Production of a Comprehensive Research Directory from Multiple Secondary Sources†

STEPHEN J. TAUBER* and ARTHUR W. ELIAS**

Informatics, Inc., Rockville, Maryland 20852

JOHN H. SCHNEIDER

National Cancer Institute, Bethesda, Maryland 20014

Received September 8, 1974

Printed, automated, and manually maintained sources of information about cancer research and control activities were identified. Data were consolidated into an automated file to characterize the organizations active in such work and to identify individual projects and scientists. A computer-formatted directory was produced, with access via geographic location, personal name, organizational name, and keyword.

The importance of informal communication channels is well recognized.[1-3] The desire to know, "Who is doing what?" also provides a market for reference works such as the "Directory of Graduate Research" in chemistry and chemical engineering[4] or "Directory of Cell Research Laboratories."[5] However, there existed no single source of information on scientific activity in cancer or in its subspecialties such as chemotherapy or carcinogenesis. Conscious of the injunction of the National Cancer Act of 1971 "to take necessary action to insure that all channels for the dissemination and exchange of scientific knowledge and information are maintained between the National Cancer Institute and other scientific, medical, and biomedical disciplines and organizations nationally and internationally,"[6] we therefore undertook to compile a comprehensive, worldwide directory of cancer research and control projects and institutions.

PRE-EXISTING SOURCES

From prior personal knowledge, use of libraries, and professional contacts, we identified several sources whose contents pertain to cancer work.[7-18] Two of these sources[7,8] are computer based, one is a private file,[9] and another is a limited-distribution offset list.[10] The considerations involved in combining the data from these sources are the subject of this publication.

The Smithsonian Science Information Exchange (SSIE) maintains a file of research summaries. Each research project financed by U.S. government funds (except classified work) is expected to supply information. In addition, information is included about projects funded by several other organizations. This file is therefore strongly biased to information about work in the United States. A July 1973 search of the entire SSIE file turned up 4,383 projects which according to their subject index entries are cancer related. All but 116 were being performed within the United States. Most of the foreign work recorded in the file, furthermore, was sponsored either by the U.S. government or by U.S. organizations such as the American Cancer Society, the Damon Runyon-Walter Winchell Cancer Fund, or the Tobacco Research Council. Only 17 of the projects were *funded* by non-U.S. sources, 13 of them by the International Atomic Energy Agency.

The Tokyo Science Museum's REGISTER System[8] maintains (in Japanese) information about Japanese research laboratories. It contains only the titles of the research projects and subject classifications. Out of a total of about 12,800 research projects, 67 were indexed in December 1972 as being cancer related.

We mention additional automated files which contain information about cancer research projects to indicate the scope of possible sources: the National Science Library file of university research supported by the Government of Canada,[19] which contains only few cancer projects; the Atomic Energy Commission Biomedical and Environmental Research Program's file of projects it sponsors,[20] which is intended *primarily* for in-house use; the IMPAC system of the National Institutes of Health Division of Research Grants,[21] which is *strictly* for in-house use; a file underlying the categorized listing of research projects sponsored by the National Cancer Institute;[22] and WHO/BRIS (World Health Organization/Biomedical Research Information Service),[23] which has been absorbed into the WHO internal information service. The 1969 WHO/BRIS data tapes survive, but their use would require underwriting the cost of bringing the requisite computer system back into operation.[23] Most automated files which contain cancer information, however, either are based on published information[24-26] or are oriented to specific *data* that result from cancer research or treatment.[27,28]

The two privately maintained files are derived by extracting the organizational affiliations of the authors from published technical literature. Macdonald and McGuffee rely on cancer research literature as it appears in print and on several secondary sources;[10] they record the organizational name and address. Cerny follows Soviet technical literature in general; he includes brief characterizations of the types of work pursued at the several institutions,[9] including about 90 working in cancer-related areas.

Printed compilations used cover research in specific areas,[11] research funded by specific organizations,[12,13] particular countries,[14,15] medical research organizations in general,[16] funding organizations,[17] and the work of an international cancer organization.[18] "Smoking and Health"[11] presents research summaries; "Cancer Research Campaign"[12] and "Imperial Cancer Research Fund"[13] list research project titles. The remaining printed sources[14-18] list organizations, with or without brief characterizations of their areas of interest. A list of institutions derived from

† Presented in part at the 168th National Meeting of the American Chemical Society, Atlantic City, N.J., Sept. 8-13, 1974.

* Address correspondence to this author at Franklin Institute Research Laboratories, Rockville, Md. 20852.

** BioSciences Information Service of Biological Abstracts, Philadelphia, Pa. 19103.

Table I. Data Items and Sort Key Fields

Data Item	Sorting priority	Field length, bytes
Name of the parent organization	4	100
Name of the 1st-level suborganization	5	100
Name of the 2nd-level suborganization	6	100
Name of the 3rd-level suborganization	7	100
Name of the 4th-level suborganization	8	100
Name of the 5th-level suborganization	9	100
Arbitrary identification number	—	
Street address or the equivalent	—	
City	3	30
State, province, or equivalent	2	50
Postal code	—	
Country	1	40
Name of the contact person for the most specific suborganization	—	
Honorific titles of the contact person	—	
Position of the contact person	—	
Telephone number of the contact person	—	
Areas of activity or interest of the most specific suborganization	—	
Title of each specific project	—	
Name of the project director		
Family name	10	40
Given names	11	20
Project director's honorific titles	—	
Names of the associate investigators	—	
Start and stop dates for the project	—	
Level of funding for the project	—	
Sponsor(s) of the project	—	
Sponsor's code(s) for project	—	
Total key length		780

WHO/BRIS[29] and the organizations listed in the "Trends in Cancer Research"[30] were included in the Macdonald-McGuffee compilation. The topical organization of "Trends" effectively precluded reconstructing the research programs of the individual institutions.

Funding organizations and research organizations generally produce reports about their programs. The reports may be general descriptions and statistics,[31,32] but they generally also contain a list of projects they sponsor or conduct, either by title alone[33,34] or with abstracts.[35,36] Much of this information does not find its way into the SSIE files, even for apparently well-represented U.S. organizations. This suggests a need for improved research activity reporting, especially for work sponsored by other than U.S. government funds. A spot check of a small sample of the American Cancer Society's tabulation of grants[37] indicated that only about one-half of the projects so sponsored are to be found in the SSIE file. Of approximately 350 research projects in progress at M. D. Anderson Hospital and Tumor Institute,[36] only 54 appeared in the SSIE file.

Research-oriented house organs also exist,[38,39] but these should be considered as part of the journal literature.

DATA COLLECTION

Data extracted from the several sources were combined onto one file card for each distinct organization, with occasional continuation cards. The working file was organized geographically so that it would be easier to recognize variants of the same organization's name, especially when English and another language were involved. Duplicates were eliminated. In case of data discrepancies, the more recent source was relied upon.

The data sought are listed in Table I. The organizational names and the position of the contact person were recorded, if available, both in English and in the national language, transliterated if necessary into the Roman alphabet. (In the few instances when organization names were available in more than one non-English national language, as for a few Swiss institutions, an arbitrary choice was made.) The street address, city, and province as well as honorific titles were recorded in the national language when possible.

Distinction between organizations was made at the level of the most specifically identified subunit for which information was available.

DATA PROCESSING

A search using the REGISTER system was conducted for projects indexed as being cancer related. The data so identified in the Museum's card file were translated from Japanese into English and then handled like the data from printed sources.

The Smithsonian Science Information Exchange file was searched via subject terms for cancer-related projects, and a subfile was copied onto magnetic tape. Data from other sources were keyboarded into the SSIE data formats with some adaptations for data items for which SSIE has made no provision. All data were reformatted and edited, via a set of MARK IV[40] programs written for the purpose, into a file better structured for our purposes: (a) five hierarchical levels of subunit are allowed for within each parent organization instead of two; (b) separate data fields are provided for family names and given names; (c) separate data fields are provided for state or province and for country; (d) duplicate fields are provided for the data items which may occur in two languages; (e) several of the data fields are substantially longer. Additional data were then keyboarded and used to update the file.

EDITS

The reformatting and editing procedure performed functions such as separating first and middle initials from the family names; separating postal codes from street addresses; segregating organization identification numbers from project numbers (both were stored in the same SSIE field); moving the names of foreign countries into a "country" field (from the field used for states of the U.S.); entering "United States of America" into the country field as appropriate; and moving foreign province names to the state field (from the field which they shared with city names).

Our data were entered into the file in upper- and lowercase characters. The SSIE data were provided in all uppercase. For uniformity of presentation in the directory, all of the data were converted entirely to upper-case.[41]

Preliminary printouts of the data indicated that additional editing was necessary to overcome some of the idiosyncrasies of data preparation and input. Those which would have seriously affected the usability of the directory were of two types: (a) inconsistent treatment of the distinct occurrences of the same data item and (b) consistent but misleading renditions of data items. Examples of the first type involve changing "U.K." and "U. K." to "UNITED KINGDOM", "UNIV. DEGLI STUDI" to "UNIVERSITA DEGLI STUDI", and "KANAGAWA-KEN" and "KANAGAWA PREFECTURE" to "KANAGAWA". Without these changes data that should be presented in close juxtaposition would have been scattered alphabetically. The second type required changes such as from "PEOPLES REPUBLIC OF CHINA" to "CHINA (PEOPLE'S REPUBLIC)" so that this country would alphabetize among the C's.

FORMATS

The remainder of the automatic data processing was concerned with *presentation* of the data in the directory according to the detailed specifications. This required due at-

PETAH TIKVA -- (IS 23) TEL-AVIV UNIVERSITY MEDICAL SCHOOL, FOGOFF-WELLCOME MEDICAL RESEARCH INSTITUTE, BEILINSON HOSPITAL OF KUPAT HOLIM; PETAH TIKVA **
5950 • CANCER CHEMOTHERAPY.

RAMAT GAN -- (IS 24) BAR ILAN UNIVERSITY; RAMAT GAN **

REHOVOT -- (IS 25) KUPAT HOLIM HOSPITALS, KAPLAN HOSPITAL; REHOVOT **
5951 • CYTOLOGY OF TUMORS.

REHOVOTH -- (IS 26) KAPLAN HOSPITAL; REHOVOTH **

REHOVOTH -- (IS 27) WEIZMANN INSTITUTE OF SCIENCE; P.O. BOX 26, REHOVOTH
5952 STUDIES OF SYNGENEIC CELLS, IMMUNOLOGICAL TECHNIQUES, INDUCING FACTORS FOR LEUKEMIA CELLS AND PRODUCING IN VITRO LYMPHOCYTES — Unknown — NIH-NCI-G-72-3890 — $ 493,617
5953 CANCER RESEARCH — J. A. GORDON — PS-60 — $ 14,000
5954 CANCER RESEARCH — C. J. HITCH — PS-64 — $ 7,922

REHOVOTH -- (IS 28) WEIZMANN INSTITUTE OF SCIENCE, CELL BIOLOGY; P.O. BOX 26, REHOVOTH
5955 ASPECTS OF NORMAL AND MALIGNANT DIFFERENTIATION OF MYOGENIC CELL LINES — D. YAFFE (with A. SHAINBERG, H. DYM, G. KESSLER) — DRG-1007C — $ 21,450

REHOVOTH -- (IS 29) WEIZMANN INSTITUTE OF SCIENCE, GENETICS; P.O. BOX 26, REHOVOTH
5956 IDENTIFICATION AND CHARACTERISATION OF SIMIAN VIRUS 40 GENES — E. WINOCOUR — DRG-1060-B — $ 25,000

REHOVOTH -- (IS 30) WEIZMANN INSTITUTE OF SCIENCE, WOLFSON INSTITUTE OF EXPERIMENTAL BIOLOGY; REHOVOTH **
5957 * CANCER CAUSATION.

TEL AVIV -- (IS 31) ICHILOV MEDICAL CENTER; TEL AVIV **

* Areas of Activity.
** Project information from this organization not available at time of listing.

Figure 1. Part of page from the directory proper.

tention to data extraction from the file, sorting of the entries, general page layout, page headings, line folding, page breaks, and accommodation of data items which occur only sometimes. Two specific features of the programming are worth noting in passing: The sort key for the directory was 780 bytes long (cf. Table I), whereas the longest key accepted by the computer library sort routine was 256 bytes; a multipass sort was therefore used, with a computer-generated intermediate sequence number used to maintain the relative positions of entries which are not otherwise distinguishable during the *later* passes through the sort routine. The multi-column formats used in the indexes were effected (a) by first sorting the entries for each page initially by calculated *row*, then by calculated *column*,[42] and (b) by then independently printing the entries for each column in a given line without advancing the printer carriage until the entry for the last column had been printed.

THE DIRECTORY

The compilation which was produced contains 6,739 individual projects or areas of activity and 5,623 organizations in 95 countries. The five most heavily represented countries account for 6,560 projects or areas of activity and 4,690 organizations (cf. Table II). The directory was printed as a limited-edition, three-volume work[43] consisting of listings for the cancer organizations and their research projects and areas of activities, a personal name index, an organizational name index, and a keyword-in-context index to the project titles and descriptions of the areas of activity.

The directory proper is arranged geographically, by country, major national subdivision, and city. The city names are used as they appear in the institutions' mail addresses.[44] The major national subdivision is the state in the United States, the province in Canada, the prefecture in Japan, etc. For many countries, especially those with few entries, the subdivisions are omitted because they are not used or were not available. The geographic arrangement was deemed advantageous because it tends to bring together variants of names of the same organization, and it permits more readily finding organizations commonly referred to by suborganizational names rather than by the name of the parent organization. For example, the Paterson Laboratories in Manchester are generally referred to as such although they are officially part of the Christie Hospital and Radium Institute; the involved relationship among the

Table II. Countries with the Largest Numbers of Entries

Country	No. of organizations	No. of projects or areas of activity
United States of America	3,975	5,849
United Kingdom	367	581
Japan	171	72
Union of Soviet Socialist Republics	94	41
France	83	10
Canada	70	12
Australia	65	17
India	63	12

group of institutions that includes Harvard University Medical School and Massachusetts General Hospital in Boston makes it difficult to know under which name to look. If the user wishes to locate all of the cities in which a given institution conducts cancer work, he does so by using the organizational name index. The University of California, for example, is listed for 13 cities, including one in New Mexico (!); the index entries for the U.S. Veterans Administration go on for several columns.

The entries are compactly arranged across the entire page so as to minimize the number of directory pages (cf. Figure 1). The individual entries consist of the following components:

1. Organizational information
   a. City in which located
   b. Names of the "parent" organization and suborganizations
   c. Mail address

and to the extent available

   d. Name and titles of a contact person
   e. Contact person's position
   f. Telephone number
   g. Area of activity

2. Project information
   a. Project title

and to the extent available

WOOD, W. B. 4231
WOOD, W. C. 2566
WOODARD, H. Q. 4086
WOODS, J. 884
WOODS, M. W. 2198
WOODS, R. 5616

WURSTER, D. H. 3344
WUTHIER, P. 770
WYARD, S. J. UK 132
6300
WYATT, J. 948
WYKLE, R. L. 5180
5189
WYNDER, E. L. 3817

WYNDER, E. L. (Contd.) 3819
3820
3821
XENOS, J. 5922
YACHNIN, S. 1491
1492

YAZDI, E. 5346
YEATES, F. A. AS 65
YEE, J. 395
YEH, C. L. 835
YEH, J. 5774
YEH, Y. 100
YEHLE, C. O. 1002
1002

YELENOSKY, R. 336
YEN, S. 2714
YEN, T. F. 376
377
YERGANIAN, G. 2644
YESNER, R. 934

YOUNG, S. UK 4
UK 96
YOUNG, V. M. 2042
2048
2050

YOUNG, V. M. (Contd.) 2061
2063
2069
2072
2082
YOUNG, V. R. 2852

2398 874
2400 3583
2401 5140
ZAVA, D. 1647
ZAVELA, D. A. 4931
ZAWADZKI, Z. A. 4931
4932
4932

Figure 2. Part of page from the personal name index.

AMERICAN HEALTH FOUNDATION
NEW YORK, NEW YORK, U.S.A. US2473
AMERICAN JOINT COMMITTEE FOR CANCER STAGING AND REPORTING
CHICAGO, ILLINOIS, U.S.A. US 876
AMERICAN MEDICAL CENTER
DENVER, COLORADO, U.S.A. US 451
AMERICAN NATIONAL RED CROSS
WASHINGTON, DISTRICT OF COLUMBIA, U.S.A. US 594
AMERICAN ONCOLOGIC HOSPITAL
PHILADELPHIA, PENNSYLVANIA, U.S.A. US3158
AMERICAN PUBLIC HEALTH ASSOCIATION
NEW YORK, NEW YORK, U.S.A. US2474
AMERICAN RADIUM SOCIETY
BALTIMORE, MARYLAND, U.S.A. US1363
AMERICAN ROENTGEN RAY SOCIETY

BEIRUT, LEBANON LE 1
AMERICUS AND SUMTER COUNTY HOSPITAL
AMERICUS, GEORGIA, U.S.A. US 753
AMHERST COLLEGE
AMHERST, MASSACHUSETTS, U.S.A. US1609
AMHERST HOSPITAL
LORAIN, OHIO, U.S.A. US2950
AMSTERDAM UNIVERSITY
AMSTERDAM, NETHERLANDS NE 1
ANALYTICAL CHEMISTRY DEPARTMENT
NES ZIONA, ISRAEL IS 17
ANATOMICAL PATHOLOGY
HOUSTON, TEXAS, U.S.A. US3611
ANATOMY
ANN ARBOR, MICHIGAN, U.S.A. US1812
BALTIMORE, MARYLAND, U.S.A. US1425
BOSTON, MASSACHUSETTS, U.S.A. US1721
BROOKLYN, NEW YORK, U.S.A. US2358
CHICAGO, ILLINOIS, U.S.A. US 915
US 955
CLEVELAND, OHIO, U.S.A. US2866
DENVER, COLORADO, U.S.A. US 475
EAST LANSING, MICHIGAN, U.S.A. US1880
FORT COLLINS, COLORADO, U.S.A. US 493
HONOLULU, HAWAII, U.S.A. US 829
HOUSTON, TEXAS, U.S.A. US3574

ANATOMY (Contd.)
IOWA CITY, IOWA, U.S.A. US1133
KANSAS CITY, KANSAS, U.S.A. US1182
LITTLE ROCK, ARKANSAS, U.S.A. US 81
LOS ANGELES, CALIFORNIA, U.S.A. US 231
NEW HAVEN, CONNECTICUT, U.S.A. US 527
NEW ORLEANS, LOUISIANA, U.S.A. US1315
NEW YORK, NEW YORK, U.S.A. US2485
PHILADELPHIA, PENNSYLVANIA, U.S.A. US3189
ROCHESTER, NEW YORK, U.S.A. US2667
SAINT LOUIS, MISSOURI, U.S.A. US2052
SALT LAKE CITY, UTAH, U.S.A. US3702
WINSTON SALEM, NORTH CAROLINA, U.S.A. US2787
ANATOMY & BIOLOGY
PHILADELPHIA, PENNSYLVANIA, U.S.A. US3246

ANGEL H. ROFFO INSTITUTE OF ONCOLOGY
BUENOS AIRES, ARGENTINA AR 22
ANIMAL & VETERINARY SCIENCE
AMHERST, MASSACHUSETTS, U.S.A. US1611
ANIMAL BIOLOGY
PHILADELPHIA, PENNSYLVANIA, U.S.A. US3257
ANIMAL DISEASES
STORRS, CONNECTICUT, U.S.A. US 555
US 558
ANIMAL DISEASES RESEARCH INSTITUTE
LETHBRIDGE, ALBERTA, CANADA CA 5
ANIMAL HEALTH UNIT
LONDON, ENGLAND, U.K. UK 100
ANIMAL SCIENCE
ITHACA, NEW YORK, U.S.A. US2448
LAFAYETTE, INDIANA, U.S.A. US1089
ANKARA UNIVERSITESI
ANKARA, TURKEY TU 1
ANN SKOLNICK INGERMAN CHAPTER
NEW YORK, NEW YORK, U.S.A. US2547
ANNIE M. WARNER HOSPITAL
GETTYSBURG, PENNSYLVANIA, U.S.A. US3111

ANTI CANCER FOUNDATION
ADELAIDE, SOUTH AUSTRALIA, AUSTRALIA AS 32
ANTI-ACID FAST BACTERIAL DISEASES RESEARCH INSTITUTE
SENDAI CITY, MIYAGI, JAPAN JA 71
ANTI-ACID-FAST BACTERIAL DISEASES RESEARCH INSTITUTE
SENDAI CITY, MIYAGI, JAPAN JA 72
ANTI-CANCER COUNCIL
BRISBANE, QUEENSLAND, AUSTRALIA AS 25
ANTI-CANCER COUNCIL OF VICTORIA
EAST MELBOURNE, VICTORIA, AUSTRALIA AS 40
ANTI-CANER COUNCIL OF VICTORIA
EAST MELBOURNE, VICTORIA, AUSTRALIA AS 41
ANTICANCER CENTRE AVELLANEDA

ARGENTINA FILANTROPICO ASISTENCIAL DE LA CITOLOGIA DEL CANCER (DAFA
BUENOS AIRES, ARGENTINA AR 5
ARGONNE CANCER RESEARCH HOSPITAL
CHICAGO, ILLINOIS, U.S.A. US 984
ARGONNE NATIONAL LABORATORY
ARGONNE, ILLINOIS, U.S.A. US 846
LEMONT, ILLINOIS, U.S.A. US1024
ARHUS TANDLAEGENOJSKOLE
AARHUS, DENMARK DA 3
ARIZONA DIVISION
PHOENIX, ARIZONA, U.S.A. US 43
ARIZONA STATE DEPARTMENT OF HEALTH
PHOENIX, ARIZONA, U.S.A. US 44
ARIZONA STATE UNIVERSITY
TEMPE, ARIZONA, U.S.A. US 52
ARIZONA TUMOR TISSUE REGISTRY
PHOENIX, ARIZONA, U.S.A. US 45
ARKANSAS BAPTIST MEDICAL CENTER
LITTLE ROCK, ARKANSAS, U.S.A. US 72
ARKANSAS CITY MEMORIAL
ARKANSAS CITY, KANSAS, U.S.A. US1157
ARKANSAS DIVISION
LITTLE ROCK, ARKANSAS, U.S.A. US 71
ARKANSAS STATE CANCER COMMISSION
LITTLE ROCK, ARKANSAS, U.S.A. US 74

Figure 3. Part of page from the organization index.

b. Principal and associate investigators' names
c. Sponsor's project identification number
d. Level of funding

The organizations are labeled for indexing purposes with organization numbers consisting of a two-letter country code and a sequential number. Each suborganization of a given organization receives a separate entry; cf. the several entries for "Weizmann Institute of Science" in Figure 1.

All projects associated with a given organization (and suborganization) are listed alphabetically by the last name of the principal investigator. The descriptions of the areas of activity and the specific projects have sequential index numbers in a single series for the entire Directory.

Each personal name appears in the personal name index, in a four-column format (cf. Figure 2). An entry for a contact person refers via the organization number (recognizable by the presence of a two-letter country code) to the *organization* with which the person is associated. An entry for a principal or associate investigator shows the number of the specific *project* with which the investigator is associated. The same individual may be both a research investigator and a contact person; cf. "S. J. Wyard" in Figure 2. He also may be associated with more than one organization; cf. "S. Young" in the same same figure.

The name of each organization and of each suborganization appears in the organization index, in a three-column format (cf. Figure 3). If the same organization or the same suborganization of a given parent organization appears several times in the directory proper under the same city, then there is only a single entry for the entire series of occurrences of that name. However, if there is a coincidence of names among suborganizations of distinct parent organizations, then distinct entries appear in the organization index; cf. "Anatomy" for Chicago in Figure 3. In any case distinct entries appear for the same organization (or for identically named organizations) in different cities; cf. "Argonne National Laboratory" in Figure 3. Very long entries are truncated after two lines, as for "Argentina Filantropico Asistencial . . . " in the same figure.

The subject index is a two-column format keyword-in-context listing of project titles and of descriptions of areas of interest (cf. Figure 4). Each entry refers to the title or description via its index number in the directory proper. The entries are right- and left-truncated without regard to word boundaries. Long keywords simply extend beyond the

ISOZYMES PAGE 759 LABELED

STUDIES OF	KARYOTYPE AND PHENOTYPE OF MALI	4560	
STUDIES OF	KARYOTYPIC STABILITY IN DOWN'S S	6507	
ING DOCTOR SERVICE IN	KENYA, UGANDA AND TANZANIA.	6046	
SIOLOGIC RESPONSES TO	KERATINIZING TISSUES	5285	
D LOCALIZATION IN THE	KERATOACANTHOMA	3975	
	KERATOACANTHOMA OF THE VERMILION	1053	
D CHARACTERIZATION OF	KERATOHYALIN GRANULES	2222	
IN DARIER'S DISEASE (	KERATOSIS POLLICULARIS)	2213	
ION OF GLYOXALASE AND	KETOALDEHYDE DEHYDROGENASE	6169	
REGULATION OF	KETOGENESIS & LIPOGENESIS IN TIS	5252	
NG AND NEOPLASTIC RAT	KIDNEY	2753	
R BIOSYNTHESIS IN RAT	KIDNEY	6366	
Y CYSTOADENOMA OF THE	KIDNEY - REPORT OF A CASE	5546	
IAL CELL TUMOR OF THE	KIDNEY AND THE HYPERTENSIVE	5136	
YCOPROTEINS BY LIVER,	KIDNEY AND TUMOR TISSUE	5280	
	KIDNEY FUNCTION AT LONG TIME	6572	
T SITES WITHIN SINGLE	KIDNEYS	2350	
TUMOR CELL	KILLING BY X-RAYS AND IMMUNIT	3661	
HOID CELLS CAPABLE OF	KILLING TARGET CELLS	6224	
ADENYLATE	KINASE IN NORMAL AND NEOPLAS	1128	
KINASE AND THYMIDINE	KINASE IN NORMAL DEVELOPING	2753	
THE INFLUENCE OF	KIND AND LEVEL OF FAT ON D	3304	
PRODUCTION OF	KINESCOPE	1355	
	KINETIC AND BIOCHEMICAL STUDI	5512	
S ON CELL POPULATION	KINETIC ASPECTS OF CHEMOTHERA	6283	
	KINETIC CONDUCTIVITY MEASUREM	6612	
OF CELL POPULATION	KINETIC MODELS	6452	
GENTS IN THE CELLULAR	KINETICS OF NORMAL AND MALIGNA	1888	
CELL	KINETICS OF OESTROGEN INDUCED	6674	
T OF LUNG TUMOURS AND	KINETICS OF PULMONARY EPITHELI	6403	
POPULATION	KINETICS OF TUMORS IN VIVO	2789	
CELL	KINETICS UNDER CHEMICAL CARCIN	6461	
	KININS IN URINE	2541	
VO AND IN VITRO USING	KNOWN CARCINOGENIC CHEMICAL	5761	
ADMINISTRATION OF	KNOWN CARCINOGENS TO VARIOU	6042	
ON OF CELL LINES WITH	KNOWN ENZYMIC LEGIONS	6727	
R MEDICAL RESEARCH IN	KUALA LUMPUR, PENANG AND IP	6048	
THE THYMUS AND	KURLOFF BODIES	6669	
	KYNETYCS YNVOLVED YN ABSORPTYO	4961	
BUNIT INTERACTIONS OF	L-ASPARAGINASE	2389	
COCHEMICAL STUDIES OF	L-ASPARAGINASE	2682	
F ACUTE LEUKEMIA WITH	L-ASPARAGINASE	3298	
ES ON CYTOTOXICITY OF	L-ASPARAGINASE	4138	
CROBIAL PRODUCTION OF	L-ASPARAGINASE ENZYMES FOR LUEKE	1607	
	L-ASPARAGINASE ENZYMOLOGY AND TH	3092	
SEPARATION OF	L-ASPARAGINASE FROM COMPLEMENT F	2261	
GENETIC STUDIES OF	L-ASPARAGINASE OF E. COLI & THE	768	
PURIFICATION G. PIG	L-ASPARAGINASES--ANTITUMOR ACTIO	4830	
	L-ASPARAGINE - BIOSYNTHESIS, HYD	2359	
TRIPHOSPHATE POOL IN	L-CELLS	2163	
NTI-TUMOR ACTIVITY OF	L-TRYPTOPHANASE OF ORGANISM (ABB	768	
PREPARATION OF	L-3-AMINO-NUCLEOSIDES AND NUCLEO	3029	
O PREPARE RADIOACTIVE	LABELED MATERIALS	4468	
USE OF RADIOACTIVELY	LABELED PREPARATIONS TO STUDY	4144	

Figure 4. Part of page from the subject index.

"window" in the middle of the column. Keywords begin with the first alphanumeric character after a blank (or the beginning of the character string); cf. "keratosis" in Figure 4.

The stop list includes not only articles, conjunctions, prepositions, single letters (which tend to be used for numbering subtitles), and nonspecific words such as "analogs", "application", "evaluate", "programs", "research", and "science", but also (since the directory deals entirely with cancer topics and chiefly with cancer in man) words such as "cancer", "neoplasia", "tumor", "human", and "man". Titles such as "Cancer Research", "Cancer Center Exploratory Studies", and "Planning for Cancer Research Program" are totally suppressed by the stop list. Words such as "combined", "conventional", "field", "large", "primary", and "structure" are *retained* in the index because of phrases such as "combined treatment", "conventional environment", "neutron field", "large bowel", "primary cancer", and "structure and activity". The word "analysis" was retained because of its meaning of chemical analysis or bioassay, and "effectiveness" for its meaning of efficacy. By contrast "operation" was placed on the stop list because it never (in this corpus) refers to surgery, only to conducting a program. Words on the stop list were also included in their British spellings to the extent that these spellings appear in the corpus.

A listing of organizations alone was also prepared as a limited-edition work.[45] The greater simplicity of the entries due to the absence of project information permitted use of a two-column format (cf. Figure 5).

FURTHER WORK

In order to produce a definitive edition of this directory a massive update would be necessary, and some refinements should be made in presentation of the data. Such a directory might be restricted to cancer *research* organizations, since information about organizations engaged solely in cancer *control* activities serves an audience different from the research community. Annoying inconsistencies carried over from the secondary sources should be eliminated, such as "Anti-Acid-Fast Bacterial Diseases Research Institute" with and without the second hyphen (cf. Figure 3) and more or less arbitrary use of "Labs", Labs.", and "Laboratories".

The directory should furthermore be printed in upper- and lower-case.[41] The program for effecting this transformation is so written that it will also expand abbreviations and make other substitutions according to a conversion table. Some subtlety is needed to avoid obtaining, e.g.,

INDIA, UTTAR PRADESH PAGE 211

KANPUR
(IN 55) MEDICAL COLLEGE
KANPUR, UTTAR PRADESH

LUCKNOW
(IN 56) COUNCIL OF SCIENTIFIC AND INDUSTRIAL RESEARCH
CENTRAL DRUG RESEARCH INSTITUTE
P.O. BOX 173
CHATTAR MANZIL PALACE
LUCKNOW, UTTAR PRADESH
DR M. L. DHAR, MSC, PH.D., DIRECTOR OF RESEARCH

(IN 57) KING GEORGE'S MEDICAL COLLEGE
LUCKNOW, UTTAR PRADESH

INDIA, WEST BENGAL

CALCUTTA
(IN 58) CHITTARANJAN NATIONAL CANCER RESEARCH CENTRE
CHITTARANJAN CANCER HOSPITAL
37, S.P. MOOKERJEE RD.
CALCUTTA, WEST BENGAL

(IN 59) INDIAN STATISTICAL INSTITUTE
CALCUTTA, WEST BENGAL

(IN 60) INSTITUTE OF POSTGRADUATE MEDICAL EDUCATION AND RESEARCH
CALCUTTA, WEST BENGAL

(IN 61) MEDICAL DEPARTMENT
SOUTH EASTERN RAILWAY GARDEN REACH
CALCUTTA, WEST BENGAL 43

(IN 62) N.R.S. MEDICAL COLLEGE
CALCUTTA, WEST BENGAL

(IN 63) R.G. KAR MEDICAL COLLEGE
CALCUTTA, WEST BENGAL

INDONESIA, --

DJAKARTA
(ID 1) INDONESIAN CANCER SOCIETY
DJALAN. HOS. TJOKROAMINTO 37
DJAKARTA

DJAKARTA (Contd.)
(ID 3) NATIONAL INSTITUTE FOR MEDICAL RESEARCH
P.O. BOX 223
DJALAN PERTJETAKAN NEGARA I
DJAKARTA
PROFESSOR J. SULIANTI SAROSO, M.D., DP.PH.
DIRECTOR

(ID 4) UNIVERSITAS INDONESIA
FACULTAS KEDOKTERAN
SALEMBA 6
DJAKARTA
PROFESSOR M. MARDJONO, M.D., DEAN

Figure 5. Part of page from the "Listing of Cancer Research and Control Organizations" (parts of each of two columns).

"Oklahoma Saint University" or "State Luke's Hospital" from the ambiguous use of the abbreviation "ST."

The major task to be accomplished in order to produce a reliable, up-to-date directory is to verify, to update, and to augment the data. The data sources were aged to varying extents, whether printed,[11-22] automated,[7,8] or manual.[9,10] The set of *organizations* engaged in cancer research and their general programs are probably stable, but the set of specific *projects* is more likely to change over the years. The sources used are furthermore—as indicated under "Pre-Existing Sources"—*known* to be incomplete. The proportion of entries for the various countries (Table II) is to some extent an artifact inasmuch as the two automated sources used concentrate on the United States and Japan, and detailed listings were available for two British funding agencies. One possible model for a survey effectively to determine all pertinent work worldwide is that conducted by WHO/BRIS, which worked its way successively to ministers of health, directors of institutions, department heads, and eventually individual researchers.[23]

Alternate arrangements, additional indexes, and companion compilations have been considered. One or more of these might be implemented with a definitive directory. For example, effort might be made to characterize major cancer research institutions along dimensions such as facilities, staff, and budget for research, education, and patient care; cooperative arrangements; publications and reporting mechanisms; and overall program. The contents of the directory could be arranged by subject matter, or subdirectories might be prepared for areas of special interest such as "immunology" or "clinical trials". An index to cities (actually prepared for a preliminary version of the directory) facilitates finding information for a country with unfamiliar subdivisions and helps with Portland Maine/Oregon types of situations. An index for funding agencies may be useful in comprehending program interests from the standpoint of support rather than performance.

ACKNOWLEDGMENTS

The staff of the Franklin Institute Research Laboratories office in Tokyo arranged for the search via REGISTER by the Tokyo Science Museum and for translation of the search results. The search of the Smithsonian Science Information Exchange file was formulated and executed by the SSIE staff. The computer programming used was done by Messrs. Roger Dailey, William J. Frankhuizen, Eugene M. Gipe, C. L. Jefferies, Richard L. Muller, Robert J. Muller, and Ronald J. Oleksa. This work was performed under NCI Contract No. C01-CO-35403.

LITERATURE CITED

(1) Menzel, H., "Planned and Unplanned Scientific Communication," in "Proceedings of the International Conference on Scientific Information, Washington, D.C., November 16–21, 1958," National Academy of Sciences–National Research Council, Washington, D.C., 1959, pp 199 ff.

(2) Rosenberg, V., "The Application of Psychometric Techniques to Determine the Attitudes of Individuals toward Information Seeking and the Effect of the Individual's Organizational Status on These Attitudes," *Proc. Am. Doc. Inst.*, **3**, 443 (1966).

(3) Lipetz, B.-A., "Information Needs and Uses," *Annu. Rev. Inf. Sci. Technol.*, **5**, 3 (1970).

(4) "Directory of Graduate Research," American Chemical Society, Washington, D.C., 1973.

(5) "Directory of Cell Research Laboratories," UNESCO, Paris, 1969.

(6) "The National Cancer Act of 1971," PL92-218, 85 Stat. 778.

(7) "Description of Services," Smithsonian Science Information Exchange, Washington, D.C., 1972 (?).

(8) "Register Service," Japan Science Foundation, Tokyo, 1973.

(9) Cerny, G., Compiler, "Compilation of Soviet Corporate Technical Authors," Informatics Inc., Rockville, Md., 1973.

(10) Macdonald, E. J., and McGuffee, V., Ed., "Institutions Participating in Cancer Research," M. D. Anderson Hospital and Tumor Institute, Houston, Tex., 1972.

(11) "1972 Directory of On-Going Research in Smoking and Health," National Clearinghouse for Smoking and Health, Bethesda, Md., 1972.

(12) "Cancer Research Campaign, 49th Annual Report 1971," British Empire Cancer Campaign for Research, London, 1972.

(13) "The Imperial Cancer Research Fund, Sixty-Ninth Annual Report and Accounts, 1970–1971," Imperial Cancer Research Fund, London, 1972.

(14) "Cancer Services Facilities and Programs in the United States," Cancer Control Program, Health Services and Mental Health Administration, Arlington, Va., 1968.

(15) Gonen, S., Ed., "Scientific Research in Israel," Center of Scientific and Technological Information, Tel-Aviv, 1969.

(16) "Medical Research Index," 4th ed., Francis Hodgson, Guernsey, Channel Islands, 1971.

(17) "Directory of European Foundations," Fondazione Giovanni Agnelli, Torino, 1969.

(18) "Manual, International Union Against Cancer," International Union Against Cancer, Geneva, 1970.

(19) Brown, J. E., personal communication, National Science Library, Ottawa, 1974.

(20) Minthorn, M. L., personal communication, Atomic Energy Commission, Germantown, Md., 1974.

(21) "Program Codes, Organization Codes and Definitions Used in Extramural Programs," Statistics and Analysis Branch, Division of Research Grants, National Institutes of Health, Bethesda, Md., 1972.

(22) Schneider, J. H., "Analysis of NCI Research Grants by Scientific Category (Fiscal Year 1970)" National Cancer Institute, Bethesda, Md., 1971.

(23) Christiansen, O. W., Personal communication to Francois Kertesz et al., World Health Organization, Geneva, 1974.

(24) Schneider, J. H., Gechman, M., and Furth, S. E., Ed., "Survey of Commercially Available Computer-Readable Bibliographic Data Bases," American Society for Information Science, Special Interest Group for Selective Dissemination of Information (SIG/SDI), Washington, D.C., 1973, ASIS-SIG/SDI-#3.

(25) Williams, M. E., and Stewart, K., Ed., "ASIDIC Survey of Information Center Services," Illinois Institute of Technology Research Institute, Chicago, Ill., 1972.

(26) Kruzas, A. T., Ed., "Encyclopedia of Information Systems and Services," 2nd ed, Edwards Brothers, Ann Arbor, Mich., 1974.

(27) Tauber, S. J., and Werner, F. L., Ed., "Information Activities and Services of the National Cancer Institute," International Cancer Research Data Bank, National Cancer Institute, Bethesda, Md., 1973, PHS Publ. No. (NIH)74-543.

(28) "Radiotherapy Patient Information System (PROCTOR III)," M. D. Anderson Hospital and Tumor Institute, Houston, Tex., 1967 (?).

(29) "List of Institutions Undertaking Cancer Research," World Health Organization, Geneva, 1968.

(30) "Trends in Cancer Research," World Health Organization, Geneva, 1966.

(31) "Division of Cancer Biology and Diagnosis Program Reports," National Cancer Institute, Bethesda, Md., 1973.

(32) "Memorial Sloan-Kettering Cancer Center. The Year: 1972," Memorial Sloan-Kettering Cancer Center, New York, 1973.

(33) "Annual Report, National Cancer Institute of Canada, 1972–1973," National Cancer Institute of Canada, Toronto, 1973.

(34) "Annual Report of the Damon Runyon-Walter Winchell Cancer Fund," Damon Runyon-Walter Winchell Cancer Fund, New York, 1973.

(35) Godden, J. O., Ed., "Cancer in Ontario 1972–1973," The Ontario Cancer Treatment and Research Foundation, Toronto, 1973.

(36) "Research Report 1972 (September 1969–August 1971)," M. D. Anderson Hospital and Tumor Institute, Houston, Tex., 1972.

(37) "American Cancer Society Research and Clinical Investigation Grants in Effect on February 1, 1973," American Cancer Society, New York, 1973.

(38) "Clinical Bulletin," Memorial Sloan-Kettering Cancer Center, New York.

(39) "The Science Reports of the Research Institutes of Tohoku University, Series C (Medicine)."

(40) "MARK IV File Management System. Reference Manual," 2nd ed., Change 4, Informatics Inc., Canoga Park, Calif., 1973.
(41) Converting all data to upper- and lower-case would provide better legibility. A program was written for such a conversion, but time constraints prevented the final test run on live data to verify its reliability.
(42) For example, if the 236 entries which are to be printed in four columns on one page of the personal name index are designated 1 to 236 in alphabetical order, then the sort sequence for printing is 1, 60, 119, 178, 2, 61, 120, 179, 3, . . . , 235, 59, 118, 177, 236.
(43) Tauber, S. J., and Elias, A. W., "Directory of Cancer Research and Control Projects and Organizations," National Cancer Institute, Bethesda, Md., 1974.
(44) *Mail* addresses were used in order to be able to use the data in mailing lists for the type of follow-up discussed under "Further Work".
(45) Tauber, S. J., and Elias, A. W., "Listing of Cancer Research and Control Organizations," National Cancer Institute, Bethesda, Md., 1974.

A Rapid Generalized Minicomputer Text Search System Incorporating Algebraic Entry of Boolean Strategies

T. L. ISENHOUR,* W. S. WOODWARD, and S. R. LOWRY

Department of Chemistry, University of North Carolina, Chapel Hill, North Carolina 27514

Received November 11, 1974

This paper presents a rapid and efficient generalized minicomputer text searching system. The system has been applied to *Chemical Condensates* and enjoys search speeds comparable to services operating on large computer systems. Complete Boolean algebraic search strategy expressions may be used as direct entries, and all forms of truncation are automatically processed. Benchmark search speeds and results are presented for realistic profiles serving varied research groups in a major university chemistry department.

INTRODUCTION

The chemical literature has grown to the extent that only very narrow fields can be exhaustively surveyed by classical means with a reasonable expenditure of the investigator's time. Chemical Abstracts Service (CAS) presently adds about 400,000 new citations per year to the literature base which it began in 1907.

Starting in June 1968, CAS has recorded *Chemical Condensates,* a citation collection including titles, references, and keywords, but not actual abstracts, on computer readable magnetic tape. Several major commercial efforts have been made to provide current awareness search services utilizing the condensates files on large computer systems.[1-4] While these approaches have achieved a certain degree of success, they have the typical disadvantages of large, centralized systems: specifically, fairly high costs for other than very routine services; locations (and attitudes) often remote from those of the users; and increasing inflexibility as the size of the routine operation grows.

Wilde and Starke have reported on a literature search system oriented toward smaller machines.[5] While their system has been in operation for several years, search times are inconveniently slow, and complex profiles require a great deal of operator effort to translate the search logic to the format used.

This paper presents a rapid and efficient generalized minicomputer search system. The system has been applied to *Chemical Condensates* and enjoys search speeds comparable to those operating on machines costing one to two orders of magnitude more. Furthermore, complete Boolean algebra search strategy expressions may be used as direct entries, and all forms of truncation are automatically processed. Benchmark search speeds and results are presented for realistic profiles serving varied research groups in a major university chemistry department.

THE SYSTEM

The computer system involved is a 64K-byte, 1.0-μsec cycle time, Raytheon 704, equipped with two Peripheral Equipment Corp. 800 bpi, 25 ips, IBM compatible magnetic tape drives, a 500-cpm card reader, and a 300-lpm line printer. Total equipment investment is about $33,000.

BACKGROUND

The result of the development and testing of this system is a simple proof that multiple-profile searches can be done rapidly and economically on a small computer system. The system developed runs directly from standard issue Chemical Abstracts tapes, handles a number of profiles simultaneously (maximum 224 per run), handles elaborate profiles, and allows all possible Boolean logic and left and right truncation of search-text fragments. Specific results of test runs are given later.

It is perhaps necessary to dispel some of the popular misconceptions about minicomputers. First, the comparison of minicomputers to major computer installations is not analogous to that of research equipment comparisons such as low-resolution and high-resolution mass spectrometers. The numbers produced by minicomputer calculations are not of lesser quality than those generated by larger installations. Usually accuracy to any degree desired can be accomplished by a trade off in time of calculation. The principal difference between large, batch-oriented computer systems and smaller machines is more, and often faster, hardware; not necessarily more accurate hardware. Also, large systems tend to have more and diversified input/output devices as well as larger and more sophisticated operating systems. However, it should be realized that the currently popular minicomputers with approximately 1-μsec cycle times, memories from 8K to 100K bytes, and standard pe-