# The Scope and Operating Efficiency of Information Centers as Illustrated by the Chemical-Biological Coordination Center of the National Research Council

By RICHARD M. DOUGHERTY

THE DESIRE to improve our information handling techniques and the quality of information services is illustrated by the growth of specialized information centers based, at least in part, on machine technology. This study of the Chemical-Biological Coordination Center, one of the first of the major mechanized science information centers, was undertaken as one step in evaluating the strengths and weaknesses of this approach, and also to suggest means by which such organizations may be made as effective as possible.

The Chemical-Biological Coordination Center of the National Research Council (hereinafter referred to as the center, or the CBCC) was established July 1, 1946. The center was an outgrowth of World War II screening programs that had generated large quantities of data on the effects of specific chemical compounds against insects and rodents. These programs were attempts to discover substitutes for compounds made unavailable by the war. The urgency for screening programs diminished with the termination of hostilities, but many individuals believed that the scientific data they generated should be made accessible to scientists in general. This led to the concept of a central clearinghouse.

*This article is based on a research project carried out in the Graduate School of Library Service, Rutgers, the State University, New Brunswick, New Jersey. The project was supported by the United States Air Force under Contract No. AF-AFOSR-62-55 monitored by the Information Services Division, Air Force Office of Scientific Research of the Office of Aerospace Research.* *Dr. Dougherty is now Head of the Acquisitions Department in the University of North Carolina Library.*

---

The CBCC's objectives were to collect, organize, and disseminate data on chemical compounds and their effects on biological systems. The center planned also to sponsor and administer a chemical screening program, like those conducted during the war and, finally, to conduct symposia and to publish reviews.

Chemical-biological correlation, the determination of broad relationships between chemical structure and biological responses, was the center's ultimate objective. This goal was stressed because funds had been solicited from several agencies on this basis, even though actual correlation studies were far from realization in 1946.

The advisory committees of the CBCC

decided that IBM tabulating equipment and punched cards would be used for storage and retrieval of data. This decision was based on two assumptions: one, the equipment would permit greater handling speed than could be achieved with any available manual system; and two, the punched cards would provide multiple access or entry to data stored on the cards. Once the system had been agreed upon, the center turned its attention to other matters—namely, the need to devise codes and notational schemes.

Devising the codes, particularly the biology code, proved more formidable than had been anticipated. The authors of the codes were scientists, but code-building was alien to their normal activities. While they were qualified to define and to organize scientific concepts and terms, building a code to be handled by punched cards was another matter.

The objective was to construct a code for classifying biological information for eventual correlation with structural and physical characteristics of chemical compounds, and to develop a notational scheme which would make possible the transfer of data onto punched cards. The duality of this objective presented serious problems.

After several false starts, a biology code was completed. It required the efforts of over sixty scientists for almost four years to develop the nucleus from which emerged the code that was finally adopted. Initial efforts were directed toward the development of a general code, but as work progressed, the apparent need to code in greater detail led some individuals to work on codes within their own fields of specialization, *i.e.,* entomology, pathology, etc. By 1948 a tentative general code was completed. The center's executive committee directed that this code be implemented by the actual coding of data. This did not, however, halt efforts to develop detailed codes for special topics.

Following a trial period, the general code was declared unsatisfactory because it did not allow for coding in sufficient detail. This caused the center again to focus its attention on detailed codes.

One year later the center invited a former staff member, who had previously been instrumental in developing a detailed code for entomological data, to return to the center to work in cooperation with the center's staff to develop a new code. Using the general code and the various detailed codes, this effort produced the foundation of an integrated biology code.

Returning again to 1946, the most frustrating problem for the code builders was that of devising a code pattern that could be used for relating chemical structure with biological activities on a single punched card. No satisfactory pattern emerged, mainly because of restrictions imposed by the punched card.

By 1948 three coding patterns were under consideration: Codes A, B, and C. Code A required the use of secondary cards, that is, if a requester required more biological data than could be provided by the primary cards (those used in conjunction with the 1948 general code), the secondary cards, which were keyed to the primary cards, could be consulted. Codes B and C took into consideration the fact that as the files expanded the practicality of sequential sorts would diminish. As an alternative they proposed that the punched cards be prefiled by subject categories and duplicate cards be produced for the multiple files. Codes B and C differed in that the former proposed to retain the chemistry and biological data on the same card, while Code C proposed to store the data on separate cards and key them together with a reference number.

To check the merits of each pattern, each coding pattern was used to code chemical and entomological data relating to 278 compounds. The efficacy of each pattern was then tested by a series of questions. The tests proved inconclusive,

but they did show that Code A, using the secondary cards, was impractical because of the excessive machine time required in the retrieval process, and because of the need for double punching in the input process. As for patterns B and C, neither was shown to be clearly superior to the other. Code C allowed for greater file expansion and for coding in detail, but Code B permitted both types of data to be stored on the same card. Pattern B was adopted.

Two years later this decision was reversed. As the revised biology code (1950-51 version) took shape, it became apparent that there would not be sufficient storage capacity for both types of data on the same card. Consequently, Code C, which had been rejected previously, was in principle now adopted. The decision at this late date, however, was expensive because data already coded and punched had to be reprocessed.

In addition to the biology code, the center also worked on a chemistry code between 1945 and 1950. This code was designed so that organic and inorganic compounds could be described by listing their constituent groupings, both functional and nonfunctional, by linear symbols. And like the biology code, it permitted the transfer of code symbols to punched cards. Although the codification of chemical structures did not prove as difficult as codification of biological systems, the chemistry code was not without its limitations. For example, the code could distinguish between structural groupings but could not designate points of attachment. Thus, after mechanical sorts, the retrieved cards had to be manually inspected by a chemist to determine whether the selected groupings actually formed the desired compound.

The chemistry code was completed and was published by the National Research Council in 1950.

While the code-building projects received priority, the center also formulated policies concerning the selection and collection of data, and developed internal procedures for handling data. Data for coding were selected on the basis of potential importance to scientists, adaptability to the CBCC files, and future usefulness to the center in making correlation studies. Originally the center aimed at collecting data on "lesser-known" compounds rather than those more frequently cited, such as DDT and Chlordane. This policy was later reversed, however, in favor of the selection criterion of uniqueness, *i.e.,* reactions of compounds unique to the center's files.

The primary sources of data for storage were from the center's screening program, unpublished reports, and selected scientific periodicals and other published materials. Coverage was intended to be broad. It included such disciplines as pharmacology, entomology, biochemistry, and medicine, among others. The number of journals containing 90 per cent of the data pertinent to the center's objectives was estimated to be two hundred and fifty. The number of journals actually processed ranged from thirty-five to fifty-five.

At first the center employed nonresident scientists to scan and abstract articles from assigned journals. The abstracts were later coded at the center by resident staff members. But because of coding errors detected during code trials conducted in 1951, the input procedure was revised. The new procedure delegated both coding and abstracting of biological data to nonresident personnel. Coding of compounds was done by the chemistry group at the center. The most drastic departure in the revised input procedure was the introduction of a double inspection. Coded data were checked and rechecked by scientific personnel.

Code sheets for recording chemical and biological data were devised. Biological data were recorded two ways: in coded form, and in abstracted form. Code symbols were also punched on standard IBM cards. The cards were then duplicated

and placed in files according to a system of predesignated subject categories. Mechanical equipment available to the center in 1947 included two types of one-column sorter, a collator, an interpreter, a tabulater, and a reproducing punch. The center later gained limited access to an IBM 101 statistical sorter.

To summarize, 1946 through 1951 could be characterized as the period of development and preparation. The codes were devised, procedures established, and policies set forth, but during this period the center had not demonstrated the practicality of its techniques, nor had it really tried. Activities emphasized the research aspects of the center's functions. At this time, however, the center assumed the role of a service agency.

An *ad hoc* committee appointed in 1950 to evaluate the center and its operations reported that the organizational phase was over, and recommended that the center should begin to demonstrate its value. This view was in essence supported by a second committee appointed in 1952. This second committee recommended specifically that the center focus its attention on providing information services. The group was not unanimously in favor of continuing the center. Some members believed that it had not shown its worth, and that operations were proving more costly than had been anticipated. But the consensus of the group after deliberation was to recommend a one-year trial period, a period which was to last three years before a successor committee was convened.

From its beginning, the center had not been able to stabilize financial support. Most of its funds were received from five agencies, four government agencies and one nonprofit foundation. The need to broaden the base of support prompted the center to approach other agencies, both governmental and private. These efforts, except for a $1,000 donation received from a private firm, proved fruitless.

In total the center received over one and one-half million dollars in direct support. Annual income averaged about $175,000. During the latter stages this gradually declined. What levels were actually needed in order to sustain and expand operations was never ascertained. Estimates advanced ranged from $400,000 to $800,000 annually.

Although the objectives to be achieved during the 1952-53 trial period were not clarified, the center converted to operational status. Sponsoring of symposia and publication of reviews, which had formerly been considered important, were discontinued. Priorities were given to three functions: storage of data, provision of information services, and operation of the screening program.

The center collected data on approximately sixty-three thousand compounds and two hundred and eighteen thousand biological responses. The punched card files contained almost one and one-half million prefiled cards. But in view of the center's intended scope of coverage, the quantity of data stored represented only a meager beginning.

The revised input procedure proved cumbersome. It created unbalanced internal work flows. For example, the biology group between January 1953 and December 1956 coded a total of one hundred and forty-eight thousand lines of data; the number of lines completely processed (coded, and inspected twice) and released for final use totalled ninety-six thousand lines. The difference of over fifty thousand lines represented a backlog of more than one year's production. This problem was never resolved.

Accurate cost studies were never conducted. One estimate placed the cost per article processed at $29.46, the cost per compound coded at $5.43, and $2.18 as the cost per line of biological data. A second estimate based on the number of code sheets completely processed and released for filing during fiscal year 1955 placed input costs at $3.67 (chemistry

and biology) per line of data processed, or $50.00 per article.

Coding errors and inconsistency of data were constant problems. The center adopted direct coding to reduce coding errors, but this proved ineffective.

Control of internal processing was not the only problem. The center had to deal with external inconsistencies which were beyond their control; these included inconsistences in scientific nomenclature, incomplete reporting of test results, and variations in testing procedures.

The center's principal means of demonstrating its value to sponsors and the scientific community at large was the provision of information. As recommended by the evaluation committees, the information service was accorded top priority. The center's facilities were extended to members of sponsoring agencies, members of official screening agencies, and other qualified scientists. Requests for information came from all parts of the scientific community. In total the center processed slightly over thirteen hundred requests. The heaviest demands were received from private companies and academic institutions, none of which contributed to the center's support.

A study of 1,025 requests received at the center between January 1953 and October 1956 shows that one-third of the requests were answered, one-quarter partially answered, and the remaining 45 per cent unanswered. The sole criterion employed in this evaluation was whether or not the information requested was supplied. For example, related information might have been useful, but such information was not considered an answer.

The time lapse between receipt of requests and sending of replies ranged from one day to more than one year. Three-fifths of the requests were handled in two weeks or less. During the latter stages of operations, a backlog of unprocessed requests accumulated, and by the middle of 1955, it amounted to almost eighty requests.

To determine whether the center made a unique bibliographic contribution, the sources used in answering questions were investigated. This analysis revealed that of the requests answered or partially answered, only one-third were based on data originating in the center's files. The remaining answers came from conventional indexes, abstracting bulletins, bibliographies, textbooks, etc. Machine utilization in the retrieval process was low. Records indicated that the machines were employed in answering from 3 to 15 per cent of the requests, and that during the latter stages the punched card files were consulted almost exclusively on a manual basis.

Retrieval costs per request was estimated by the center variously at $60.00 and $150.00—the first figure based on unit costs, the second on over-all operational costs. The analysis of information requests described above showed that input costs, based on 1955 cost figures, per successful use of the files was approximately $1,850.

A primary objective of the center had been to develop techniques for performing chemical-biological correlation studies. Between 1953 and 1956 almost fifty questions involving chemical-biological correlations were received at the center, but no correlation studies were undertaken. The reason most commonly cited was insufficient data.

The screening program was intended to facilitate the preliminary testing or screening of compounds on a variety of plants, animals, and microorganisms by making compounds available to scientists, to collect unpublished data, and to disseminate these data. In total, over ten thousand different compounds were offered to screening agencies and others. Screeners selected 55,000 samples, of which the center was able to supply 75 per cent.

The center received approximately forty-two thousand lines of data as a result of screening activities. Sponsors or sponsor-related organizations accounted for 40

per cent of the data received. Seventy-five per cent of the data were eventually published in the center's publication, the *Summary Tables of Biological Tests*. Four companies reported in 1955 that they had compounds in pilot plant or some stage of commercial development as a result of the program. There was no evidence, however, that many compounds were developed commercially. The major deterrent to commercial development, as expressed by most individuals, was inadequate patent protection.

Costs of the screening program were relatively fixed. Over-all costs per compound ranged between $30,000 and $35,000. Input costs per line of screening data were estimated to be $.85. This figure did not include costs of administering the program, handling compounds, correspondence, etc. With these additional factors included, the cost per line was found to be over $6.00, which was almost double the cost of processing data from the literature.

Early in 1957, the NAS-NRC announced the termination of the center. Inability to attract stable, long-range financing was cited by the academy as the basis for its decision. Reaction ran the gamut from disgust to complete agreement. Some believed the center was accomplishing an important job and should have been allowed to continue. They pointed to the screening program, the development of the codes, and the information service as positive achievements. Conversely, others felt the center had been unable to define or limit its objectives and scope of operations, which ultimately led to dilution of programs.

While the center failed, its design and operation typified those of information centers now in existence. This is particularly true with reference to the intellectual and mechanical skills which are required to operate such organizations. The investigation of CBCC showed that the operation of information centers such as the CBCC requires four categories of skills which are not necessarily provided by a staff consisting entirely of scientists; these skills are subject specialization, bibliographic competence, knowledge in depth of the devices and mechanisms available for achievement of bibliographic operations, and administrative ability.

The development of codes requires specialized subject knowledge. To satisfy these substantive intellectual requirements, it is necessary to employ individuals with specialized subject backgrounds. Equally important in code development, however, is the formulation and standardization of definitions, cross-referring of synonyms, and the construction of notational schemes. These tasks require the skills of the indexer and code builder, and these skills are not necessarily provided by men with specialized subject knowledge.

Bibliographic competence can be divided into two aspects. The first is the ability to achieve control of the literature. To achieve bibliographic control, it is necessary to have a thorough knowledge of the bibliographic structure of areas pertinent to an organization's interests. The second is the ability to convert information (by coding, indexing, abstracting, etc.) into a form suitable for storage, retrieval, and dissemination. These functions require competence in bibliographic principles and associated skills, such as indexing, code construction, abstracting, cataloging, and the provision of reference services.

While the codes (this would include any technique employed for analyzing subject content) provide the intellectual framework within which data can be organized, stored, and retrieved, the procedures and devices adopted for storage and retrieval provide the "mechanical" framework for handling data. This function requires knowledge of information handling systems, both manual and machine, and of the capabilities and limitations of each system, so that selection of an optimum system can be based on a

*(Continued on page 20)*

can be more flexible, directed to and even varied according to the needs of its users, but it must have a consistent practice or it will not avoid inconsistencies, partly caused—as LC says of its own headings—by "varying theories of subject heading practice over the years."

In the writer's view, and of course he is not alone, a new body of theory and practice must be laid down and taught not only as the basis for the specific entry of the dictionary catalog, divided or undivided, but also as the basis of any indexing or information retrieval. He does not accept the view that documentation or information retrieval must be, or is, essentially different from library subject cataloging, but he does think that librarians who have neglected their own half-solved problems of cataloging are at least as much to blame for the opposition of information retrieval and library cataloging as the amateurs, the engineers, and the chemists are, with their usually ignorant and prejudiced assumptions of what is covered by the word cataloging. He is sceptical of theorizing with little relation to the proven theory and practice of the past, and of new names for old things such as documentation and information retrieval for subject indexing, descriptors for subject headings, and so on; he thinks Cutter in his specific entry definitions and rules and Kaiser with his concrete-process breakdown made permanent progress, independently but essentially on the same lines. Theirs were major steps towards logical subject indexing, whether mechanized or not, and whether arrangements of subject names such as theirs are used directly or indirectly.

Cutter's theory and practice was exemplified in LC cataloging in its subject headings, though with some unexpected deviations from the master. The writer thinks a development of Cutter's practice, and Kaiser's, should be exemplified in an ideal subject headings list, the compilation of which would be the inductive corrective of the deductive approach, from principles. But this would be detail at the technical level. The cataloging in question is American, and even as it is—though we may receive it critically—we receive it with admiration and gratitude. Whatever might be done to improve it would have to be done in America, and whatever is done or not done, we hope that we may continue to receive it with admiration and gratitude. ■■

# Information Centers . . .

*(Continued from page 12)*

comparative study of alternative approaches. Without comparative studies it is not possible to demonstrate objectively the practicability of any system.

Finally, the study underscored the importance of administration. An information center, like any other organization, is subject to the principles of administration. While this is axiomatic, and certainly not startling, its importance has again been demonstrated. In the absence of sound administration, it will be difficult, if not impossible, to blend the varied skills described above into an integrated, effective organization. ■■