HERMES D. KREILKAMP

The National Agricultural Library's Data Base: AGRICOLA

The origin and development of AGRICOLA, the National Agricultural Library's data base in Beltsville, Maryland, are summarized and its system and resources explained. Evaluations made at the Universities of Arizona, California, Florida, Georgia, Minnesota, Wisconsin, and Pennsylvania are presented, together with observations about the data base made at several international research centers.

AGRICOLA (AGRICultural On-Line Access) is the new name of the National Agricultural Library's (NAL) data base, formerly called CAIN (CAtaloging and INdexing system of NAL). Not long ago a professor of forestry, using NAL's CAIN, wrote a six-page letter to the Steenbock Memorial Library of the University of Wisconsin. In it he indicated that two bibliographic searches conducted for him by the library had increased his (manually prepared) file of references by 63 percent. This remarkable addition to a carefully prepared bibliography indicates the potential of data bases such as AGRICOLA to expand bibliographic research.[1]

Besides the growing realization of the tremendous resources data bases add to any bibliographic research, there is also a growing conviction that widening access to these data bases will play an important part in the economic and social growth of our nation and of national economies throughout the world. International cooperation in collecting and disseminating information is well under way in the fields of chemistry, physics, medicine, and nuclear energy; but it is still generally unsatisfactory in the coverage of agricultural literature, where over 200,000 documents are produced every year.

AGRICOLA, NAL's outstanding data base, is part of our national effort toward a more adequate bibliographic control of these documents. It is now on-line at over thirty-five of our nation's universities and is playing an important part in assisting agricultural researchers and educators of future farmers.

AGRICOLA contains, first of all, the titles of all monographs and serials contained in the National Agricultural Library at Beltsville, Maryland, one of the three national libraries in this country and a depository of books and information on agricultural science gathered from all over the world for more than 100 years.[2]

The Origins of AGRICOLA

In 1964 a computer group was formed at NAL to develop methods for automating the burgeoning collection of materials in the library.[3] In 1967 the group devised a system for preparing subject and author indexes for a bibliography of agriculture. At about the same time, the contract for programs for the U.S. Department of Agriculture's (USDA) *Pesticides Documentation Bulletin* became operational. In 1969 these programs were modified, and additional programs were written in order to process all agricultural data at

Hermes D. Kreilkamp is research librarian, St. Joseph's College, Rensselaer, Indiana.

NAL within one set of parameters. These became the CAIN system, with their data placed on tape and made available to anyone wishing to purchase them. They became in this way the first major agricultural machine-readable and readily available data base.

Records from the Food and Nutrition Information and Educational Materials Center (developed to promote the national school lunch program) were added in 1973. These records differ from the other bibliographic data first entered into the data base insofar as they included not only titles and the names of authors of articles and monographs, but also abstracts of all materials on these subjects, abstracts which are also searchable in the data base.

In June 1973 the federal government awarded a contract to Lockheed Missiles and Space Company to place the data from the CAIN tapes on-line (i.e., to provide an interactive bibliographic search-and-retrieval service for anyone wanting to search the data base from 1972 onward). In this agreement the tapes were provided to Lockheed by NAL. Lockheed provided for the conversion of the data to Lockheed's format for computer storage and processing for on-line retrieval, plus the software to access the data base, and leased to NAL terminals which were installed at two NAL locations.[4] This procedure provided greater and more economical access to the data base for citizens via a nationally available commercial information system.[5]

AGRICOLA is on-line today, not only with Lockheed, but with System Development Corporation (SDC). The advantage of such an arrangement is that the U.S. Department of Agriculture can have its data base on-line all day (without constructing its own retrieval system). AGRICOLA is designed primarily as a document locator and a means of bibliographic control for NAL. Its master record today contains over 850,000 items which have been accessioned sequentially according to the year they were processed at NAL. The important searchable data in this data base are:

1. *Subject elements:* (a) *titles* (always in English); (b) *subject category codes* (a maximum of two for items selected from seventy possible classifications); (c) *subject terms* (assigned at present only to monographs); and
2. *Author and citation elements:* (a) *personal and corporate authors;* (b) *journal title;* (c) *document type and language* and the availability of a translation.[6]

On-line today, these items are accessioned by means of an inverted file, on their call numbers, titles, authors, or subject terms.

Scope of AGRICOLA

AGRICOLA includes all items cataloged and indexed by NAL. By-products include catalog cards for eight separate catalogs, photocopy for a monthly book catalog, and a "sale tape" for subscribers to the data base and for the printing of the *Bibliography of Agriculture.* Approximately 135,000 items are added to these tapes each year, 120,000 of which are journal articles from 1,200 journals, for which all the articles are entered, and from 4,000 other journals from which selected articles are chosen.[7]

The subjects of these items range from those which are essentially agricultural—such as agricultural economics, rural sociology, agricultural products, animal industry, engineering, pesticides, plant science, soils, and fertilizers—to such as are included because of their relevance to agriculture: botany, chemistry, entomology, forestry, food and nutrition, law, water resources, and economics in general.

In January 1976, over 5,000 records of the American Agricultural Economics Data Base were added as one of three subunits of AGRICOLA; search approaches to this file include article titles, personal authors, corporate au-

thors, source codes, key words, abstracts, NAL call number, and publication date.[8] All the above data are transcribed upon tapes which are nine-track, 800/1600 b.p.i., blocked two in EBCDIC, with standard IBM 360 header and trailer labels.

Data for these tapes are inputted into the system via the staff of NAL, from the Cataloging Section and the Indexing Section. Although earlier NAL used IBM Hollerith cards for reading information into the data base, since 1974 data are keyed directly to a disk, using CRTs and a minicomputer.

Pre-stored formats are called to the CRT from the minicomputer in entering the data onto the disk, and the data are entered in normal upper and lower case without diacritical marks. An average of four formats is needed to enter one item. The data are then transmitted by telephone lines to the Washington computer center at USDA, where the processing is done in batch mode. The data are checked by computer against an authority file, and errors in regard to the number of characters in a field, etc., are indicated on the CRT screen by a blinking light. By this method of editing and processing the documents, the items are processed twenty-four days quicker under the present system than they were under the original system using punched cards. The edit and update system is such that the fields can be deleted or changed in whole or in part, and this is something which is accomplished at each update run.

Three main types of output which are constantly updated are: a master file; activity notices (every action submitted or system generated is reported); and error notices (system-discovered omissions). There are four major modes of publication: NAL's catalog cards, selected weekly, sorted, and distributed; the NAL book catalog, printed monthly, with listings by main entry and by authors (the index portions of which are cumulative semiannually); the catalog of the Food and Nutrition Information and Educational Materials Center; and bibliographies prepared for printing via Linotron.

AGRICOLA tapes are currently available in two forms: one from NAL, which publishes it monthly; and the other from Oryx Press, which publishes an abstracted form of these to which keywords are added (from a thesaurus) to the titles of articles included on the tapes. The version from Oryx Press excludes foreign titles and monographs and includes the serial article titles only. Whereas AGRICOLA is written in EBCDIC in its own format, Oryx publishes its abstracts in ASCII or BCD, written in MARC II format. Whereas AGRICOLA's software is written in COBOL, that of Oryx is in COBOL and also IBM Assembler.

Appraisals of AGRICOLA

There have been in recent years a number of studies evaluating the CAIN (now AGRICOLA) data base at universities or national research centers running SDI services or retrospective searches.

A Swedish study of SDI services (1971–1972) showed the average relevance of hits retrieved from the CAIN data base to be regularly 70 percent. The same study noted also the broad coverage of agricultural literature characteristic of CAIN and particularly of East European literature. This study considered the main limitation of CAIN to be its lack of more subject category codes. Yet users generally were very much satisfied with it and relied on it for their current awareness needs, even though they paid for the service. They noted, however, that CAIN needed to be supplemented by other sources for more complete coverage.[9]

The University of Florida, which had for several years been conducting SDI services utilizing CAIN for 220 profiles, observed a constantly increasing usage of CAIN and also noted its 70 percent

relevancy on hits retrieved from the data base. Users of CAIN at Florida also voiced the desirability of more enrichment of titles and wanted more coverage of food science and agricultural economics.[10]

In the Netherlands, the Center for Agricultural Publishing and Documentation (PUDOC) conducted a study running SDI profiles against the CAIN tapes and those of the *Bibliography of Agriculture* and concluded that the CAIN tapes were preferable because of their inclusion of monographs and foreign language versions of titles (items not included in the *Bibliography of Agriculture*). The same study also noted the good coverage of East European literature and expressed the desire to see CAIN titles supplemented by subject descriptors and more category codes. Also noted was the need to supplement CAIN with other sources.[11]

The Office of Computing Services at the University of Georgia at Athens has its own system of searching twenty commercial tape services for some 1,660 users. It noted that the users of the CAIN data base were growing by roughly 200 each year. Retrospective searches of CAIN were run in batches (of about twenty) at two-week intervals. Users of CAIN at Georgia also noted the good, broad coverage of agricultural literature and the low acquisition costs; yet they wanted to see the CAIN titles supplemented by more subject descriptors and more subject categories. At Georgia, too, it was noted that CAIN needed to be supplemented by other sources as well (e.g., by *Biological Abstracts* (BIOSIS today)).[12]

The Agricultural Research Service (ARS) at Beltsville, Maryland, runs SDI profiles for a user group of over 1,000—of which 600 were run against CAIN. But here also, CAIN was always searched in combination with other data bases, such as *Biological Abstracts*, since it is generally recognized that although CAIN had references lacking in others, it was never complete. ARS uses a modified version of the University of Georgia's software for its services—software capable of searching a wide variety of data bases—and noted CAIN's particularly strong coverage of foreign literature in the field of agriculture. But ARS expressed criticism of the time lag between the publication of some articles and their appearance in CAIN. Campey noted, however, that ARS still ran more SDI profiles against the CAIN tapes than any other organization studied, despite these criticisms.[13]

Norwegian use of the CAIN data base afforded apparently the least user satisfaction, due perhaps to the fact that users were required to construct their own profiles and to initiate profile reviews without the assistance of professional interface. Perhaps another reason was that the software used in Norway placed severe limits on the size and complexity of profiles. The relevance of hits retrieved from CAIN in Norway ran only 30 percent so that users, on the whole, felt dissatisfied.[14]

The University of California had CAIN SDI services in operation since January 1972 and has had CAIN on-line since October 1973. Seven profiles run there against both forms of service, from January through May 1974, showed that although it took more library staff to provide on-line services with CAIN, much greater service was delivered (if service is measured by the amount of literature searched). The average relevance of hits from SDI searches (in an earlier period) was 70 percent, and user reaction to CAIN was again generally favorable.[15]

General conclusions drawn from Campey's study were, first, that there appeared to be a definite correlation in many instances between user satisfaction with CAIN and the amount of professional interface service provided to assist the user in questioning the data base. Wherever such professional assistance was lacking, user experience with

NAL's data base was usually unsatisfactory. For this reason, probably, the U.S. Department of Agriculture has begun on-line training courses in the use of AGRICOLA for its own personnel and land-grant university librarians.[16]

Campey concluded that the data base provides good, but not complete, coverage of agricultural literature and that if comprehensive coverage is to be achieved, the search of NAL's data base must be supplemented by alternative services, including those which cover the agricultural literature within other scientific disciplines. As to the suitability of this data base for SDI services, Campey concluded that it can be used satisfactorily as an economically justifiable SDI service and for general coverage of the current agricultural literature. It appeared to be the best available machine-readable data base anywhere in the world, needing, however, to be supplemented by alternative data bases.

Campey noted that although there were data bases which appeared to be outperforming CAIN for SDI services, even in these cases there was still evidence of extensive and expanding usage of NAL's data base for SDI purposes. This suggested it is providing an essential service to our nation's economy.

These conclusions appear to be borne out also by other recent studies. Douglas Leadenham conducted a study of CAIN on-line in 1975 at the University of Arizona. In this study CAIN's record was compared with the performance record of other manual and on-line data bases and was found ranking second only to BIOSIS. CAIN on-line was shown to be far easier to use than its printed equivalent, the *Bibliography of Agriculture*, and gave better results.[17]

This study also indicated the desirability of cooperation between CAIN and the Commonwealth of Agricultural Bureaux (CAB). The suggestion was made that if these data bases were combined, one would obtain "complete" worldwide coverage of agricultural literature. It is interesting to note that Lockheed recently has put CAB on-line with AGRICOLA, thus providing the kind of coverage Leadenham suggested.

At the University of Wisconsin, a study of CAIN on-line revealed an even higher relevancy rate (72.8 percent) than those previously encountered.[18]

Another, at the University of Minnesota, suggested several ways CAIN might be improved still further. On-line (with Lockheed) CAIN has right-truncation power, enabling users to search for roots (e.g., ENZYM). Reich and Hearth suggested in this study the desirability of adding a program which would allow for left as well as right truncation. The word from Lockheed, although unofficial, was that it had plans to add this feature to its system and also to improve its present system so that it could accept an instruction which generates more than 100 roots (the limit of its present truncation power).[19]

Oyler and McKay at the University of Wisconsin have noted how CAIN illustrated the critical nature of the interplay that goes on between the searcher and the user in querying the system and the need for knowledge of the system and its thesaurus to use it efficiently.[20]

A study of on-line use of NAL's CAIN data base at Colorado State University, performed in two phases, noted that user satisfaction was raised from 40 to 60 percent, due not only to the increasing skill of terminal operators but especially to the fact that in the second phase of the operation, users were required to be present during the processing of their searches.[21]

Conclusions

A number of conclusions may be drawn from all these evaluations of CAIN, now AGRICOLA. The first is that a large data base such as NAL's can be operated successfully and commercially on-line, especially when available in a system also comprehending other

large data bases on-line. The trend of research in the future undoubtedly will be to search data bases of a similar nature, even though this may involve some overlapping and duplication of items retrieved.

There seems to be a definite future for automatic indexing such as that used in AGRICOLA, although the experience of CAIN users indicated the desirability of enriching certain key words with scientific or popular names which might suggest themselves for greater clarity. The tapes provided by NAL have proved their multiple advantages not only in the variety of searches made on them via computers but also in the variety of outputs, many of which admit of a variety of commercial marketing as well. Adding subject descriptors to bibliographic titles may add to the accessibility of data, but it also will add to the cost of data bases.

User satisfaction with a data base such as AGRICOLA depends to a large extent not only upon the skill of terminal operators, but also on the presence of the user and the interface of the two in formulating questions.

As the study of CAIN on-line at the University of Pennsylvania concluded, the searching of data bases is probably best left to a librarian familiar with the computer language, commands, and codes as well as the nature of a particular data base and the strategies necessary to use it to best advantage. "Working together, the scientist and librarian are likely to be more efficient than either one alone."[22]

References

1. J. V. Caswell and D. K. Oyler, "Specialized Data Base Utilization," *Agricultural Library Information Notes* (cited hereafter as *ALIN*) 2:3 (March 1976).
2. J. Caponio and L. Moran, "CAIN: A Computerized Literature System for the Agricultural Sciences," *Journal of Chemical Information and Computer Sciences* 15:3 (1975).
3. V. J. Van Dyke and N. L. Ayer, "Multipurpose Cataloging and Indexing System (CAIN) at the National Agricultural Library," *Journal of Library Automation* 5:1 (March 1972). For many of the details of this data base the author is indebted directly to Mr. Van Dyke.
4. *Quarterly Bulletin of the International Association of Agricultural Librarians and Documentalists* 18:4 (1973).
5. As Cuadra has noted, the key to operating any data base successfully today is to have a large number of users to share the high costs of computer time; but to do this, one needs also to provide access to many data bases. See his article "SDC Experiences with Large Data Bases," *Journal of Chemical Information and Computer Sciences* 15:1 (1975).
6. L. H. Campey, *User Reactions to CAIN* (Luxembourg: Commission of the European Communities, 1974).
7. Charles L. Gilreath, *CAIN On-Line User's Guide* (Beltsville, Md.: National Agricultural Library, 1975).
8. *ALIN* 3:2 (Feb. 1977).
9. Campey, "User Reactions to CAIN," p.4–5.
10. Ibid., p.5–6, 27–30.
11. Ibid., p.6–7, 30–36.
12. Ibid., p. 7–9, 36–38.
13. Ibid., p.9–10, 38–41.
14. Ibid., p.10–12, 41–42.
15. E. C. Jestes, *A Comparison of Sub-Current Awareness and On-Line Retrospective Bibliographic Search Services from the CAIN Data Base* (Davis, Calif.: University of California, University Library, 1975).
16. *Annual Report 1975* (Beltsville, Md.: National Agricultural Library, 1976).
17. D. J. Leadenham, *Comparison of CAIN On-Line to Other Agricultural Indexes at the University of Arizona* (Tucson, Ariz.: University of Arizona, 1975).
18. D. K. Oyler and M. W. McKay, *An Analysis of CAIN On-Line* (Madison, Wis.: Steenbock Memorial Library; Beltsville, Md.; National Agricultural Library, 1975).
19. P. Reich and F. E. Hearth, "Evaluation of CAIN On-Line, St. Paul Campus Libraries, University of Minnesota," *ALIN* 2:1 (Jan. 1976). See also *ALIN* 1:9 (Sept. 1975) for the study at Auburn University.
20. Oyler and McKay, *An Analysis of CAIN On-Line*, p.15.
21. C. L. Gifford, "CAIN On-Line Testing and Assistance at Colorado State University," *ALIN* 2:4 (April 1976).
22. Keith E. Roe, Vladimir Micuda, and Robert S. Seeds, "Literature Searching with the CAIN On-Line Bibliographic Data Base," *Bioscience* 25:12 (Dec. 1975).