt to NLM not merely because of this report, important though it may be, but because it again demonstrates the responsiveness of NLM to a need, it's recognition that the system efficiency must be measured and monitored. . . . We are now looking to MEDLARS II and, on the basis of this report, we should be nothing but optimistic.

In a set of formal comments delivered to NLM, Cyril Cleverdon (1968a) had this praise for Lancaster:

> In the first place, the evaluation has been carried out by Mr. Lancaster in a manner that is beyond praise. His application and integrity of purpose have been outstanding, and even the most casual reading of the final report must indicate the amazing amount of work he has done. (p.10)

Lancaster's work was cited favorably in numerous mentions in the *Annual Review of Information Science and Technology (ARIST)*, some of which are included here. In 1970, Lancaster himself was invited to write the chapter on evaluation, an honor earned by the excellence of his work at NLM and his reputation in the field. Of the MEDLARS evaluation, Lancaster (Lancaster & Gillespie, 1970) discussed it as the first major evaluation of a large national information system and noted that Cleverdon (1968b) had already utilized its detailed examples for a paper on procedures for evalu-

ating a retrieval system at various stages of development. Lancaster closed his review with a call for more evaluations of actual operating systems: "It is in the conduct of evaluation programs applied to working systems that most effort, we feel, needs to be applied in the future" (p. 63).

Cleverdon's (1971) review in the ARIST chapter on Design and Evaluation of Information Systems cited the utility of data obtained by Lancaster for further studies of subject retrieval. In the chapter on Information Science Applications in Medicine (Caceres, Weihrer & Pulliam, 1971), the authors write: "These evaluations of MEDLARS as a model will be most helpful in the implementation of future systems" (p.332).

The MEDLARS evaluation was important at one level because it was the first comprehensive evaluation of the first large-scale operational information storage and retrieval system—two very powerful firsts. But surely the clarity and completeness of Lancaster's published reports of the study, including the detailed reasoning behind certain design decisions and system recommendations, were responsible for the widespread reference to the evaluation as a landmark study and for the general acknowledgment of the high quality of his research as a systems evaluator.

Evaluation of On-Line Searching in MEDLARS (AIM-TWX) by Biomedical Practitioners

Lancaster conducted a second evaluation of NLM's MEDLARS retrieval system in 1970–71, this time for its innovation as one of the earliest online services. AIM-TWX was an experimental service developed as an exploration of future online capabilities. Lancaster's report was published in 1972 and entitled "Evaluation of On-Line Searching in MEDLARS (AIM-TWX) by Biomedical Practitioners" (Lancaster, 1972b).

In this section, I begin with a discussion of the general information environment, then describe the study, its results, and its impact at NLM and elsewhere.

The Information Environment

One disadvantage of the MEDLARS Demand Search Service was its slowness. The time was usually three to six weeks from the submission of the request, through search strategy formulation by an NLM analyst, processing in the computer, and mailing the bibliography to the patron (Miles, 1982). To take advantage of new computer developments for online retrieval systems, the NLM embarked on the development of MEDLARS II. An early result was a practical online bibliographic system named AIM-TWX. This provided access to the Abridged *Index Medicus* (AIM) database using the Teletypewriter Exchange Network (TWX) as the communication system. The AIM-TWX database contained approximately 100,000 citations, comprised primarily of articles published in the prior five years in one hundred English-language journals in clinical medicine. One innovation, in addition to the online access mode, was that users could either

enter search terms directly or they could search the MeSH vocabulary to locate appropriate search terms.

When AIM-TWX was opened as an experimental system to a select group of users across the country in June 1970, it became the first national medical information service available from any teletypewriter or TWX terminal in the country, for the price of a telephone call (Dee, 2007). The goal was to determine the need for and usefulness of such services.

Conducting the Evaluation

The purpose of the investigation was "to determine how effectively biomedical practitioners, with a minimum of introduction to the system, can conduct on-line searches to satisfy their own information needs" (Lancaster, 1972b). This was another of Lancaster's highly important studies, not only because it involved an innovative retrieval system, but also because it was looking at the end user's direct experience with the system. Study results were also important for the further development of the MEDLARS II system, from which MEDLINE was introduced in October 1971 as NLM's full online retrieval system.

Method Searches used in the study were conducted by biomedical practitioners at four MEDLARS centers during the three-month period from November 1970 through February 1971. Searchers were provided with a brief standardized description of how to use the system.

The users conducted searches on their own, although the analysts were on hand to answer questions related to purely technical problems, not those related to searching. The trained search analysts also structured and conducted parallel searches on the same subject, for comparison purposes. Relevance judgments were obtained for both searches, and recall and precision measures were calculated. An online questionnaire captured data about the role of the searcher, prior searching experience, the purpose of the search, and the value of results. Unit costs were also calculated, another innovation and particular contribution of this study. Results between the practitioners and search analysts were compared. Characteristics of "worst" and "best" searches were also compared through a thorough detailed analysis, as in the earlier MEDLARS Demand Search Service evaluation.

Results Over the three-month study period, forty-eight test searches were completed. Precision was calculated on forty-five of the forty-eight searches, with an average precision measure of 63 percent. Lancaster considered it quite encouraging that a group of end users with minimal exposure to the system should be able to achieve precision of greater than 60 percent. Recall was calculated on thirty-six of the forty-eight searches, with an average recall measure of 57 percent. Lancaster noted that this was about the same as the recall results for the earlier MEDLARS study of the batch system, and that it appeared entirely satisfactory for the users,

half of whom indicated they were looking for a few citations only. Unit costs were calculated for thirty-nine searches, obtained by dividing the total time at the terminal by the number of relevant citations retrieved. The average unit cost was 3.4 minutes, which was deemed reasonable considering that most searches were somewhat complex, requiring coordination between two or more aspects. Assessments of the value of citations to the user were also obtained for thirty-nine of the forty-eight searches, with 67 percent being rated of major or considerably high value.

In comparing characteristics of the best and worst searches, Lancaster found that the best searches, those with high recall and high precision, had two features in common: they involved relatively simple relationships, and the terms from the search request statement map fairly directly into MeSH headings. The worst searches were of two types: requests for which little actually existed in the database; or requests involving more sophisticated search techniques due to complex conceptual relationships or because the appropriate MeSH term was not obvious.

Lancaster made the following observations about use of the service by biomedical practitioners: searches are effective when the conceptual relationships are not complex; users are successful in using relatively simple approaches; searches are effective when the MeSH terms match the user's request terms closely; lack of entry vocabulary is a problem; users' failure to recognize all possible approaches to retrieval is a problem; interactive features of system are little used; and few users choose the "print full" option.

Conclusions Following the study, Lancaster concluded that many biomedical practitioners could exploit AIM-TWX profitably with minimal introduction to the system and without using a trained MEDLARS analyst. He also concluded that AIM-TWX met a definite need, noting that most of the searches could not have been conducted in *Index Medicus* due to the required combination of conceptual relationships.

On the whole, he found the results to be "surprisingly good" (Lancaster, 1972b, p. 11). Precision rates were high, recall rates were comparable to trained analysts searching the offline MEDLARS service, and the cost in time was reasonable. Although the users did not perform quite as well as the trained analysts to whom they were compared, they were not expected to do so. Lancaster suggested they would probably not use the printed *Index Medicus* as effectively as trained professionals either. To facilitate use of the system and improve performance, Lancaster suggested providing a brief, clear, and well-illustrated booklet that describes how to use the system. He envisioned a booklet that would present the essentials of the system and not attempt to cover every feature and command option.

Recommended improvements included the following: making it more forgiving of simple typographical errors; removing duplicate records in a search session; and allowing the use of entry terms rather that MeSH

terms only. Lancaster (1972b) believed that many problems would be solved by "a well constructed network of cross-references and an adequate entry vocabulary" to facilitate more sophisticated search strategies.

Lancaster also offered some general conclusions about online searching of MEDLARS, based on his analysis of searches. Reflecting on the differences between the historical user group of trained search analysts and the new group of end user searchers, he offered ideas of what a future online system should offer to best support direct use by the biomedical practitioner. He recommended that we "strive to produce improved systems that are more user-oriented and that will help the user to attain higher levels of success" (Lancaster, 1972b, p. 14). Suggested approaches to doing this included generating spontaneous displays of related vocabulary terms, providing tallies to show a term's usage in indexing, providing the option of viewing a term's definition and permissible subheadings, and providing information on the frequency of particular MeSH term/subheading pairs. Additional suggestions addressed facilitating use of logical operators by designing an entry screen that resembles a simple form with logical operators already displayed and in place.

Looking further to the future, as he is always so good at doing, Lancaster envisioned that, ultimately, end user systems should allow input of natural language search requests and avoid the necessity for use of Boolean operators. He also stressed that an extensive entry vocabulary would be needed to allow the necessary mapping from natural language to the controlled vocabulary terms used for indexing the documents. He envisioned Boolean logic being replaced by systems based on term-weighting and ranking algorithms, writing that "Boolean search equations are unnecessary and are probably undesirable in mechanized retrieval systems" (Lancaster, 1972b, p. 17).

The evaluation report closed with further thoughts on the directions Lancaster believed online systems should go in the long run. Written more than thirty-five years ago, his words reflect his signature forward-thinking attitude toward system design and are as true today as they were at the time:

> We should always look for ways of improving retrieval systems and making them more attractive to potential users. The philosophy that "the system is used, therefore it is good" is a very shallow one. We must not assume that a system having appeal today will always retain this appeal. There is a certain novelty factor about AIM-TWX that is at least partly responsible for the very favorable acceptance it has in most quarters. But novelty wears off and system designers cannot afford to rest too long on their laurels. In the past, users have been required to adapt to the information system. In the future systems must be designed that adapt to the users. (Lancaster, 1972b, p. 18)

Impact of AIM-TWX Study

Guided by experience with the test, NLM went on to plan an online system that would accommodate ten times as many searches as MEDLARS each year at one-tenth the cost. This new service, named MEDLINE (for MEDLARS onLINE), began trial runs in the library in October 1971 and was opened to a selected group of institutions in December 1971. Patrons immediately turned to it, as they could obtain lists of citations within minutes. NLM discontinued the MEDLARS Demand Search Service in January 1973 (Miles, 1982).

The visibility and importance of the AIM-TWX study in the field is clear from numerous citations in ARIST, some examples of which are included here. In chapters on "The User Interface in Interactive Systems," reviewers highlighted the value of the AIM-TWX study to interface designers, citing Lancaster's numerous suggestions for added features beneficial to end users (Bennett, 1972) and his contributions to the understanding of how to improve user system interfaces (Martin, 1973). In a chapter on "Document Description and Representation," Batten (1973) notes the success of inexperienced searchers in the AIM-TWX study as an indicator that the heuristic search capability of online systems may allow for item representations that are less intensively descriptive than batch systems. In a chapter on "Economics of Information," Michael Cooper (1973) noted the contribution of the AIM-TWX study to developing a measure of search cost. Martin (1973) also commented on the cost aspect, writing "In the AIM-TWX study, Lancaster [has] made it possible for future researchers to address the question of costs directly, assuming that retrieval is satisfactory" (p. 212).

Other Evaluation Research

Lancaster's record of achievement in the evaluation of information storage and retrieval systems extends beyond the MEDLARS evaluations, of course. He was also sought after as a consultant, advisor, evaluator, and designer, and also continued to conduct research and write in the areas of inquiry of interest to him. Throughout the 1970s and 1980s, he was invited to give numerous lectures, seminars, and workshops throughout the world on topics including evaluation of national information systems, indexing and abstracting, thesaurus construction, information retrieval techniques, and evaluation criteria and methods. Lancaster also wrote four evaluation-related articles for the *Encyclopedia of Library and Information Science:* "Evaluation and Testing of Information Retrieval Systems" (1972a), "On-Line Information Systems" (1977c), "Pertinence and Relevance" (1977d), and "Precision and Recall" (1978). Included here are some examples of his work on evaluation of bibliographic retrieval systems.

Evaluations of Information System Use and Function

Soon after the completion of the AIM-TWX study Lancaster used the same evaluation framework for an evaluation of the online Epilepsy Abstracts Retrieval System (EARS) the National Institute of Neurological Diseases and Stroke (NINDS) (Lancaster, 1971a; Lancaster, Rapport, & Penry, 1972). EARS contained approximately 8,000 abstracts and allowed free text searching. Neurology specialists at six U.S. medical centers conducted their own online searches, and parallel searches on the same topic were conducted by experienced neurologists at NINDS. The results of forty-seven searches were evaluated in terms of recall, precision, and general user satisfaction, and compared against the results of the experienced searcher. A detailed analysis of factors affecting the success and failure was also conducted.

As with AIM-TWX, Lancaster concluded that reasonably successful searching could be done by inexperienced searchers. Martin (1973) highlighted the EARS study in his ARIST review of the user interface literature, writing:

> Recall and precision failures were attributed to the fact that users did not cover all approaches when they formulated requests. Since free-text searching is inherently difficult, improved instruction and the user of online searching aids (e.g., a thesaurus or synonym groups) would have improved performance. (p. 211)

In another review for ARIST, Bennett (1972) noted the detailed analysis and examples that are hallmarks of a Lancaster evaluation, writing, "Ten pages of examples provide much of value for the designer responsive to the challenge of the redevelopment cycle" (p. 183).

In 1974, Lancaster (1974) evaluated the applicability of an online bibliographic search system to the National Instructional Materials Information System (NIMIS). In 1976 and 1977, he was a member of an international study team appointed by UNESCO to assess the impact of the AGRIS international information program, including the AGRINDEX database, on the worldwide dissemination and availability of agricultural information (Badran, et. al., 1977; Lancaster & Martyn, 1978). Lancaster (1977a) went on to develop a set of guidelines for UNESCO on the evaluation of information systems and services. In 1977, he did an evaluation of the French PASCAL bibliographic retrieval system for the Centre national de la recherche scientifique (CNRS) (Lancaster, 1977e). Throughout the 1980s, he was involved in the evaluation for UNESCO of the United Nations Environment Programme's information program. In 1994, he evaluated the searching of databases on CD-ROM by end users (Lancaster, et al., 1994). In the 1990s, Lancaster worked again with the research and development staff of the National Library of Medicine, this time on a design for the evaluation of MedIndEx, a prototype expert system for medical indexing (Lancaster et al, 1996).

Lancaster was frequently invited to contribute papers to edited books, including those dealing with evaluation. He contributed a piece on evaluation in the environment of an operating information service for *Information Retrieval Experiment* by Karen Spärck Jones (Lancaster, 1981). He contributed a paper on some limitations of methods for evaluation of information services for the FID publication *Theoretical Problems in Informatics* (Lancaster & Rapp, 1981). In a paper for *Perspectives in Information Management*, he reexamined issues surrounding natural language versus controlled vocabularies in searching (Lancaster, 1989).

Lancaster also offered an early view of the future of the indexing and abstracting systems that he evaluates. Writing in 1982 before the first electronic journals, he outlined the possible steps in an evolution from a predominantly paper-based publishing environment to one that is predominantly electronics-based, with the disappearance of the printed journal and the secondary databases as we know them (Lancaster & Neway, 1982). In this paper, Lancaster also foresaw the current movement toward interactive publications, writing:

> The most important point to be made is that the entire character of primary publications is likely to change rather drastically and that electronic capabilities will have a radical effect on the way that information is presented, perhaps leading to a situation in which much narrative text is replaced by alternative modes of presentation and publications become "interactive," the user being able to manipulate and interact with the data presented. In other words, future electronic publications may look less like present publications than like the more sophisticated programs now existing within systems for computer-aided instruction or, to use a more extreme analogy, like the electronic game. (p. 187)

Cost Studies

Lancaster is also well known for his research on cost-effectiveness and cost-benefit analysis as they relate to information storage and retrieval systems. Following the early evaluations of operational systems such as MEDLARS, "it became obvious that evaluation must be more directed toward operational decisions" (King, 1978, p.2).

Lancaster and Climenson (1968) followed up on the MEDLARS evaluation with an analysis of the economic efficiency of the system. They distinguished between evaluating only user satisfaction, which addresses operating efficiency, versus evaluating the efficiency of the means to satisfy user requirements, which addresses economic efficiency. The trade-offs between operating efficiency and economic efficiency in determining the most economical path to follow are described, including pay-off factors, break-even points, and diminishing returns. The paper considers these factors in relation to key retrieval systems components: the acquisition subsystem, the indexing subsystem, the index language, the searching subsystem, and the equipment subsystem.

King and Lancaster (1969) developed a conceptual framework for the cost/performance/benefits approach to evaluation. In his review for ARIST, Lancaster (1970) summarized it as follows:

> Cost refers to input of resources to a system; performance relates to attributes directly controlled by the system, such as recall, precision, and speed of response; benefits are the consequences of system performance in terms of value, return on investment, effect on the behavior of the user, effect on other systems, and non-quantifiable consequences such as interactions with other systems. (p. 62)

Lancaster (1971b) published an important paper on cost-effectiveness analysis in JASIS, in which he emphasized the distinction between cost effectiveness and cost-benefit analysis of information systems. He also described approaches to doing cost-effectiveness analysis for various system components, including coverage, indexing, index language, search process, and hardware. Donald King selected the cost-effectiveness paper for inclusion in his 1978 compilation entitled *Key Papers in the Design and Evaluation of Information Systems* (King, 1978). King described it as a "classic paper" that "bridges the early methods and criteria of evaluation and newer approaches" (p.10). He summarized its content as follows:

> Lancaster takes some complex mathematical evaluation models developed at Westat, Inc., and elsewhere, and describes these concepts in simple terms. The paper describes indexing and search system in terms of the cost and effectiveness of functions performed by the system such as acquisition and storage, identification and location, and presentation. Factors are listed which relate performance of the functions to costs and benefits. (p. 10)

In an ARIST chapter on "Costs, Budgeting, and Economics of Information Processing," Wilson (1972) wrote, "The design of storage and retrieval systems has reached a state where Lancaster can summarize in a check-off list the items required for cost-effectiveness analysis" (p. 43). He also noted that Lancaster sets a "modest goal" for cost-effectiveness analysis—"to serve as a useful tool in the decision-making process" (p. 43).

Books

Lancaster is of course very well known for his excellent books on information retrieval systems, all of which have been favorably reviewed and some of which have received awards within the profession. Most of the books were intended primarily as texts for use in schools of library and information science, but they were always of interest to a much wider audience. While the topics span the field of information retrieval, most include content on evaluation principles and methods even when evaluation is not the main focus of the book. All share the same secrets of success—clarity, relevance, balance, and a practical approach within a theoretical framework.

In Lancaster's first book, *Information Retrieval Systems: Characteristics, Testing, and Evaluation* (Lancaster, 1968b), he is praised for the clarity of writing. Saul Herner writes in his foreword to the book,

> All too rarely in this complex field of information science . . . are the practitioners able to make themselves clear . . . Mr. Lancaster is a welcome exception . . . He has furnished us with perhaps the most complete and authoritative statement extant about where we are in this rapidly evolving field, how we got there, and . . . what directions the field is likely to take in the future.

The American Society for Information Science recognized its significant contribution by awarding it Best Information Science Book in 1970.

The book was written for students of library and information science as well as practitioners concerned with system design, operation, and evaluation. In the preface, Lancaster describes the book as being concerned primarily with the intellectual factors that affect the performance of all retrieval systems: indexing, vocabulary control, search strategy, and user-system interaction. Writing from the viewpoint of evaluator, he emphasizes measurement of system performance against satisfaction of user requirements. The content of the book is drawn heavily from Lancaster's work on the ASLIB Cranfield Project and his association with Cyril Cleverdon. Examples from the MEDLARS study are also provided, particularly in the sections dealing with controlled vocabulary, indexing, user-system interaction, and evaluation of operating efficiency.

To give an idea of the state-of-the-art in information systems at that time, I quote one of Lancaster's comments on the feasibility of automated searching due to fast processing speeds:

> For example, using a Honeywell 800 Computer, and associated peripheral devices, the MEDLARS system at the National Library of Medicine can compare a batch of 40 highly complex search formulations against a file of 700,000 document descriptions, producing, for each search, a printout of citations of all items satisfying the search logic, in about eight hours of processing time. (Lancaster, 1968a, p. 47)

The second edition, published eleven years later in 1979, was expanded in scope to be even more suitable as an introductory text book. It includes significantly more content related to evaluation, and the MEDLARS system is described in detail, comprising much of an entire chapter in the book.

In *Measurement and Evaluation of Library Services* (Lancaster, 1977b), Lancaster broadened the scope of his evaluation texts beyond the automated information storage and retrieval systems that were used by libraries, and addressed the functions of the library itself. A single chapter addresses information retrieval and literature searching, while the rest of the book covers topics such as catalog use, reference service, the collection, document delivery, technical services, and library automation. The book covers a wide range of evaluation methods for assessing how well the

library satisfies the needs of its users. In his foreword to the book, Herb Goldhor wrote that "Professor Lancaster is one of the best-qualified and competent people to write this book." He also predicted that it would quickly become a standard reference in the profession—which it did.

Another classic, *Vocabulary Control for Information Retrieval* was published in 1972 (Lancaster, 1972c), with a 2nd edition in 1986 (Lancaster, 1986). Lancaster's introduction describes the book as dealing with

> the properties of vocabularies for indexing and searching document collections: the construction, organization, display, and maintenance of these vocabularies; and the vocabulary as a factor affecting the performance of retrieval systems. (p. vii)

The MEDLARS system, its MeSH vocabulary, and results of the MEDLARS evaluation are used throughout the book to illustrate various principals and practices.

Information Retrieval On-line (Lancaster & Fayen, 1973) was named ASIS Best Information Science Book in 1974. Bourne and Hahn (2003) describe it as a "major milestone in the literature of online systems" that "functioned for years as a textbook, handbook, and encyclopedia on all aspects of online retrieval systems" (p. 2). The section on performance evaluation listed six criteria for assessing the performance of information retrieval systems, which are now quite familiar: coverage, recall, precision, response time, user effort, and form of output.

The first edition of *Indexing and Abstracting in Theory and Practice* (Lancaster, 1991) received the Best Information Science Book award for 1992 from the American Society for Information Science. Subsequent editions were published in 1998 and 2003 (Lancaster, 1998; Lancaster, 2003). The book focuses primarily on principles of indexing and abstracting, but evaluation concepts are addressed throughout in individual chapters, particularly those on consistency, quality, and text searching. There is also one full chapter specifically devoted to evaluation aspects, in which Lancaster addresses the role of indexing and abstracting in four principal criteria for evaluating bibliographic databases—coverage, retrievability, predictability, and timeliness.

In the third edition's chapter on quality of indexing, Lancaster (2003) refers to some of his later work for NLM, writing

> In a study performed for the National Library of Medicine, I developed a method of evaluating the quality of indexing for MEDLINE by comparing the work of indexers against a "standard," this being a set of terms agreed upon by highly experienced indexers. (p. 95–96)

The chapter on text searching discusses natural language versus controlled vocabulary in detail, including evaluation studies to determine the relative merits of each. This chapter also describes the use of a post-controlled vocabulary, in which the system's controlled vocabulary is used

as a search aid, but is not actually used to index documents. Lancaster has written about the promise of postcontrolled vocabularies in natural language systems in both editions of *Vocabulary Control for Information Retrieval* (1972c, 1986) and in a paper on natural language retrieval (Lancaster, Rapport, and Penry, 1972).

Information Retrieval Today (Lancaster & Warner, 1993) expanded and updated the content of *Information Retrieval Systems: Characteristics, Testing, and Evaluation*, including material on automatic indexing, CDROM databases, linguistics, semantics, hypertext, expert systems, and developments in evaluation and quality control of information retrieval systems.

Continued Benefit and Impact at NLM

Re-Cap of Tenure at NLM

Lancaster's CV modestly and straightforwardly lists his employment at NLM as information systems specialist from 1965–68. But this vastly understates his roles while employed at NLM, and of course cannot represent the strength of his continued relationship with NLM after he left its employ. In the three years at NLM, he served as Information Systems Evaluator first in the Information Systems Division, then in the Research and Development Program. Following completion of the MEDLARS evaluation, he was promoted to deputy chief of the Bibliographic Services Division, then further promoted to special assistant to the associate director for Library Operations.

The NLM continued to benefit from Lancaster's insights and evaluation expertise following the early evaluations. He returned to give seminars, teach courses, serve as consultant, and conduct evaluation studies. More generally, he continued to heighten awareness and understanding of NLM services through his writing; through his inspiration of students' interest in systems evaluation and development; and through his encouragement of students to join the NLM through its associate fellowship program, many of whom went on to assume important leadership roles at NLM and in the field of medical librarianship.

As important to many, he continued his relationship as professional colleague and friend.

Remembrances from NLM

Wilf continues to be held in high regard by current and former NLM staff, who offer some remembrances on this occasion honoring his work.

Grace Smoley (formerly Jenkins, McCarn) is a former Chief of Bibliographic Services Division at NLM and now retired. She writes:

> I recall Wilf Lancaster with great admiration for his professional expertise and wonderful personal qualities. He had an amazing ability to speak, teach, and write in a straight, non-jargon way that was a delight. His landmark evaluations are a testament to his organizational ability

> and creativity. On a personal level, I remember Wilf as a kind and generous person who went out of his way to support his staff, co-workers, and students. My late (deceased) husband Davis McCarn always spoke of Wilf with the highest regard also. Wilf made a real difference to the library and information world in helping to get the whole online searching and indexing of literature into the mainstream. NLM and the library world would not have been the same without him.

Becky Lyon is Deputy Associate Director for Library Operations. She writes:

> Although I wasn't at NLM when Wilf conducted the MEDLARS evaluation, I recall learning all about it from Wilf as a library school student at the University of Illinois. I was one of 10 students that year who received a fellowship under an NLM grant to attend the U of I library school to study biomedical librarianship. Wilf was the project director and he carefully mentored all 10 of us, guiding us in appropriate career directions. I was encouraged by him to apply for the NLM Associate Program following graduation and was selected for the 1972–73 program. Throughout my years at NLM and in other libraries, I have always appreciated that Wilf steered me to NLM and the care that Wilf took in sending his best and brightest to our Associate program.

Sheldon Kotzin, Associate Director for Library Operations, did not work directly with Wilf at NLM, but recalls that the MEDLARS evaluation was considered of great importance at the time and was taken seriously by senior management at NLM.

Dan Tonkery, former Chief of the Technical Services Division at NLM and current Vice President of Business Development at Ebsco, Inc., writes:

> My first encounter with Wilf Lancaster occurred as a student in library school at the University of Illinois in 1969, where I frequently cited his research in a number of my papers. My first personal encounter came in the spring of 1970, when I had been elected to be President of the Student Group in Library School and was involved in the hiring of new faculty. Wilf had applied for a position in the library school and went through the interview process where I had an opportunity to meet him. I was also able to take an active part in his hiring through participation in the discussions and the vote.
>
> During the summer I had the privilege of taking two courses from him, Systems Analysis and Design and Information Storage and Retrieval, both two of my favorite courses in Library School. I graduated from the University of Illinois, was selected to be in NLM's Associate Program, and entered that program in September 1970.
>
> During my ten years at NLM I had an opportunity to work on a variety of projects under Dr. Joseph Leiter, and several of those projects involved meetings and discussions with Wilf Lancaster. He was a frequent visitor to NLM and I had the great fortune to be involved in many of those sessions. Wilf was a personal favorite of Dr. Leiter and he would give him tasks to complete that supported Joe's positions. Joe often needed an expert's view on NLM data and Wilf's analysis was just what the good Dr. needed. Frequently Wilf's analysis was a valuable

> tool to prove Joe's position when he was trying to get Marty Cummings' attention or approval.

Rose Marie Woodsmall, longtime employee of NLM who worked on AIM-TWX, MEDLINE, Grateful Med, and PubMed, and is now retired, writes:

> When I came to the National Library of Medicine in July of 1967, everyone was talking about the MEDLARS evaluation that was just concluding. It seems to me now that it was the first time I had ever heard about evaluation in an information setting, and it made a big impression on me that led to a career-long interest in such studies. My printed copy of the 1968 report is one of the few things that I did not pass on when I retired from NLM in 2002. The other connection that comes to mind when I think of Wilf is the NLM Library Associate Program. When our selection committee would meet, one of the first things that was inevitably said was "So did Wilf send us a candidate this year?"—the assumption being that he had made our job easier by recommending a stellar candidate. Thanks, Wilf, for leading us to all of those good librarians and information scientists.

Betsy Humphreys, Deputy Director of NLM, was not at NLM during the time of the MEDLARS evaluation, but had the pleasure of taking Lancaster's continuing education class on the design and evaluation of library services at NLM in 1979. And she still has the book within easy grasp on her bookshelf.

Kent Smith, former NLM Deputy Director, and NLM Executive Officer at the time of the MEDLARS evaluation, recalls that NLM Director Martin Cummings insisted that these important recommendations be implemented where appropriate.

Barbara Rapp, Chief of NLM's Office of Planning and Analysis, writes:

> At the University of Illinois, Wilf inspired a great interest in vocabulary control and the design and evaluation of information retrieval systems. I have drawn heavily on his teachings and publications throughout my career, returning often to the familiar, well-worn and faded book jackets of the indispensable books. They have served me well in many roles—as student, developer, professor, indexer, technical support manager, training coordinator, and program analyst. As a mentor he was also a great influence, and I am grateful for his push to NLM's door through the Associate Program in 1978.

Sally Sinn (MLS '73), former Deputy Chief of NLM's Technical Services Division, writes:

> My time at NLM did not overlap with Wilf's, but I took his Thesaurus Construction course when I attended University of Illinois library school in which he referred often to the work done on information retrieval based upon NLM's MEDLARS system. He was an enthusiastic promoter of NLM and MEDLARS and encouraged promising students to consider applying to the NLM Associate Program. I believe his high regard for the quality of NLM's products and services continued all through his career.

Susanne M. Humphrey, Information Scientist in NLM's Lister Hill Center, was formerly on staff of the Bibliographic Services Division, Library Operations, during Lancaster's tenure from 1965–68. She writes:

> It was a pleasure to work with Wilf in 1993 regarding a design for evaluating MedIndEx, a prototype knowledge-based computer-assisted indexing system developed at NLM's Lister Hill Center. This work was done as part of a six-month contract on which Wilf was co-principal investigator and for which I was the NLM Project Officer. Two specific tasks come to mind that were Wilf's responsibility: determining the gold standard indexing and devising the scoring method to compare the quality of indexing using the system against the standard. The first task required achieving a consensus on the part of experienced indexers as to what was the best MEDLINE indexing for each of thirty test articles which then was reviewed by three NLM revisers. Wilf accomplished this easily and quickly, as was his typical style, with a minimum of sessions with the indexers. The second was developing a scoring method to compare the quality of MEDLINE indexing produced by MedIndex against the standard, including positive and negative values so that an indexer should be penalized for not using a term that appears in the standard and also for using a term that does not appear there. The scoring of MEDLINE is unusually complicated because three types of term—main headings, subheadings, and check tags—exist, and the first two of these can be weighted ("starred") to indicate they represent a "central" concept discussed in the document; moreover, a main heading can have several subheadings, some starred and some not. Despite the complexity, Wilf came up with an original scoring algorithm that was easily programmed in Perl. Wilf's contribution to this contract is reflected in a 1993 NTIS report and a 1996 article in JASIS on evaluating interactive knowledge-based systems.

Tamas Doszkocs, a computer scientist at NLM, considers Wilf as one of the true giants in information science:

> I have memories of awe and admiration for Wilf as his student (and later as a CLIS faculty member) at the University of Maryland. I should also add that that Wilf's classic "Vocabulary Control for Information Retrieval" has influenced my work to this day, including my recent projects on universal meta-search and discovery systems (see http://allplus.com).

References

Adams, S. (1965). MEDLARS: Performance, problems, possibilities. *Bulletin of the Medical Library Association, 53*(2), 139–151.

Armitage, B. (1969). [Review of the book *Evaluation of the Medlars demand search service*]. *Library Association Record, 71*(1).

Badran, O. A., Haman, J., Lancaster, F. W., & Martyn, J. (1977). *Report on the Independent Appraisal of AGRIS* (SC/77/WS/20). Paris: UNESCO.

Batten, W. E. (1973). Document description and representation. *Annual Review of Information Science and Technology, 8*, 43–68.

Bennett, J. L. (1972). The user interface in interactive systems. *Annual Review of Information Science and Technology, 7*, 159–196.

Bourne, C. P., & Hahn, T. B. (2003). *A history of online information systems 1963–1978*. Cambridge, MA: MIT Press.
Brandhorst, W. T., & Eckert, P. W. (1972). Document retrieval and dissemination systems. *Annual Review of Information Science and Technology, 7*, 379–437.
Caceres, C. A., Weihrer, A. L., & Pulliam, R. (1971) Information science applications in medicine. *Annual Review of Information Science and Technology, 6*, 325–367.
Cleverdon, C.W. (1968a). *Comments on the evaluation of MEDLARS*. Internal NLM Document, MEDLARS Collection, NLM Archives.
Cleverdon, C. W. (1968b). The critical appraisal of information retrieval systems. Paper presented at *The International Congress of the International Federation for Documentation*, Moscow, September 1968.
Cleverdon, C.W. (1971). Design and evaluation of information systems. *Annual Review of Information Science and Technology, 6*, 41–73.
Cooper, M. D. (1973). The economics of information. *Annual Review of Information Science and Technology, 8*, 5–40.
Cummings, M. M. (1968). Internal NLM Document. Notes for MEDLARS Evaluation Advisory Committee, January 15–16, 1968. Cummings Collection, NLM Archives.
Cummings, M.M. (1969). Internal NLM Document. Memo to Director, NIH, April 3, 1969. MEDLARS Collection, NLM Archives.
Dee, C. R. (2007). The development of the Medical Literature Analysis and Retrieval System (MEDLARS). *Journal of the Medical Library Association, 95*(4), 416–425.
Evans, G. (1968). [Review of the book *Evaluation of the Medlars demand search service*]. *Journal of Documentation, 24*(4), pp. 320–323.
Harley, A. J., & Lancaster, F. W. (1969). *Structure and uses of vocabulary in MEDLARS II*. Silver Spring, MD: Computer Sciences Corporation.
Herner, S., Lancaster, F. W., & Johanningsmeier, W. F. (1965). A case study in the application of Cranfield system evaluation techniques. *Journal of Chemical Documentation, 5*(2), 92–96.
Jenkins, G. T. (1972). The MEDLARS demand search quality control program. *Bulletin of the Medical Library Association, 60*(3), 423–426.
Karel, L., Austin, C. J., & Cummings, M. M. (1965). Computerized bibliographic services for biomedicine: Library-based automated storage, retrieval, and publication of literature citations is feasible. *Science, 148*(3671), 766–772.
King, D. W., (Ed.). (1978). *Key papers in the design and evaluation of information systems*. White Plains, NY: Knowledge Industry Publications, Inc.
King, D. W., & Lancaster, F. W. (1969). Costs, performance and benefits of information systems. *Proceedings of the American Society for Information Science, 6*, 501–505.
Lancaster, F. W. (1964). *Project SHARP (Ships Analysis and Retrieval Project), information storage and retrieval system: Evaluation of indexing procedures and retrieval effectiveness*, NAVSHIPS 250–210–3. Washington, DC: Department of the Navy, Bureau of Ships.
Lancaster, F. W. (1967a). Internal NLM Document. Memo to Dr. Norman Schumway, January 17, 1967. MEDLARS Collection, NLM Archives.
Lancaster, F. W. (1967b). Internal NLM Document. Memo to Miss Winifred Sewell, January 17, 1967. MEDLARS Collection, NLM Archives.
Lancaster, F. W. (1968a). *Evaluation of the MEDLARS Demand Search Service*. Bethesda, MD: National Library of Medicine.
Lancaster, F. W. (1968b). *Information retrieval systems: Characteristics, testing, and evaluation*. New York: John Wiley & Sons.
Lancaster, F. W. (1969a). MEDLARS: Report on the evaluation of its operating efficiency. *American Documentation, 20*(2), 119–142.
Lancaster, F. W. (1969b). Evaluating the performance of a large computerized information system. *Journal of the American Medical Association, 207*(1), 114–120.
Lancaster, F. W. (1970). *Principles of MEDLARS*. Bethesda, MD: National Library of Medicine.
Lancaster, F. W. (1971a). *An evaluation of EARS (Epilepsy Abstracts Retrieval System) and factors governing its effectiveness*. Report to the National Institute of Neurological Disease and Stroke. Bethesda, MD: National Institute of Neurological Diseases and Stroke.
Lancaster, F. W. (1971b). The cost-effectiveness analysis of information retrieval and dissemination systems. *Journal of the American Society for Information Science, 22*(1), 12–27.

Lancaster, F. W. (1972a). Evaluation and testing of information retrieval systems. In *Enclyclopedia of Library and Information Science* (Vol. 8, pp. 234–259). New York: Marcel Dekker.

Lancaster, F. W. (1972b). *Evaluation of on-line searching in MEDLARS (AIM-TWX) by biomedical practitioners.* (Occasional Papers series, no. 101). Urbana, IL: University of Illinois Graduate School of Library Science.

Lancaster, F. W. (1972c). *Vocabulary control for information retrieval.* Washington, DC: Information Resources Press.

Lancaster, F. W. (1974). *Evaluation of on-line bibliographic searching systems in terms of their suitability for application in the National Instructional Materials Information System (NIMIS).* Bedford, MA: QEI, Inc.

Lancaster, F. W. (1977a). *Guidelines for the evaluation of information systems and services.* Paris: UNESCO.

Lancaster, F. W. (1977b). *Measurement and evaluation of library services.* Washington, DC: Information Resources Press.

Lancaster, F. W. (1977c). On-Line information systems. In *Enclyclopedia of Library and Information Science* (Vol. 20, pp. 394–405). New York: Marcel Dekker.

Lancaster, F. W. (1977d). Pertinence and relevance. In *Enclyclopedia of Library and Information Science* (Vol. 22, pp. 70–76). New York: Marcel Dekker.

Lancaster, F. W. (1977e). *Results of an evaluation of the PASCAL System of CNRS: A report to the Bureau National d'Information Scientifique et Technique.* Paris: Bureau National d'Information Scientifique et Technique.

Lancaster, F. W. (1978). Precision and Recall. In *Enclyclopedia of Library and Information Science* (Vol. 23, pp. 170–180). New York: Marcel Dekker.

Lancaster, F. W. (1979). *Information retrieval systems: Characteristics, testing, and evaluation* (2nd ed.). New York: John Wiley & Sons.

Lancaster, F. W. (1981). Evaluation within the environment of an operating information service. In K. Spärck Jones (Ed.), *Information Retrieval Experiment.* London, Buttersworths, pp. 105–127.

Lancaster, F.W. (1986). *Vocabulary control for information retrieval* (2nd ed.). Washington, D.C.: Information Resources Press.

Lancaster, F. W. (1989). Natural language versus controlled language: a new examination. In C. Oppenheim, C.L. Citroen, & J.M. Griffiths (Eds.), *Perspectives in Information Management* (Vol. 1, pp. 1–23). London: Butterworths.

Lancaster, F. W. (1991). *Indexing and abstracting in theory and practice.* Champaign, IL: University of Illinois, Graduate School of Library and Information Science.

Lancaster, F. W. (1998). *Indexing and abstracting in theory and practice* (2nd ed.). Champaign, IL: University of Illinois, Graduate School of Library and Information Science.

Lancaster, F. W. (2003). *Indexing and abstracting in theory and practice* (3rd ed.). London, England: Facet Publishing.

Lancaster, F. W., & Climenson, W. D. (1968). Evaluating the economic efficiency of a document retrieval system. *Journal of Documentation, 24*(1), 16–40.

Lancaster, F. W., Elzy, C., Zeter, M. J., Metzler, L., & Yuen, M. L. (1994). Comparison of the results of end user searching with results of two searching by skilled intermediaries. *RQ, 33*(3), 370–387.

Lancaster, F. W. & Fayen, E. G. (1973). *Information Retrieval On-lin*e. Los Angeles: Wiley-Becker and Hayes.

Lancaster, F. W., & Gillespie, C. J. (1970). Design and evaluation of information systems. *Annual Review of Information Science and Technology, 5*, 33–70.

Lancaster, F. W., & Martyn, J. (1978). Assessing the benefits and promise of an international information program (AGRIS). *Journal of the American Society for Information Science, 29*(6), 283–288.

Lancaster, F. W., & Mills, J. (1964). Testing indexes and index language devices: The ASLIB Cranfield Project. *American Documentation, 15*(1), 4–13.

Lancaster, F. W., & Neway, J. M. (1982). The future of indexing and abstracting services. *Journal of the American Society for Information Science, 33*(3), 183–189.

Lancaster, F.W., & Rapp, B.A. (1981). Some limitations of methods used in the evaluation of information services. In A.I. Mikhailov & Y.A. Shreider (Eds.), *Theoretical Problems of Informatics* (FID 591, pp. 9–29). Moscow: International Federation for Documentation.

Lancaster, F. W., Rapport, R. L., & Penry, J. K. (1972). Evaluating the effectiveness of an online, natural language retrieval system. *Information Storage and Retrieval, 8*(5), 223–245.
Lancaster, F. W., Ulvila, J. W., Humphrey, S. M., Smith, L. C., Allen, B., & Herner, S. (1996). Evaluation of interactive knowledge-based systems: overview and design for empirical testing. *Journal of the American Society for Information Science*, 47(1), 57–69.
Lancaster, F. W., & Warner, A. (1993). *Information retrieval today*. Arlington, VA: Information Resources Press.
Leiter, J. P. (1967). Internal NLM Document. Memo to NLM Director, January 17, 1967. Cummings Collection, NLM Archives.
Martin, T. H. (1973). The user interface in interactive systems. *Annual Review of Information Science and Technology, 8*, 203–220.
MEDLARS Evaluation Project (1966, October). *NLM News*, p. 3.
MEDLARS Orientation Held (1968, December). *NLM News*, pp. 2–3.
MEDLARS Search Analysts Meet (1968, November). *NLM News*, pp. 2–3.
Miles, W. D. (1982). A *History of the National Library of Medicine: The Nation's Treasury of Medical Knowledge*. Bethesda, MD: National Library of Medicine.
New MEDLARS Training Programs Under Way. (1968, May). *NLM News*, p. 3.
Share Your Training Aids (1968, November). *NLM News*, p. 2.
Wilson, J. H. (1972). Costs, budgeting, and economics of information processing. *Annual Review of Information Science and Technology, 7*, 39–68.

Barbara A. Rapp is chief of the Office of Planning and Analysis at the National Library of Medicine (NLM). She has more than twenty years experience at NLM, including positions as coordinator of the postgraduate Associate Fellowship Program, manager of user services in the National Center for Biotechnology Information, and operations research analyst in the Office of Planning and Evaluation, where she participated in a major study of the impact of MEDLINE for clinical decision making. Prior to NLM, Dr. Rapp was on the faculty of the School of Library and Information at the Catholic University of America, where she taught courses in information systems and managed the program in health sciences librarianship. She has numerous presentations and publications on scientific databases and information retrieval systems. She is also an active member of the Medical Library Association.